SOLUTIONS TO SELECTED EXERCISES

for

Essential Mathematics for Convex Optimization

Fatma Kılınç-Karzan, Tepper School of Business, Carnegie Mellon University

Arkadi Nemirovski, H. Milton Stewart School of Industrial and Systems Engineering, Georgia Institute of Technology

Exercises from Part I

5.1 Elementaries

Exercise I.1. Mark in the following list the sets which are convex:

1. $\{x \in \mathbf{R}^2 : x_1 + i^2 x_2 \leq 1,\ i = 1, \ldots, 10\}$

   Solution: convex

2. $\{x \in \mathbf{R}^2 : x_1^2 + 2ix_1x_2 + i^2x_2^2 \leq 1,\ i = 1, \ldots, 10\}$

   Solution: convex. Here is an equivalent description where convexity is evident: $\{x : |x_1 + ix_2| \leq 1,\ i = 1, \ldots, 10\}$.

3. $\{x \in \mathbf{R}^2 : x_1^2 + ix_1x_2 + i^2x_2^2 \leq 1,\ i = 1, \ldots, 10\}$

   Solution: convex (it is the intersection of ellipses)

4. $\{x \in \mathbf{R}^2 : x_1^2 + 5x_1x_2 + 4x_2^2 \leq 1\}$

   Solution: nonconvex

5. $\left\{x \in \mathbf{R}^{10} : x_1^2 + 2x_2^2 + 3x_3^2 + \ldots + 10x_{10}^2 \leq 1000x_1 - 999x_2 + 998x_3 - \ldots + 992x_9 - 991x_{10}\right\}$

   Solution: convex (ellipsoid)

6. $\{x \in \mathbf{R}^2 : \exp\{x_1\} \leq x_2\}$

   Solution: convex

7. $\{x \in \mathbf{R}^2 : \exp\{x_1\} \geq x_2\}$

   Solution: nonconvex

8. $\{x \in \mathbf{R}^n : \sum_{i=1}^n x_i^2 = 1\}$

   Solution: nonconvex

9. $\{x \in \mathbf{R}^n : \sum_{i=1}^n x_i^2 \leq 1\}$

   Solution: convex

10. $\{x \in \mathbf{R}^n : \sum_{i=1}^n x_i^2 \geq 1\}$

    Solution: nonconvex

11. $\{x \in \mathbf{R}^n : \max_{i=1,\ldots,n} x_i \leq 1\}$

    Solution: convex

12. $\{x \in \mathbf{R}^n : \max_{i=1,\ldots,n} x_i \geq 1\}$

    Solution: nonconvex, except for $n = 1$

13. $\{x \in \mathbf{R}^n : \max_{i=1,\ldots,n} x_i = 1\}$

    Solution: nonconvex, except for $n = 1$

14. $\{x \in \mathbf{R}^n : \min_{i=1,\ldots,n} x_i \leq 1\}$

Solution: nonconvex, except for $n = 1$

15. $\{x \in \mathbf{R}^n : \min\limits_{i=1,\dots,n} x_i \geq 1\}$

    Solution: convex

16. $\{x \in \mathbf{R}^n : \min\limits_{i=1,\dots,n} x_i = 1\}$

    Solution: nonconvex, except for $n = 1$

Exercise I.2. Mark by **T** those of the following claims which are always true:

1. The linear image $Y = \{Ax : x \in X\}$ of a linear subspace X is a linear subspace. *Solution:* **T**
2. The linear image $Y = \{Ax : x \in X\}$ of an affine subspace X is an affine subspace. *Solution:* **T**
3. The linear image $Y = \{Ax : x \in X\}$ of a convex set X is convex. *Solution:* **T**
4. The affine image $Y = \{Ax + b : x \in X\}$ of a linear subspace X is a linear subspace.
5. The affine image $Y = \{Ax + b : x \in X\}$ of an affine subspace X is an affine subspace. *Solution:* **T**
6. The affine image $Y = \{Ax + b : x \in X\}$ of a convex set X is convex. *Solution:* **T**
7. The intersection of two linear subspaces in $\mathbf{R}^n$ is always nonempty. *Solution:* **T**
8. The intersection of two linear subspaces in $\mathbf{R}^n$ is a linear subspace. *Solution:* **T**
9. The intersection of two affine subspaces in $\mathbf{R}^n$ is an affine subspace.
10. The intersection of two affine subspaces in $\mathbf{R}^n$, when nonempty, is an affine subspace. *Solution:* **T**
11. The intersection of two convex sets in $\mathbf{R}^n$ is a convex set. *Solution:* **T**
12. The intersection of two convex sets in $\mathbf{R}^n$, when nonempty, is a convex set. *Solution:* **T**

Exercise I.3. Prove that the relative interior of a simplex with vertices $y^0, \dots, y^m$ is exactly the set

$$\left\{\sum_{i=0}^m \lambda_i y_i : \ \lambda_i > 0, \ \sum_{i=0}^m \lambda_i = 1\right\}.$$

Solution: The claim is evident for the standard simplex $\Delta_m := \{x \in \mathbf{R}^m_+ : \ \sum_i x_i \leq 1\}$. Moreover, the set $\Delta := \mathrm{Conv}\{y^0, \dots, y^m\}$ is the image of Δ_m under the affine mapping

$$x \mapsto A(x) = y^0 + \sum_{i=1}^m x_i(y^i - y^0) : \mathbf{R}^m \to \mathbf{R}^{\dim(y)},$$

which is a one-to-one affine correspondence between $\mathbf{R}^m$ and $\mathrm{Aff}\{y^0, \dots, y^m\}$, and such a correspondence clearly maps the relative interiors of convex sets in the argument space onto the relative interiors of their images in the image space.

Exercise I.4 Which of the following claims is true:

1. The set $X = \{x : Ax \leq b\}$ is a cone if and only if $X = \{x : Ax \leq 0\}$.
2. The set $X = \{x : Ax \leq b\}$ is a cone if and only if $b = 0$.

Solution: The claim in item 1 is correct, while the claim in item 2 is not. Let us show that the claim in item 1 is correct. We immediately see that if $X = \{x : Ax \leq 0\}$, then X is clearly a cone. To see the other direction, suppose that the set $X = \{x : Ax \leq b\}$ is a cone. Then $0 \in X$, so that $b \geq 0$, and therefore the set $\overline{X} := \{x : Ax \leq 0\}$ is contained in X. Moreover, for any $x \in X$, as X is a cone we have that $tx \in X$ for all $t > 0$, and so $Ax \leq t^{-1}b$ for all $t > 0$. Then, by taking the limit of both sides of this latter inequality as $t \to +\infty$ we conclude that $Ax \leq 0$. Therefore, $X \subseteq \overline{X}$, the bottom line being that $X = \overline{X} = \{x : Ax \leq 0\}$.
A counterexample for the claim in item 2 is, e.g., $A = [1; 1]$, $b = [0; 1]$, so that $Ax \leq b$ is the system of two univariate linear inequalities $x \leq 0$, $x \leq 1$; here the solution set is a cone, but $b \neq 0$.

Exercise I.5 Suppose $\mathbf{K}$ is a closed cone. Prove that the set $X = \{x : Ax - b \in \mathbf{K}\}$ is a cone if and only if $X = \{x : Ax \in \mathbf{K}\}$.

Solution: Follows the same argument as in Exercise I.4.1.

Exercise I.6. Prove that if M is a nonempty convex set in $\mathbf{R}^n$ and $\epsilon > 0$, then for every norm $\|\cdot\|$ on $\mathbf{R}^n$, the ϵ-neighborhood of M, i.e., the set

$$M_\epsilon = \left\{ y \in \mathbf{R}^n : \inf_{x \in M} \|y - x\| \leq \epsilon \right\},$$

is convex.

Solution: Consider any $y', y'' \in M_\epsilon$ and any $\lambda \in [0;1]$; we should prove that $y := \lambda y' + (1-\lambda)y'' \in M_\epsilon$. As $y', y'' \in M_\epsilon$, using the definition of the set M_ϵ we deduce that for every $\delta > 0$ there exist $x'_\delta \in M$ and $x''_\delta \in X$ such that $\|y' - x'_\delta\| \leq \epsilon + \delta$ and $\|y'' - x''_\delta\| \leq \epsilon + \delta$. Hence,

$$\begin{aligned} \|y - \underbrace{[\lambda x'_\delta + (1-\lambda)x''_\delta]}_{:=x_\delta}\| &= \|\lambda[y' - x'_\delta] + (1-\lambda)[y'' - x''_\delta]\| \\ &\leq \lambda\|y' - x'_\delta\| + (1-\lambda)\|y'' - x''_\delta\| \\ &\leq \lambda(\epsilon + \delta) + (1-\lambda)(\epsilon + \delta) = \epsilon + \delta. \end{aligned}$$

Also, as M is convex, $x_\delta \in M$. Thus, we see that for every $\delta > 0$ there is a point $x_\delta \in M$ such that $\|y - x_\delta\| \leq \epsilon + \delta$, and since $\delta > 0$ is arbitrary, we conclude that $\inf_{x \in M} \|y - x\| \leq \epsilon$, that is, $y \in M_\epsilon$.

Exercise I.7. Which of the following claims are always true? Explain why/why not.

1. The convex hull of a bounded set in $\mathbf{R}^n$ is bounded.

   Solution: yes.

2. The convex hull of a closed set in $\mathbf{R}^n$ is closed.

   Solution: not necessarily. Consider the set $X := \{x \in \mathbf{R}^2 : x_2 \geq |x_1|^{-1}, x_1 \neq 0\}$. Note that X is closed yet its convex hull is the open half-plane $\{x \in \mathbf{R}^2 : x_2 > 0\}$.

3. The convex hull of a closed convex set in $\mathbf{R}^n$ is closed.

   Solution: yes. And the color of white horse of Alexander the Great is "white."

4. The convex hull of a closed and bounded set in $\mathbf{R}^n$ is closed and bounded.

   Solution: yes, see Corollary I.2.5.

5. The convex hull of an open set in $\mathbf{R}^n$ is open.

   Solution: yes

Exercise I.8. Let A, B be nonempty subsets of $\mathbf{R}^n$. Consider the following claims. If the claim is always (i.e., for every data satisfying premise of the claim) true, give a proof; otherwise, give a counter example.

1. If $A \subseteq B$, then $\mathrm{Conv}(A) \subseteq \mathrm{Conv}(B)$.
   Solution: evidently true.
2. If $\mathrm{Conv}(A) \subseteq \mathrm{Conv}(B)$, then $A \subseteq B$.
   Solution: evidently false. Consider $n = 1$, $A = \{1, 2, 3\}$, $B = \{1, 3\}$.
3. $\mathrm{Conv}(A \cap B) = \mathrm{Conv}(A) \cap \mathrm{Conv}(B)$.
   Solution: evidently false. Consider $n = 1$, $A = \{0, 2\}$, $B = \{1, 3\}$, resulting in $\mathrm{Conv}(A \cap B) = \varnothing$ and $\mathrm{Conv}(A) \cap \mathrm{Conv}(B) = [1, 2]$.
4. $\mathrm{Conv}(A \cap B) \subseteq \mathrm{Conv}(A) \cap \mathrm{Conv}(B)$.
   Solution: evidently true, since $A \cap B \subseteq A$, we have $\mathrm{Conv}(A \cap B) \subseteq \mathrm{Conv}(A)$. Similarly, $\mathrm{Conv}(A \cap B) \subseteq \mathrm{Conv}(B)$.
5. $\mathrm{Conv}(A \cup B) \subseteq \mathrm{Conv}(A) \cup \mathrm{Conv}(B)$.
   Solution: evidently false. Consider $n = 1$, $A = \{0\}$, $B = \{1\}$.
6. $\mathrm{Conv}(A \cup B) \supseteq \mathrm{Conv}(A) \cup \mathrm{Conv}(B)$.
   Solution: evidently true: since $A \cup B$ contains A, we have $\mathrm{Conv}(A \cup B) \supseteq \mathrm{Conv}(A)$, and similarly $\mathrm{Conv}(A \cup B) \supseteq \mathrm{Conv}(B)$.
7. If A is closed, so is $\mathrm{Conv}(A)$.
   Solution: false, see Remark I.2.6.

8. If A is closed and bounded, so is $\mathrm{Conv}(A)$.
   Solution: true, see Corollary I.2.5.
9. If $\mathrm{Conv}(A)$ is closed and bounded, so is A.
   Solution: evidently false. Consider $A = [0, 1/2) \cup \{1\}$.

Exercise I.9. Let A, B, C be nonempty subsets of $\mathbf{R}^n$ and D be a nonempty subset of $\mathbf{R}^m$. Consider the following claims. If the claim is always (i.e., for every data satisfying premise of the claim) true, give a proof; otherwise, give a counter example.

1. $\mathrm{Conv}(A \cup B) = \mathrm{Conv}(\mathrm{Conv}(A) \cup B)$.

   Solution: true. Since $A \subseteq \mathrm{Conv}(A)$ and $B \subseteq B$, we have $(A \cup B) \subseteq (\mathrm{Conv}(A) \cup B)$ and so $\mathrm{Conv}(A \cup B) \subseteq \mathrm{Conv}(\mathrm{Conv}(A) \cup B)$. To see the other direction, note that the set $\mathrm{Conv}(A \cup B)$ clearly contains both $\mathrm{Conv}(A)$ and B, that is, $\mathrm{Conv}(A \cup B) \supseteq (\mathrm{Conv}(A) \cup B)$. Moreover, $\mathrm{Conv}(A \cup B)$ is convex, implying $\mathrm{Conv}(A \cup B) \supseteq \mathrm{Conv}(\mathrm{Conv}(A) \cup B)$.

2. $\mathrm{Conv}(A \cup B) = \mathrm{Conv}(\mathrm{Conv}(A) \cup \mathrm{Conv}(B))$.

   Solution: true. Applying the preceding part twice, we get

   $$\mathrm{Conv}(A \cup B) = \mathrm{Conv}(\mathrm{Conv}(A) \cup B) = \mathrm{Conv}(B \cup \mathrm{Conv}(A)) = \mathrm{Conv}(\mathrm{Conv}(B) \cup \mathrm{Conv}(A)).$$

3. $\mathrm{Conv}(A \cup B \cup C) = \mathrm{Conv}(\mathrm{Conv}(A \cup B) \cup C)$.

   Solution: true. Applying the first part of this exercise, we get

   $$\mathrm{Conv}(A \cup B \cup C) = \mathrm{Conv}((A \cup B) \cup C) = \mathrm{Conv}(\mathrm{Conv}(A \cup B) \cup C).$$

4. $\mathrm{Conv}(A \times D) = \mathrm{Conv}(A) \times \mathrm{Conv}(D)$.

   Solution: true. Indeed, $A \times D \subseteq \mathrm{Conv}(A) \times \mathrm{Conv}(D)$. Moreover, $\mathrm{Conv}(A) \times \mathrm{Conv}(D)$ is convex, implying that $\mathrm{Conv}(A \times D) \subseteq \mathrm{Conv}(A) \times \mathrm{Conv}(D)$. To see the reverse direction, consider a point $z \in (\mathrm{Conv}(A) \times \mathrm{Conv}(D))$. Then, $z = [\sum_i \lambda_i a^i; \sum_j \mu_j d^j]$ for some weights $\lambda_i \geq 0$ summing up to 1 and $a^i \in A$, and for some weights $\mu_j \geq 0$ summing up to 1 and $d^j \in D$. Hence, $z = \sum_{i,j} \lambda_i \mu_j [a^i; d^j]$, and since $\lambda_i \mu_j \geq 0$ and $\sum_{i,j} \lambda_i \mu_j = 1$, we see that $z \in \mathrm{Conv}(A \times D)$. Thus, $\mathrm{Conv}(A) \times \mathrm{Conv}(D) \subseteq \mathrm{Conv}(A \times D)$.

5. When A is convex, to get the set $\mathrm{Conv}(A \cup B)$ (which is always the set of convex combinations of several points from A and several points from B), it suffices to take convex combinations of points with *at most one of them* taken from A, and the rest taken from B. Similarly, if A and B are both convex, to get $\mathrm{Conv}(A \cup B)$, it suffices to add to $A \cup B$ all convex combinations of pairs of points, one from A and one from B.

   Solution: Both claims are true. Indeed, $\mathrm{Conv}(A \cup B)$ is the set of all convex combinations of finite collections of points, some from A and the rest from B. Consider such a collection $z = \sum_{i \in I} \lambda_i a^i + \sum_{j \in J} \mu_j b^j$, where I, J are sets of indices, λ_i are nonnegative and $a^i \in A$, $i \in I$, μ_j are nonnegative and $b^j \in B$, $j \in J$, and the total sum of all λ_i and μ_j is 1. Justifying the first claim boils down to verifying that when A is convex, we can restrict I to be of cardinality 0 or 1. Indeed, if $\sum_{i \in I} \lambda_i = 0$, z is convex combination of points from B, and if $\alpha := \sum_{i \in I} \lambda_i > 0$, we can write $\sum_{i \in I} \lambda_i a^i = \alpha a$, where $a := \sum_{i \in I} \frac{\lambda_i}{\alpha} a^i$ is a point from A (since A is convex), that is, z can be represented as convex combination $\alpha a + \sum_{i \in J} \mu_i b^i$ of a collection where one point is from A, and all remaining points are from B, as required.
   Similarly, to justify the second claim, we should verify that when A and B are convex, the above z is either a point from A, or from B, or a convex combination of two points, one from A and one from B. When $\alpha := \sum_{i \in I} \lambda_i = 0$ or $\beta := \sum_{i \in J} \mu_i = 0$, the initial representation of z is in fact the representation of the point as convex combination of points from B, resp., from A, that is, either is a point from B, or a point from A, or both. And when $\alpha > 0$ and $\beta > 0$, we have, same as in the first claim, $z = \alpha a + \beta b$ with $a \in A$, $b \in B$, and of course $\alpha + \beta = 1$. That is, z is convex combination of a point from A and a point from B.

6. Suppose A is a set in $\mathbf{R}^n$. Consider the affine mapping $x \mapsto Px + p : \mathbf{R}^n \to \mathbf{R}^m$, and the image of A under this mapping, i.e., the set $PA + p := \{Px + p : x \in A\}$. Then, $\mathrm{Conv}(PA + p) = P\,\mathrm{Conv}(A) + p$.

Solution: trivially true. Here is the justification:

$$\begin{aligned}
\mathrm{Conv}\{PA+p\} &= \left\{\sum_i \lambda_i y^i :\ \lambda_i \geq 0,\ \sum_i \lambda_i = 1,\ y^i \in PA+p,\ \forall i\right\} \\
&= \left\{\underbrace{\sum_i \lambda_i (Px^i+p)}_{=P(\sum_i \lambda_i x^i)+p} :\ \lambda_i \geq 0,\ \sum_i \lambda_i = 1,\ x^i \in A,\ \forall i\right\} \\
&= \left\{Px+p :\ x = \sum_i \lambda_i x^i, \lambda \geq 0,\ \sum_i \lambda_i = 1\right\} \\
&= P\,\mathrm{Conv}(A) + p.
\end{aligned}$$

7. Consider an affine mapping $y \mapsto P(y) : \mathbf{R}^m \to \mathbf{R}^n$ where $P(y) := Py+p$. Recall that given a set $X \in \mathbf{R}^n$, its inverse image under the mapping $P(\cdot)$ is given by $P^{-1}(X) := \{y \in \mathbf{R}^m :\ P(y) \in X\}$. Then, $\mathrm{Conv}(P^{-1}(A)) = P^{-1}(\mathrm{Conv}(A))$.

   Solution: clearly false. Consider $m = n = 1$, $Px+p \equiv 0$, and $A = \{-1, 1\}$. Note that in this case as $0 \notin A$ we have $P^{-1}(A) = \varnothing$ and so $\mathrm{Conv}(P^{-1}(A)) = \varnothing$. On the other hand, $\mathrm{Conv}(A) = [-1, 1]$ and so $0 \in \mathrm{Conv}(A)$ and $P^{-1}(\mathrm{Conv}(A)) = \mathbf{R}^m$.

8. Consider an affine mapping $y \mapsto P(y) : \mathbf{R}^m \to \mathbf{R}^n$ where $P(y) := Py+p$. Then, $\mathrm{Conv}(P^{-1}(A)) \subseteq P^{-1}(\mathrm{Conv}(A))$.

   Solution: clearly true. Consider any $z \in \mathrm{Conv}(P^{-1}(A))$; then z is a convex combination of points from $P^{-1}(A)$, that is, $Pz+p$ is a convex combination of points from A.

Exercise I.10 Let $X_1, X_2 \in \mathbf{R}^n$ be two nonempty sets, and define $Y := X_1 \cup X_2$ and $Z := \mathrm{Conv}(Y)$. Consider the following claims. If the claim is always (i.e., for every data satisfying premise of the claim) true, give a proof; otherwise, give a counter example.

1. Whenever X_1 and X_2 are both convex, so is Y.

   Solution: Obviously false. Take $n = 1$, and $X_1 := \{-1\}$ and $X_2 := \{+1\}$.

2. Whenever X_1 and X_2 are both convex, so is Z.

   Solution: Obviously true by definition of Z.

3. Whenever X_1 and X_2 are both bounded, so is Y.

   Solution: Obviously true.

4. Whenever X_1 and X_2 are both bounded, so is Z.

   Solution: Obviously true.

5. Whenever X_1 and X_2 are both closed, so is Y.

   Solution: Obviously true - closedness is preserved by taking finite unions.

6. Whenever X_1 and X_2 are both closed, so is Z.

   Solution: This is false as Z is not necessarily closed. Indeed, this claim is not valid even when X_1, X_2 are nonempty polyhedral, but not bounded, sets. For example, by selecting $n = 2$, $X_1 := \{x \in \mathbf{R}^2 : x_1 \geq 0, x_2 = 0\}$ and $X_2 := \{[0; 1]\}$, we see that the set $\mathrm{Conv}(X_1 \cup X_2)$ is not polyhedral, but its closure is.

7. Whenever X_1 and X_2 are both compact, so is Y.

   Solution: Obviously true – Y is closed and bounded along with X_1 and X_2.

8. Whenever X_1 and X_2 are both compact, so is Z.

   Solution: Obviously true – by previous item, Y is compact, so that Z is compact by Corollary I.2.5.

9. Whenever X_1 and X_2 are both polyhedral, so is Y.

Solution: Obviously false. Take $n = 1$, and $X_1 := \{-1\}$ and $X_2 := \{+1\}$.

10. Whenever X_1 and X_2 are both polyhedral, so is Z.

    Solution: This is false as Z is not necessarily closed, see solution to item 6, and closedness for a polyhedral set is a must.

11. Whenever X_1 and X_2 are both polyhedral and bounded, so is Y.

    Solution: Obviously false. Take $n = 1$, and $X_1 := \{-1\}$ and $X_2 := \{+1\}$.

12. Whenever X_1 and X_2 are both polyhedral and bounded, so is Z.

    Solution: This claim is indeed true, see solution to Exercise I.22.2 for a proof.

Exercise I.11. Consider two families of convex sets given by $\{F_i\}_{i\in I}$ and $\{G_j\}_{j\in J}$. Prove that the following relation holds:

$$\mathrm{Conv}\left(\bigcup_{i\in I,\, j\in J}(F_i \cap G_j)\right) \subseteq \mathrm{Conv}\left(\bigcup_{j\in J}\left[G_j \cap \mathrm{Conv}(\cup_{i\in I}F_i)\right]\right).$$

Solution: Note that for all $j \in J$ and for all $i' \in I$, we have

$$(F_{i'} \cap G_j) \subseteq [G_j \cap (\cup_{i\in I}F_i)] \subseteq [G_j \cap \mathrm{Conv}(\cup_{i\in I}F_i)]\,,$$

and so for all $j \in J$

$$\bigcup_{i'\in I}(F_{i'} \cap G_j) \subseteq [G_j \cap \mathrm{Conv}(\cup_{i\in I}F_i)]\,.$$

By first taking the union of both sides over $j \in J$ and then taking the convex hull of the resulting sets, we arrive at the desired relation.

Exercise I.12. Let C_1, C_2 be two nonempty conic sets in $\mathbf{R}^n$, i.e., for each $i = 1, 2$, for any $x \in C_i$ and $t \geq 0$, we have $t \cdot x \in C_i$ as well. Note that C_1, C_2 are not necessarily convex. Prove that

1. $C_1 + C_2 \neq \mathrm{Conv}(C_1 \cup C_2)$ may happen if either C_1 or C_2 (or both) is nonconvex.

   Solution: Let C_1 be the origin in $\mathbf{R}^2$, and C_2 be the union of nonnegative rays of the coordinate axes. Here both sets are nonempty and conic, their sum is C_2, and the convex hull of their union (which is C_2) is the first quadrant.

2. $C_1 + C_2 = \mathrm{Conv}(C_1 \cup C_2)$ always holds if C_1, C_2 are both convex.

   Solution: When C_1, C_2 are nonempty and convex, we have by Exercise I.9.5 that $\mathrm{Conv}(C_1 \cup C_2) = \{x = \alpha y + (1-\alpha)z : y \in C_1, z \in C_2, \alpha \in [0,1]\}$, whence $\mathrm{Conv}(C_1 \cup C_2) = C_1 \cup C_2 \cup \{x = \alpha y + (1-\alpha)z : y \in C_1, z \in C_2, \alpha \in (0,1)\} = C_1 \cup C_2 \cup \big(\cup_{\alpha\in(0,1)}[\alpha C_1 + (1-\alpha)C_2]\big)$. When C_1, C_2, in addition to being nonempty and convex, are also conic, for $\alpha \in (0,1)$ it holds $\alpha C_1 + (1-\alpha)C_2 = C_1 + C_2$, so that the above computation results in $\mathrm{Conv}(C_1 \cup C_2) = C_1 \cup C_2 \cup [C_1 + C_2]$. The latter union is just $C_1 + C_2$, since $C_1 + C_2$ contains both C_1 and C_2 (as a nonempty conic set contains the origin).

3. The equality $C_1 \cap C_2 = \bigcup_{\alpha\in[0,1]}(\alpha C_1 \cap (1-\alpha)C_2)$ always holds if C_1, C_2 are both convex.

   Solution: We have

   $$\cup_{\alpha\in[0,1]}[\alpha C_1 \cap (1-\alpha)C_2] = [0 \cdot C_1 \cap 1 \cdot C_2] \cup [1 \cdot C_1 \cap 0 \cdot C_2] \cup \big(\cup_{\alpha\in(0,1)}[\alpha C_1 \cap (1-\alpha)C_2]\big),$$

   and the set in parentheses $\big(\big)$ is just $C_1 \cap C_2$ due to the conicity of C_1, C_2. Besides this, as it was mentioned when solving item 2, $0 \in C_1 \cap C_2$, so that $[0 \cdot C_1 \cap C_2] = [C_1 \cap 0 \cdot C_2] = \{0\} \subset C_1 \cap C_2$. The bottom line is that $\cup_{\alpha\in[0,1]}[\alpha C_1 \cap (1-\alpha)C_2] = C_1 \cap C_2$, as claimed.

Exercise I.13. Let $X \subseteq \mathbf{R}^n$ be a convex set with $\mathrm{int}\, X \neq \varnothing$, and consider the following set

$$\mathbf{K} := \mathrm{cl}\,\{[x;t] :\ t > 0,\, x/t \in X\}\,.$$

Prove that the set $\mathbf{K}$ is a closed cone with a nonempty interior.

Solution: $\mathbf{K}$ is what was in section 1.5 called closed conic transform of X; it was shown in section 1.5 that $\mathbf{K}$ is a closed cone. When $\overline{x} \in \mathrm{int}\, X$, we clearly have $[\overline{x};1] \in \mathrm{int}\,\mathbf{K}$, so that $\mathrm{int}\,\mathbf{K} \neq \varnothing$ whenever $\mathrm{int}\, X \neq \varnothing$.

5.2 Around ellipsoids

Exercise I.14. Verify each of the following statements:

1. Any ellipsoid $E \in \mathbf{R}^n$ is the image of the unit Euclidean ball $B_n = \{x \in \mathbf{R}^n : \|x\|_2 \le 1\}$ under a one-to-one affine mapping. That is, $E \subset \mathbf{R}^n$ can be represented as $E = \{x : (x-c)^\top C(x-c) \le 1\}$ with $C \succ 0$ and $c \in \mathbf{R}^n$ if and only if it can be represented as $E = \{c + Du : u \in B_n\}$ with nonsingular D, and in the latter representation D can be selected to be symmetric positive definite.

   Solution: Let $E = \{x : (x-c)^\top C(x-c) \le 1\}$ with $C \succ 0$. Then, by defining $H := C^{1/2}$ (see section D.1.5) we have

   $$E = \{x : (H(x-c))^\top \underbrace{(H(x-c))}_{:=u} \le 1\} = \{x = c + \underbrace{H^{-1}}_{=:D} u : u^\top u \le 1\}$$

   where $D = H^{-1} \succ 0$ as $C \succ 0$. For the other direction, given a nonsingular D, to say that $x = c + Du$ with some u satisfying $\|u\|_2 \le 1$, is the same as to say that $\|D^{-1}(x-c)\|_2 \le 1$, that is, the same as to say that $(x-c)^\top \underbrace{D^{-\top}D^{-1}}_{=:C}(x-c) \le 1$ (by definition, $D^{-\top} = (D^{-1})^\top$), and $C := D^{-\top}D^{-1}$ is symmetric positive definite since D^{-1} is nonsingular.

2. Given $C \succ 0$, $D \succ 0$ and $c, d \in \mathbf{R}^n$, the ellipsoid $E_C := \{x : (x-c)^\top C(x-c) \le 1\}$ is contained in the ellipsoid $E_D := \{x : (x-c)^\top D(x-c) \le 1\}$ if and only if $C \succeq D$. If the ellipsoid E_C is contained in the ellipsoid $E'_D = \{x : (x-d)^\top D(x-d) \le 1\}$, then $C \succeq D$.

   Solution: The first claim: Setting $x = y + c$, we should prove that with positive definite C, D, the implication $y^\top Cy \le 1 \implies y^\top Dy \le 1$ holds true if and only if $C \succeq D$. By homogeneity, the implication in question is the same as the relation

   $$\forall (s, y : s > 0, y^\top Cy \le s) : \quad y^\top Dy \le s,$$

   which for $C \succ 0$ is exactly the same as $C \succeq D$.
   The second claim: Suppose $E_C \subseteq E'_D$. Then, using part 1,

   $$\begin{array}{rll} u^\top u \le 1 & \iff & c + C^{-1/2}u \in E_C \\ & \implies & c + C^{-1/2}u \in E'_D \\ & \implies & (C^{-1/2}u + c - d)^\top D(C^{-1/2}u + c - d) \le 1 \\ & \implies & (u+f)^\top \underbrace{(C^{-1/2}DC^{-1/2})}_{=:H}(u+f),\ f := C^{1/2}(c-d). \end{array}$$

   Applying the resulting inequality to $-u$ in the role of u, we conclude that

   $$u^\top u \le 1 \implies (f \pm u)^\top H(f \pm u) \le 1,$$

   whence

   $$u^\top u \le 1 \implies 1 \ge \frac{1}{2}\left((f+u)^\top H(f+u) + (f-u)^\top H(f-u)\right) = u^\top Hu + f^\top Hf \ge u^\top Hu,$$

   where the concluding inequality is due to $H \succ 0$ (implied by $C, D \succ 0$). Hence, we arrive at

   $$I \succeq H = C^{-1/2}DC^{-1/2},$$

   implying, after multiplying both sides from the right and from the left by the symmetric matrix $C^{1/2}$, that $C \succeq D$, as claimed.

3. For a set $U \subset \mathbf{R}^n$, let $\mathrm{Vol}(U)$ be the ratio of the n-dimensional volume of U and the n-dimensional volume of the unit ball B_n. Then, for an n-dimensional ellipsoid E represented as $\{x = c + Du : \|u\|_2 \le 1\}$ with nonsingular D we have

   $$\mathrm{Vol}(E) = |\mathrm{Det}(D)|,$$

and when E is represented as $\{x : (x-c)^\top C(x-c) \le 1\}$ with $C \succ 0$, we have

$$\mathrm{Vol}(E) = \mathrm{Det}^{-1/2}(C).$$

Solution: The first relation is evident – one-to-one affine transformation $u \mapsto c + Du$ multiplies n-dimensional volumes by $|\mathrm{Det}(D)|$. Using item 1, we see that the second representation of E is equivalent to the first representation with $D := C^{-1/2}$, so that the second representation of $\mathrm{Vol}(E)$ is readily given by the first one.

Exercise I.15. Given $C \succ 0$, an ellipsoid $\{x : (x-a)^\top C(x-a) \le 1\}$ is the solution set of quadratic inequality $x^\top Cx - 2(Ca)^\top x + (a^\top Ca - 1) \le 0$. Prove that the solution set E of any quadratic inequality $f(x) := x^\top Cx - c^\top x + \sigma \le 0$ with positive *semi*definite matrix C is convex.

Solution: Let $x, y \in E$ and $\lambda \in [0,1]$. Then,

$$\begin{aligned}
&f(\lambda x + (1-\lambda)y)\\
&= \big(\lambda^2 x^\top Cx + \lambda(1-\lambda)x^\top Cy + \lambda(1-\lambda)y^\top Cx + (1-\lambda)^2 y^\top Cy\big)\\
&\quad - \lambda c^\top x - (1-\lambda)c^\top y + \lambda\sigma + (1-\lambda)\sigma\\
&= \lambda(x^\top Cx - c^\top x + \sigma) + (1-\lambda)(y^\top Cy - c^\top y + \sigma) - \underbrace{\lambda(1-\lambda)((x-y)^\top C(x-y))}_{\ge 0 \text{ due to } C \succeq 0}\\
&\le \lambda \underbrace{f(x)}_{\le 0} + (1-\lambda)\underbrace{f(y)}_{\le 0} \le 0.
\end{aligned}$$

That is, $\lambda x + (1-\lambda)y \in E$.

5.3 Truss Topology Design

Exercise I.16. [First acquaintance with Truss Topology Design]
Preamble. What follows is the first exercise in a "Truss Topology Design" (TTD) series ((other exercises in it are I.18, III.9, IV.11, IV.28). The underlying "real life" mechanical story is simple enough to be told and rich enough to illustrate numerous constructions and results presented in the main body of our textbook – ranging from Caratheodory Theorem to semidefinite duality, demonstrating on a real life example how the theory works.
Trusses. Truss is a mechanical construction, like railroad bridge, electric mast, or Eiffel Tower, composed of thin elastic *bars* linked with each other at *nodes* – points from physical space (3D space for spatial, and 2D space for planar trusses).

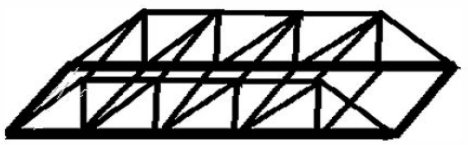


Figure 5.1. Pratt Truss Bridge

When truss is subject to external load – collection of forces acting at the nodes – it starts to deform, so that the nodes move a little bit, leading to elongations/shortenings of bars, which, in turn, result in reaction forces. At the equilibrium, the reaction forces compensate the external ones, and the truss capacitates certain potential energy, called *compliance*. Mechanics models this story as follows.

- The nodes form a finite set $p_1, \ldots, p_K$ of distinct points in physical space $\mathbf{R}^d$ ($d = 2$ for planar, and $d = 3$ for spatial constructions). Virtual displacements of the nodes under the load are somehow restricted by "support conditions;" we will focus on the case when some of the nodes "are fixed" – cannot move at all (think about them as being in the wall), and the remaining "are free" – their virtual displacements form the entire $\mathbf{R}^d$. A virtual displacement v of the nodal set can be identified with a vector of dimension $M = dm$, where m is the number of free nodes;

v is block vector with m d-dimensional blocks, indexed by the free nodes, representing physical displacements of these nodes.

- There are N bars, i-th of them linking the nodes with indexes α_i and β_i (with at least one of these nodes free) and with volume (3D or 2D, depending on whether the truss is spatial or planar) t_i.
- An external load is a collection of physical forces – vectors from $\mathbf{R}^d$ – acting at the free nodes (forces acting at the fixed nodes are of no interest – they are suppressed by the supports). Thus, an external load f can be identified with block vector of the same structure as a virtual displacement – blocks are indexed by free nodes and represent the external forces acting at these nodes. Thus, displacements v of the nodal set and external loads f are vectors from the space $\mathcal{V}$ of *virtual displacements* – M-dimensional block vectors with m d-dimensional blocks.
- The bars and the nodes together specify the symmetric positive semidefinite $M \times M$ *stiffness matrix* A of the truss. The role of this matrix is as follows. A displacement $v \in \mathcal{V}$ of the nodal set results in reaction forces at free nodes (those at fixed nodes are of no interest – they are compensated by supports); assembling these forces into M-dimensional block-vector, we get a *reaction*, and this reaction is $-Av$. In other words, the potential energy capacitated in truss under displacement $v \in \mathcal{V}$ of nodes is $\frac{1}{2}v^\top Av$, and reaction, as it should be, is the minus gradient of the potential energy as a function of v [2]. At the equilibrium under external load f, the total of the reaction and the load should be zero, that is, the equilibrium displacement satisfies

$$Av = f \tag{5.1}$$

  Note that (5.1) may be unsolvable, meaning that the truss is crushed by the load in question. Assuming the equilibrium displacement v exists, the truss at equilibrium capacitates potential energy $\frac{1}{2}v^\top Av$; this energy is called *compliance* of the truss w.r.t. the load. Compliance is convenient measure of rigidity of the truss with respect to the load, the less the compliance the better the truss withstands the load.

Let us build the stiffness matrix of a truss. As we have mentioned, the reaction forces originate from elongations/shortenings of bars under displacement of nodes. Consider i-th bar linking nodes with initial – prior to the external load being applied – positions $a_i = p_{\alpha_i}$ and $b_i = p_{\beta_i}$, and let us set

$$d_i = \|b_i - a_i\|_2,\ e_i = [b_i - a_i]/d_i.$$

Under displacement $v \in \mathcal{V}$ of the nodal set,

- positions of the nodes linked by the bar become $a_i + \underbrace{v^{\alpha_i}}_{da}$, $b_i + \underbrace{v^{\beta_i}}_{db}$, where v^γ is γ-th block in v – the displacement of γ-th node
- as a result, elongation of the bar becomes, in the first-order in v approximation, $e_i^\top[db - da]$, and the reaction forces caused by this elongation by Hooke's Law[3] are

$$\begin{array}{cc} d_i^{-1}S_i e_i e_i^\top[db - da] & \text{at node \# } \alpha_i \\ -d_i^{-1}S_i e_i e_i^\top[db - da] & \text{at node \# } \beta_i \\ 0 & \text{at all remaining nodes} \end{array}$$

  where $S_i = t_i/d_i$ is the cross-sectional size of i-th bar. It follows that *when both nodes linked by i-th bar are free, the contribution of i-th bar to the reaction is*

$$-t_i \mathfrak{b}_i \mathfrak{b}_i^\top v,$$

[2] This is called *linearly elastic* model; it is the linearized in displacements approximation of the actual behavior of a loaded truss. This model works the better the smaller are the nodal displacements as compared to the inter-nodal distances, and is accurate enough to be used in typical real-life applications.

[3] Hooke's Law says that the magnitude of the reaction force caused by elongation/shortening of a bar is proportional to $Sd^{-1}\delta$, where S is bar's cross-sectional size (area for spatial, and thickness for planar truss), d is bar's (pre-deformation) length, and δ is the elongation. With units of length properly adjusted to bars' material, the proportionality coefficient becomes 1, and this is what we assume from now on.

where $\mathfrak{b}_i \in \mathcal{V}$ is the vector with just two nonzero blocks:
— the block with index α_i – this block is $e_i/d_i = [b_i - a_i]/\|b_i - a_i\|_2^2$, and
— the block with index β_i – this block is $-e_i/d_i = -[b_i - a_i]/\|b_i - a_i\|_2^2$.
It is immediately seen that when just one of the nodes linked by i-th bar is free, the contribution of i-th bar to the reaction is given by similar relations, but with one, rather than 2, blocks in $\mathfrak{b}_i$ – the one corresponding to the free among the nodes linked by the bar.

The bottom line is that *The stiffness matrix of a truss composed of N bars with volumes t_i, $1 \le i \le N$, is*

$$A = A(t) := \sum_i t_i \mathfrak{b}_i \mathfrak{b}_i^\top,$$

where $\mathfrak{b}_i \in \mathcal{V} = \mathbf{R}^M$ are readily given by the geometry of nodal set and the indexes of nodes linked by bar i.

Truss Topology Design problem. In the simplest Truss Topology Design (TTD) problem, one is given

- a finite *set of tentative nodes* in 2D or 3D along with support conditions indicating which of the nodes are fixed and which are free, and thus specifying the linear space $\mathcal{V} = \mathbf{R}^M$ of virtual displacements of the nodal set,
- the *set of N tentative bars* – unordered pairs of (distinct from each other) nodes which are allowed to be linked by bars, and the total volume $W > 0$ of the truss,
- An external load $f \in \mathcal{V}$.

These data specify, as explained above, vectors $\mathfrak{b}_i \in \mathbf{R}^M$, $i = 1, \ldots, N$, and the stiffness matrix

$$A(t) = \sum_{i=1}^N t_i \mathfrak{b}_i \mathfrak{b}_i^\top = B\,\mathrm{Diag}\{t_1, \ldots, t_N\} B^\top \in \mathbf{S}^M \qquad [B = [\mathfrak{b}_1, \ldots, \mathfrak{b}_N]]$$

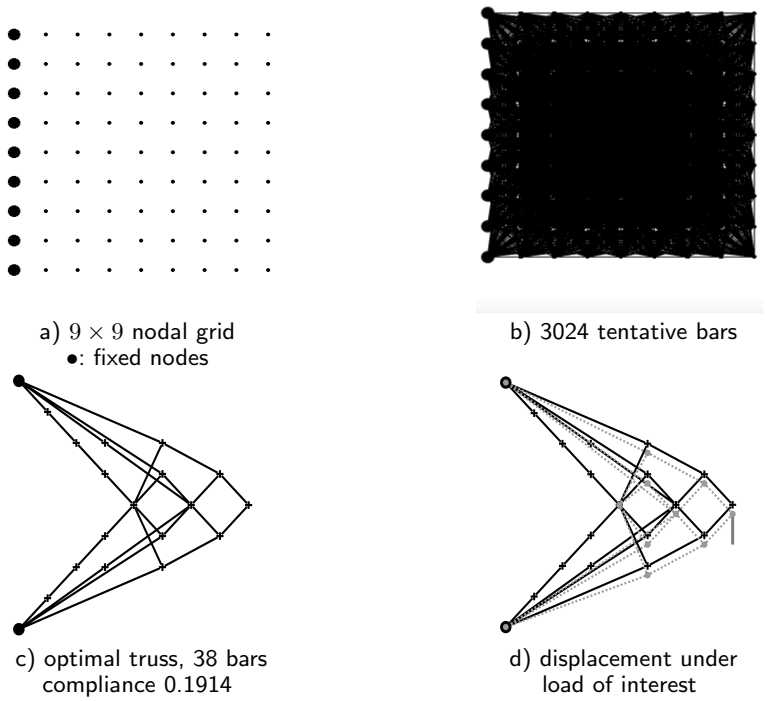
of truss, which under the circumstances can be identified with vector $t \in \mathbf{R}^N_+$ of bar volumes. What we want is to find the truss of given volume capable to "withstand best of all" the given load, that is, the one that minimizes the corresponding compliance.

When applying the TTD model, one starts with dense grid of tentative nodes and broad list of tentative bars (e.g., by allowing to link by a bar every pair of distinct from each other nodes, with at least one of the nodes in the pair free). At the optimal truss yielded by the optimal solution to the TTD problem, many tentative bars (usually vast majority of them) get zero volumes, and significant part of the tentative nodes become unused. Thus, TTD problem in fact is not about sizing – it allows to recover optimal structure of the construction, this is where "Topology Design" comes from.

To illustrate this point, here is a toy example (it will be our guinea pig in the entire series of TTD exercises):

Console design: We want to design a 2D truss as follows:

- The set of tentative nodes is the 9×9 grid $\{[p;q] \in \mathbf{R}^2 : p, q \in \{0, 1, \ldots, 8\}\}$ with the 9 most-left nodes fixed and remaining 72 nodes free, resulting in $M = 144$-dimensional space $\mathcal{V}$ of virtual displacements
- The external load $f \in \mathcal{V} = \mathbf{R}^{144}$ is a single-force one, with the only nonzero force $[0; -1]$ applied at the 5-th node of the most-right column of nodes.
- We allow for all pairwise connections of pairs of distinct from each other nodes, with at least one of these nodes free, resulting in $N = 3204$ tentative bars
- The total volume of truss is $W = 1000$.



a) 9×9 nodal grid
●: fixed nodes

b) 3024 tentative bars

c) optimal truss, 38 bars
compliance 0.1914

d) displacement under
load of interest

Figure 4.1. Console. d): positions of the bars and nodes before and after (in gray) deformation. The vertical segment starting at the right-most node: the external force.

Important: *From now on, speaking about TTD problem, we always make the following assumption:*

$\mathfrak{R}$: $\sum_{t=1}^{N} \mathfrak{b}_i \mathfrak{b}_i^\top \succ 0.$

Under this assumption, the stiffness matrix $A(t) = \sum_i t_i \mathfrak{b}_i \mathfrak{b}_i^\top$ associated with truss $t > 0$ is positive definite, so that such a truss can withstand whatever load f.

You can verify numerically that this is the case in Console design as stated above.

After this lengthy preamble (to justify its length, note that it is investment to a series of exercises, rather than just one of them), let us pass to the exercise per se. Consider a TTD problem.

1. Prove that truss $t \geq 0$ (recall that we identify truss with the corresponding vector of bar volumes) is capable to carry load f if and only if the quadratic function

$$F(v) = f^\top v - \frac{1}{2} v^\top A(t) v$$

is bounded from above, and that whenever this takes place,

- the maximum of F over $\mathcal{V}$ is achieved
- the maximizers of F are exactly the equilibrium displacements v – those with

$$A(t)v = f,$$

  and for such a displacement, one has

$$[\max F =]\ F(v) = \frac{1}{2}v^\top A(t)v = \frac{1}{2}v^\top f$$

- the maximum value of F is exactly the compliance of the truss w.r.t. the load f

Solution: Observe, first, that a quadratic function

$$G(v) = g^\top v - \frac{1}{2}v^\top Av : \mathbf{R}^M \to \mathbf{R}$$

with $A \succeq 0$ attains its maximum if and only if it is bounded from above, and that the maximizers v of G are exactly the solutions v to the Fermat equation

$$[\nabla G(v) =]\ \ g - Av = 0.$$

Indeed, invoking eigenvalue decomposition $A = U\,\mathrm{Diag}\{\lambda\}U^\top$ of A (here U is orthogonal, and $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_M$ are the eigenvalues of A; note that $\lambda_M \geq 0$ since $A \succeq 0$) and representing $v \in \mathbf{R}^M$ by the vector $\overline{v} = U^\top v$ of coordinates of v in the eigenbasis of A, we get

$$G(v) = \overline{f}^\top \overline{v} - \frac{1}{2}\sum_{i=1}^M \lambda_i \overline{v}_i^2. \qquad [\overline{f} = U^\top f]$$

We conclude that G is bounded from above if and only if $\overline{f}_i = 0$ for all i such that $\lambda_i = 0$, and in this case G attains its maximum, the maximizers being exactly v's such that $\lambda_i \overline{v}_i = \overline{f}_i$ for all i, or, which is the same, all v's such that $Av = f$. For such a v, if any,

$$G(v) = \sum_i [\overline{f}_i \overline{v}_i - \frac{1}{2}\lambda_i \overline{v}_i^2] = \frac{1}{2}\sum_i \lambda_i \overline{v}_i^2 = \frac{1}{2}v^\top Av = \frac{1}{2}v^\top f.$$

It remains to note that by definition of the compliance of truss t w.r.t. load f, this compliance is finite if and only if the equation $Av = f$ in variables f has a solution $v = v_f$, in which case the compliance is $\frac{1}{2}v_f^\top Av_f$. ■

Note: From the above analysis, it follows that our original definition of compliance indeed makes sense – while the equilibrium displacement v – the one such that $Av = f$ – when exists, not necessarily is uniquely defined by A and f, the analysis we have just carried out shows that when v_f exists, the quantity $\frac{1}{2}v_f^\top Av_f$ *is* uniquely defined by A and f.

2. Prove that a real τ is an upper bound on the compliance of truss $t \geq 0$ w.r.t. load f if and only if the symmetric matrix

$$\mathcal{A} = \left[\begin{array}{c|c} B\,\mathrm{Diag}\{t\}B^\top & f \\ \hline f^\top & 2\tau \end{array}\right],\ B = [\mathfrak{b}_1, \ldots, \mathfrak{b}_N]$$

   is positive semidefinite. As a result, pose the TTD problem as the optimization problem

$$\mathrm{Opt} = \min_{\tau, r}\left\{\tau : \left[\begin{array}{c|c} B\,\mathrm{Diag}\{t\}B^\top & f \\ \hline f^\top & 2\tau \end{array}\right] \succeq 0, t \geq 0, \sum_i t_i = W\right\} \tag{5.2}$$

   Prove that the problem is solvable.

Solution: As we have already seen, the compliance $\mathcal{C}$ of truss t w.r.t. f does not exceet a real τ iff $\tau \geq \sup_v F(v)$, that is, setting $A = A(t) = B\,\mathrm{Diag}\{t\}B$,

$$\begin{array}{rl} & \tau \geq \mathcal{C} \\ \Longleftrightarrow & \tau \geq f^\top v - \frac{1}{2} v^\top A v \,\forall v \\ \Longleftrightarrow & \tau - f^\top v + \frac{1}{2} v^\top A v \geq 0 \,\forall v \\ \Longleftrightarrow & 2\tau - 2f^\top v + v^\top A v \geq 0 \,\forall v \\ \Longleftrightarrow & 2\tau s^2 + 2s f^\top u + u^\top A u \geq 0 \,\forall (s \neq 0, u) \text{ [look at } v = -u/s] \\ \Longleftrightarrow & 2\tau s^2 + 2s f^\top u + u^\top A u \geq 0 \,\forall [u; s] \in \mathbf{R}^{M+1} \text{ [by continuity]} \\ \Longleftrightarrow & \left[\begin{array}{c|c} A & f \\ \hline f^\top & 2\tau \end{array}\right] \succeq 0 \qquad \blacksquare \end{array}$$

To prove that the problem is solvable, note that $BB^\top \succ 0$, implying that every $t > 0$ such that $\sum_i t_i = W$ can be augmented by large enough τ to yield a feasible solution. Thus, (5.2) is feasible. Since for every feasible solution to the problem τ is nonnegative, the objective is below bounded on the (nonempty!) feasible set, so that the infimum Opt of the value of s objective at feasible solutions is nonnegative real. We can find sequence $[t^j, \tau^j]$ of feasible solutions with $\tau^j \to \mathrm{Opt}$ as $j \to \infty$. By feasibility, t^j form a bounded sequence, so that passing to a subsequence, we can assume that $\lim_{j\to\infty}[t^j; \tau^j]$ exists; clearly, this limit is a feasible solution, and the τ-component of this solution is Opt, implying that this solution is optimal. ■

3. [computational study]

   3.1. Solve the Console problem numerically and reproduce the numerical results presented above.

   3.2. Resolve the problem with the set of all possible tentative bars reduced to the subset of "short" bars connecting neighboring nodes only:

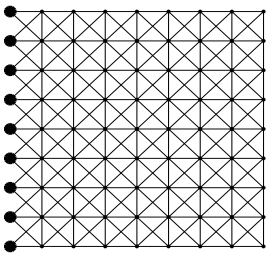
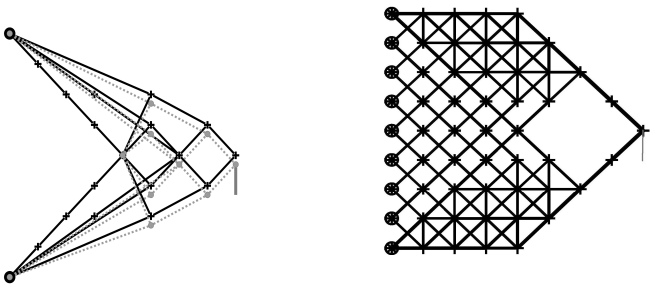


Figure 5.3. 262 "short" tentative bars

and compare the resulting design and compliance to those in the previous item.

Solution: 3.2: Here are our results:



"Long-bar truss:" 38 bars, compliance 0.1914 "Short-bar truss:" 128 bars, compliance 0.2903

The vertical segments starting at the right-mpst nodes: the external force.

5.4 Around Caratheodory Theorem

Exercise I.17. Prove the following statement: Let $X \subset \mathbf{R}^n$ be nonempty. Then

1. if a point x can be represented as a convex combination of a collection of vectors from X, then the collection can be selected to be affinely independent.
2. if a point x can be represented as a conic combination of a collection of vectors from X, then the collection can be selected to be linearly independent,

Note that the claims above are refinements, albeit minor ones, of the Caratheodory Theorem (plain and conic, respectively). Indeed, when $M = \mathrm{Aff}(X)$ and m is the dimension of M, every affine independent collection of points from X contains at most $m+1$ points (Proposition A.44), so that the first claim implies that if $x \in \mathrm{Conv}(X)$, then x is a convex combination of at most $m+1$ points from X; however, the vectors participating in such a combination are not necessarily affinely independent, so that the first claim provides a bit more information than the plain Caratheodory's Theorem. Similarly, if $L = \mathrm{Lin}(X)$ and $m = \dim L$, then every linearly independent collection of vectors from X contains at most $m \le n$ points, that is, the second claim implies the Caratheodory's Theorem in conic form, and provides a bit more information than the latter theorem.

Solution: 1: For $x \in \mathrm{Conv}(X)$, let $x = \sum_{i\in I} \lambda_i x_i$ be the shortest – with the minimum possible cardinality of I – representation of x as a convex combination of points from X, and let us verify that the vectors x_i participating in this representation form an affinely independent collection. Assuming otherwise, there exists a nontrivial collection of reals δ_i, $i \in I$, such that $\sum_i \delta_i x_i = 0$ and $\sum_i \delta_i = 0$, and we can proceed exactly as in the proof of Caratheodory's Theorem: setting $\lambda_i(t) = \lambda_i + t\delta_i$, we have $\sum_i \lambda_i(t) = 1$ and $\sum_i \lambda_i(t) x_i = x$ for all t, and since not all δ_i are zeros and their sum is 0, some of $\lambda_i(t)$ for large t become negative, implying, due to $\lambda_i(0) \ge 0\,\forall i$, that for some t^* all $\lambda_i(t^*)$ are nonnegative, and some of them vanish, contradicting the assumption that number of terms in our initial representation of x as a convex combinations of points from X is the minimum possible.

2: Similarly, for $x \in \mathrm{Cone}(X)$, let $x = \sum_{i\in I} \lambda_i x_i$ be the representation of x as a conic combination of points from X with the minimum possible number of terms, and let us prove that the vectors x_i participating in this representation form a linearly independent collection. This indeed is so when $I = \varnothing$. Now let I be nonempty, and assume, for contradiction, that the vectors x_i, $i \in I$, are linearly dependent, so that $\sum_i \delta_i x_i = 0$ for a nontrivial collection $\delta_i, i \in I$. Passing, if necessary, from δ_i to $-\delta_i, i \in I$, we may assume that some of δ_i are strictly negative. Setting $\lambda_i(t) = \lambda_i + t\delta_i$, we have that $\sum_i \lambda_i(t) x_i = x$ for all t, $\lambda_i(0) \ge 0$, $i \in I$, and some of $\lambda_i(t)$ become negative for large t. It follows that there exists the largest $t = t^*$ for which all $\lambda_i(t)$ still are nonnegative, and for $t = t^*$ some of $\lambda_i(t)$ vanish, implying that $x = \sum_{i\in I} \lambda_i(t^*) x_i$ is a representation of x as a conic combination of x_i with some of the coefficients equal to 0, contradicting the minimality of the original representation. ■

Exercise I.18.[4] Consider TTD problem, and let N be the number of tentative bars, M be the dimension of the corresponding space of virtual displacements $\mathcal{V}$, and f be an external load. Prove that if truss $t \ge 0$ can withstand load f with compliance $\le \tau$ for some given real τ, then there exists truss $\bar{t}$ of the same total volume as t with compliance w.r.t. f at most τ and at most $M+1$ bars of positive volume.

Solution: Denoting by v the equilibrium displacement of (nodes of) truss t under load f, by the results of Exercise I.16.1 we have

$$\sum_{i=1}^{N} t_i \underbrace{[\mathfrak{b}_i \mathfrak{b}_i^\top v]}_{g_i} = f \;\&\; \frac{1}{2} f^\top v \le \tau$$

Denoting w the volume of t and assuming w.l.o.g. that $w > 0$, we see that f/w is a convex combination, with coefficients t_i/w, of vectors $g_i \in \mathbf{R}^M$. By Caratheodory Theorem, f/w is a convex combination of the same vectors g_i with coefficients s_i such that at most $M+1$ of these coefficients are positive. It follows that setting $\bar{t}_i = w s_i$, we get truss $\bar{t}$ of the same volume as t, with at most $M+1$ bars of positive

[4] Preceding exercise in the TTD series is I.16.

volume, such that $\sum_i \bar{t}_i \mathfrak{b}_i \mathfrak{b}_i^\top v = f$, so that v is the equilibrium displacement of truss $\bar{t}$ under load f. Consequently, the compliance of truss $\bar{t}$ w.r.t. load f is $\frac{1}{2} f^\top v \le \tau$. ■

Exercise I.19.

1. Prove that if a system of linear equations $Ax = b$ with n variables and m equations has a nonnegative solution, it has a nonnegative solution with at most m positive entries.

   Solution: Let $A_1, \ldots, A_n$ be the columns of A. Nonnegative solutions to the system $Ax = b$ are exactly the vectors of coefficients in representation of $b \in \mathbf{R}^m$ as a conic combination of $A_1, \ldots, A_n$. Then, by conic version of Caratheodory's Theorem (Fact I.2.7), if b admits such a representation, it admits such a representation with at most m of A_i's involved.

2. Let $V_1, \ldots, V_n$ be n nonempty sets in $\mathbf{R}^m$, and define

   $$\overline{V} := \mathrm{Conv}(V_1 + V_2 + \ldots + V_n).$$

   1. Prove that

      1. Taking direct product commutes with taking convex hull:

         $$\mathrm{Conv}(V_1 \times ... \times V_n) = \mathrm{Conv}(V_1) \times ... \times \mathrm{Conv}(V_n)$$

         Solution: Applying induction in n, it suffices to verify that for nonempty $U, V \subset \mathbf{R}^n$ it holds $\mathrm{Conv}(U \times V) = \mathrm{Conv}(U) \times \mathrm{Conv}(V)$. When $[u^i; v^i] \in U \times V$ and $\lambda_i \ge 0$, $\sum_i \lambda_i = 1$, we have $\sum_i \lambda_i [u^i; v^i] = [\sum_i \lambda_i u^i; \sum_i \lambda_i v^i] \in \mathrm{Conv}(U) \times \mathrm{Conv}(V)$, implying that $\mathrm{Conv}(U \times V) \subseteq \mathrm{Conv}(U) \times \mathrm{Conv}(V)$. Vice versa, if $[u; v] \in \mathrm{Conv}(U) \times \mathrm{Conv}(V)$, then $u = \sum_i \lambda_i u^i$ with $u^i \in U$, $\lambda_i \ge 0$, $\sum_i \lambda_i = 1$, and $v = \sum_j \mu_j v^j$ with $v^j \in V$, $\mu_j \ge 0$, $\sum_j \mu_j = 1$, whence $[u; v] = [\sum_{i,j} \lambda_i \mu_j u^i; \sum_{i,j} \lambda_i \mu_j v^j] = \sum_{i,j} \lambda_i \mu_j [u^i; v^j] \in \mathrm{Conv}(U \times V)$ due to $\lambda_i \mu_j \ge 0$ and $\sum_{i,j} \lambda_i \mu_j = 1$. Thus, $\mathrm{Conv}(U) \times \mathrm{Conv}(V) \subseteq \mathrm{Conv}(U \times V)$.

      2. Taking affine image commutes with taking convex hull: if $V \subset \mathbf{R}^n$ is nonempty and $x \mapsto \mathcal{A}(x) = Ax + b : \mathbf{R}^n \to \mathbf{R}^m$ is an affine mapping, then define $\mathcal{A}(V) := \{\mathcal{A}(x) : x \in V\}$ and show that

         $$\mathrm{Conv}\left(\mathcal{A}(V) := \{\mathcal{A}(x) : x \in V\}\right) = \mathcal{A}(\mathrm{Conv}(V))$$

         Solution: Evident.

      3. Conclude from the previous two items that taking weighted sum of sets commutes with taking convex hull:

         $$\mathrm{Conv}\left(\lambda_1 V_1 + ... + \lambda_n V_n := \{v = \sum_i \lambda_i v_i : v_i \in V_i, i \le n\}\right) = \lambda_1 \mathrm{Conv}(V_1) + ... + \lambda_n \mathrm{Conv}(V_n) \qquad [\lambda_i \in \mathbf{R}]$$

         In particular,

         $$\overline{V} = \mathrm{Conv}(V_1) + \ldots + \mathrm{Conv}(V_n).$$

         Solution: Note that $\lambda_1 V_1 + ... + \lambda_n V_n = \mathcal{A}(V_1 \times ... \times V_n)$, where $\mathcal{A}[v^1; ...; v^n] = \sum_i \lambda_i v^i$, and apply subsequently the first and the second of the preceding items.

         Note: The last three claims remain true when convex hull is substituted with affine span (but not when it is substituted with conic hull or linear span – compare $\mathrm{Cone}(\{1\}) \times \mathrm{Cone}(\{1\})$ and $\mathrm{Lin}(\{1\}) \times \mathrm{Lin}(\{1\})$ with $\mathrm{Cone}(\{1\} \times \{1\})$ and $\mathrm{Lin}(\{1\} \times \{1\})$, respectively.

   2. Prove *Shapley-Folkman Theorem*:

      Let $x \in \overline{V}$. Then, there exists a representation of x such that

      $$x = x_1 + \ldots + x_n, \quad x_i \in \mathrm{Conv}(V_i),$$

      in which at least $n - m$ of x_i's belong to the respective sets V_i.

   Comment: Shapley-Folkman Theorem says, informally, that when $n \gg m$, summing up n nonempty sets in $\mathbf{R}^m$ possesses certain "convexification property" – every point from the

convex hull $\overline{V}$ of the sum of our sets is the sum of points x_i with all but m of them belonging to V_i rather than to $\mathrm{Conv}(V_i)$, and only $\leq m$ of the points belonging to V_i "fractionally," that is, belonging to $\mathrm{Conv}(V_i)$, but not to V_i. This nice fact has numerous useful applications.

Solution: Let $x \in \overline{V}$. Then, by the previous part, $x = \sum_{i=1}^{n} \left(\sum_{k=1}^{K} \lambda_{i,k} x_{i,k} \right)$ with some K, $x_{i,k} \in V_i$, and $\lambda_{i,k} \geq 0$, $\sum_{k=1}^{K} \lambda_{i,k} = 1$, $i \leq n$. Hence, nK reals $\lambda_{i,k}$ form a nonnegative solution to the system of $m+n$ linear equations

$$(a) \quad \sum_k \lambda_{i,k} = 1, \quad 1 \leq i \leq n$$

$$(b) \quad \sum_i \sum_k \lambda_{i,k} x_{i,k} = x.$$

By the first item of this exercise, this system has a nonnegative solution with at most $m+n$ nonzero entries; let us denote this solution by $\{\overline{\lambda}_{i,j}\}$. We can partition the equations in (a) into two groups, the ones in which exactly one of the variables participating in the equation takes a positive value in the solution $\{\overline{\lambda}_{i,j}\}$, and the equations in which two or more variables participating in the equation take positive values in $\{\overline{\lambda}_{i,j}\}$. Let d be the number of equations in the latter set. Every one of the remaining $n-d$ equations in (a) involves at most one, and therefore exactly one, variable which at our solution gets positive value, and this positive values is, of course, 1. Since every one of the variables $\lambda_{i,k}$ enters exactly one of the equations (a), we conclude that the total number of positive $\overline{\lambda}_{i,k}$'s (which, as we remember, is at most $m+n$) is at least $2d + (n-d) = n+d$, implying that $d \leq m$. Thus, for at least $n-m$ values of i all but one of $\overline{\lambda}_{i,k}$'s, $k = 1, \ldots, K$, are zeros, and the remaining one equals to 1. In other words, all but at most m of the n sums $\sum_k \overline{\lambda}_{i,k} x_{i,k}$, $i = 1, \ldots, n$, are just points from the respective sets V_i. It remains to recall that $x = \sum_{i=1}^{n} \sum_k \overline{\lambda}_{i,k} x_{i,k}$.

Exercise I.20. Caratheodory's Theorem in its plain and its conic forms are "existence" statements: if a point $x \in \mathbf{R}^m$ is a convex, respectively conic, combination of points $x^1, \ldots, x^N$, then *there exists* a representation of x of the same type which involves at most $(m+1)$, respectively, m, terms. Extract from the proofs of the theorems *algorithms* for finding these "short" representations at the cost of solving at most N solvable systems of linear equations with at most N variables and m equations each.

Solution: For the sake of definiteness, consider plain Caratheodory Theorem (conic case can be treated in exactly the same fashion). Proof of Theorem, on immediate inspection, is based on the following observation:

> Given representation $x = \sum_{i=1}^{K} \lambda_i x^i$ with $\lambda_i > 0$, $i \leq K$ and $\sum_i \lambda_i = 1$, in the case of $K > m+1$ and finding a nontrivial solution δ to the homogeneous system of linear equations
>
> $$\sum_{i=1}^{K} \delta_i x^i = 0, \quad \sum_{i=1}^{K} \delta_i = 0.$$
>
> we can convert, by simple computation, the initial representation into a new one, of the same form, which assigns positive weights to at most $K' \leq K-1$ of x^i's.

Recall that we are given representation of x as convex combination of $K = N$ of x^i's. If $K > m+1$, we can apply the above construction to represent x as a convex combination of $K' < K$ of x^i's. If $K' > m+1$, we can iterate this update, with K' in the role of K, to represent x as convex combination of at most $K'' < K'$ of x^i's. Proceeding in this way, we in at most N steps will represent x as convex combination of at most $m+1$ of x^i's.

Exercise I.21. Prove *Kirchberger's Theorem*:

> Consider two sets of finitely many points $X = \{x^1, \ldots, x^k\}$ and $Y = \{y^1, \ldots, y^m\}$ in $\mathbf{R}^n$ such that $k + m \geq n+2$ and all the points $x^1, \ldots, x^k, y^1, \ldots, y^m$ are distinct. Assume that for any subset $S \subseteq X \cup Y$ which contains $n+2$ points the convex hulls of the sets $X \cap S$ and $Y \cap S$ do not intersect: $\mathrm{Conv}(X \cap S) \cap \mathrm{Conv}(Y \cap S) = \varnothing$. Then, the convex hulls of X and Y also do not intersect: $\mathrm{Conv}(X) \cap \mathrm{Conv}(Y) = \varnothing$.

Hint: Assume for contradiction that $\mathrm{Conv}(X) \cap \mathrm{Conv}(Y) \neq \varnothing$, so that

$$\sum_{i=1}^{k} \lambda_i x^i = \sum_{j=1}^{m} \mu_j y^j \qquad (*)$$

for certain nonnegative λ_i, $\sum_{i=1}^k \lambda_i = 1$, and certain nonnegative μ_j, $\sum_{j=1}^m \mu_j = 1$, and look at the expression of this type with the minimum possible total number of nonzero coefficients λ_i, μ_j.

Solution: Following the hint, assume for the contradiction that $\mathrm{Conv}(Y)$ and $\mathrm{Conv}(X)$ do intersect, so that the relation $(*)$ holds for appropriately chosen λ_i, μ_j satisfying

$$\lambda_i \geq 0, \quad \mu_j \geq 0, \quad \sum_i \lambda_i = \sum_j \mu_j = 1. \qquad (**)$$

And, among the collection of weights λ_i, μ_j satisfying $(*)$ and $(**)$, let us select one that has the smallest in the total number of positive λ_i, μ_j. Without loss of generality, we may assume that in this collection of weights, the positive weights are the first p of λ_i's and the first q of μ_j's. Note that by the premise of Kirchberger's Theorem, $p + q > n + 2$. Now consider the following system of $n + 2$ equations with $p + q > n + 2$ unknowns:

$$\begin{aligned} \sum_{i=1}^{p} \delta_i x^i - \sum_{j=1}^{q} \theta_j y^j &= 0, \\ \sum_i \delta_i &= 0, \\ \sum_j \theta_j &= 0. \end{aligned}$$

As this is a homogeneous system of linear equations and the number of unknowns is greater than the number of equations, the system has a nontrivial solution δ, θ. Setting $\lambda_i(t) = \lambda_i + t\delta_i$, $i \leq p$, and $\mu_j(t) = \mu_j + t\theta_j$, $j \leq q$, we have for all t:

$$\sum_i \lambda_i(t) x^i = \sum_j \mu_j(t) y^j, \quad \sum_i \lambda_i(t) = 1, \quad \sum_j \mu_j(t) = 1.$$

For $t = 0$, all the coefficients $\lambda_i(t)$, $\mu_j(t)$ are positive. Since $\sum_i \delta_i + \sum_j \theta_j = 0$ and not all δ_i, θ_j are zeros, among the reals δ_i, θ_j at least one should be negative.

Hence, for large enough $t > 0$ some of the coefficients $\lambda_i(t)$, $\mu_j(t)$ will be negative. Consequently, there exists the largest $t = t_*$ for which all $\lambda_i(t)$, $\mu_j(t)$ are nonnegative; among $\lambda_i(t_*)$, $\mu_j(t_*)$, there is clearly at least one zero, and we see that the coefficients $\lambda_i(t_*)$, $\mu_j(t_*)$ satisfy $(*)$, $(**)$, and the total number of positive among them is $< p + q$, which is a contradiction.

Exercise I.22 [Follow-up to Shapley-Folkman Theorem]

1. Let $X_1, \ldots, X_K$ be nonempty convex sets in $\mathbf{R}^n$, and define $X := \bigcup_{k \leq K} X_k$. Prove that

$$\mathrm{Conv}(X) = \left\{ x = \sum_{k=1}^{K} \lambda_k x^k : \ \lambda_k \geq 0,\, x^k \in X_k,\, \forall k \leq K,\, \sum_{k=1}^{K} \lambda_k = 1 \right\}.$$

Solution: Let $\overline{X}$ be the set on the right hand side. As $X = \bigcup_{k \leq K} X_k$, based on the definition of $\overline{X}$ it is clear that $\mathrm{Conv}(X) \supseteq \overline{X}$. So, all we need to show is that $\mathrm{Conv}(X) \subseteq \overline{X}$. To this end consider any $x \in \mathrm{Conv}(X)$, and so $x = \sum_{s=1}^{S} \mu_s y^s$ with $\mu_s \geq 0$, $y^s \in X$, $s \leq S$, and $\sum_s \mu_s = 1$. We can clearly split the index set $\{1, 2, \ldots, S\}$ into K non-overlapping subsets S_k, $k \leq K$ (some of these subsets can be empty) in such a way that $s \in S_k$ implies that $\mu_s > 0$ and $y^s \in X_k$. For k with nonempty S_k, let us set $\lambda_k := \sum_{s \in S_k} \mu_s$ and define $x^k := \sum_{s \in S_k} \frac{\mu_s}{\lambda_k} y^s$. By definition of λ_k and the fact that $y^s \in X_k$ for all $s \in S_k$, we see that x^k is a convex combination of points from X_k, and as X_k is a convex set we conclude that $x^k \in X_k$. For k with empty S_k, let us set $\lambda_k = 0$ and select somehow x^k in the (nonempty!) set

X_k. As a result, we get $x = \sum_{s=1}^S \mu_s y^s = \sum_{k=1}^K \lambda_k x^k$ with $x^k \in X_k$, $\lambda_k \geq 0$, and $\sum_{k=1}^K \lambda_k = 1$. This shows that $x \in \overline{X}$ as desired.

2. Let X_k, $k \leq K$, be nonempty bounded polyhedral sets in $\mathbf{R}^n$ given by polyhedral representations:

$$X_k = \left\{x \in \mathbf{R}^n : \ \exists u^k \in \mathbf{R}^{n_k} \text{ such that } P_k x + Q_k u^k \leq r_k\right\}.$$

Define $X := \bigcup_{k \leq K} X_k$. Prove that the set $\mathrm{Conv}(X)$ is a polyhedral set given by the polyhedral representation

$$\mathrm{Conv}(X) = \left\{ x \in \mathbf{R}^n : \begin{array}{ll} \exists x^k \in \mathbf{R}^n,\ u^k \in \mathbf{R}^{n_k},\ \lambda_k \in \mathbf{R},\ \forall k \leq K: & \\ \quad P_k x^k + Q_k u^k - \lambda_k r_k \leq 0,\ k \leq K & (a) \\ \quad \lambda_k \geq 0,\ \sum_{k=1}^K \lambda_k = 1 & (b) \\ \quad x = \sum_{k=1}^K x^k & (c) \end{array} \right\}. \qquad (*)$$

Does the claim remain true when the assumption of boundedness of the sets X_ks is lifted?

Solution: Let us temporary denote by $\widehat{X}$ the right hand side set in $(*)$ and set $\overline{X} := \mathrm{Conv}(X)$. We need to show that $\overline{X} = \widehat{X}$. Recall from item 1 that

$$\overline{X} = \left\{x = \sum_{k=1}^K \lambda_k x^k : \ \lambda_k \geq 0,\, x^k \in X_k,\, \forall k \leq K,\ \sum_{k=1}^K \lambda_k = 1\right\}.$$

Let $\widetilde{X}$ be the set of all vectors representable as convex combinations, *with positive coefficients*, of vectors from $X_1, \ldots, X_K$. Note that $\widetilde{X} \subseteq \overline{X}$.

Observe that

$$\widetilde{X} = \left\{ x \in \mathbf{R}^n : \begin{array}{ll} \exists x^k \in \mathbf{R}^n,\ u^k \in \mathbf{R}^{n_k},\ \lambda_k \in \mathbf{R},\ \forall k \leq K: & \\ \quad P_k x^k + Q_k u^k - \lambda_k r_k \leq 0,\ k \leq K & (a') \\ \quad \lambda > 0,\ \sum_{k=1}^K \lambda_k = 1 & (b') \\ \quad x = \sum_{k=1}^K x^k & (c') \end{array} \right\}. \qquad (!)$$

Indeed, when x belongs to the right hand side set in (!), we have $y^k := \lambda_k^{-1} x^k \in X_k$ due to $P_k y^k + Q_k[\lambda_k^{-1} u^k] \leq r_k$ and $x = \sum_k \lambda_k y^k$. Vice versa, when $x \in \widetilde{X}$, we have $x = \sum_k \lambda_k y^k$ with positive λ_k summing up to 1 and $y^k \in X_k$. The latter means that there exist v^k such that $P_k y^k + Q_k v^k \leq r_k$. Setting $x^k = \lambda_k y^k$, $u^k = \lambda_k v^k$, we ensure validity of (a') – (c'), so that x belongs to the right hand side set in (!).

Next, we claim that $\widehat{X} = \mathrm{cl}\, \overline{X}$. First, observe that $\widetilde{X}$ is dense in $\overline{X}$, meaning that every point $x \in \overline{X}$ is the limit of a sequence of points from $\widetilde{X}$. Indeed, consider any $x \in \overline{X}$, i.e., $x = \sum_k \lambda_k x^k$ with nonnegative λ_k summing up to 1 and $x^k \in X_k$ for all k. Then, we have $x = \lim_{i \to \infty} \sum_k \frac{\lambda_k + 1/i}{1 + K/i} x^k$, and the points in the right hand side sequence belong to $\widetilde{X}$. Now, observe that $\widehat{X}$ is closed (it is polyhedrally representable and thus polyhedral) and moreover $\widetilde{X}$ is dense in $\widehat{X}$. Indeed, by (!) we have $\widetilde{X} \subseteq \widehat{X}$. On the other hand, let us fix somehow $\overline{x}^k \in X_k$ and $\overline{\lambda}_k > 0$ such that $\sum_k \overline{\lambda}_k = 1$, and let $\overline{u}^k$ be such that $P_k \overline{x}^k + Q_k \overline{u}^k \leq r_k$. Given $x \in \widehat{X}$, there exist x^k, u^k and λ_k satisfying (a) – (c). For all $i = 1, 2, \ldots$, setting

$$\begin{aligned} x^{k,i} &:= (1 - 1/i) x^k + (1/i) \overline{x}^k, \\ u^{k,i} &:= (1 - 1/i) u^k + (1/i) \overline{u}^k, \\ \lambda_{k,i} &:= (1 - 1/i) \lambda_k + (1/i) \overline{\lambda}_k, \end{aligned}$$

we ensure that $P_k x^{k,i} + Q_k u^{k,i} - \lambda_{k,i} r_k \leq 0$, $\lambda_{k,i} > 0$, $\sum_i \lambda_{k,i} = 1$, implying that $x^{(i)} := \sum_k x^{k,i} \in \widetilde{X}$. As $i \to \infty$, we clearly have $x^{(i)} \to x$, so that $\widetilde{X}$ indeed is dense in $\widehat{X}$. The latter combines with closedness of $\widehat{X}$ to imply that the $\widehat{X}$ is the closure of $\widetilde{X}$, and the latter set, due to the fact that $\widetilde{X}$ is dense in $\overline{X}$, is the same as the closure of $\overline{X}$. Thus, $\widehat{X} = \mathrm{cl}\, \overline{X}$.

It remains to note that since X_k are bounded, $\overline{X}$ is closed. This is immediate: assuming that $x = \lim_{i \to \infty} \sum_k \lambda_{k,i} x^{k,i}$ with nonnegative $\lambda_{k,i}$, $\sum_k \lambda_{k,i} = 1$, and $x^{k,i} \in X_k$, boundedness of X_k, $k \leq K$, allows to find a subsequence $i_1 < i_2 < \ldots$ of indexes such that for some λ_k and x^k, $k \leq K$, it holds

$\lambda_{k,i_s} \to \lambda_k$ and $x^{k,i_s} \to x^k$ for every k as $s \to \infty$. Since X_k are polyhedral and thus closed, we have $x^k \in X_k$, and of course $\lambda_k \geq 0$, $\sum_k \lambda_k = 1$, that is, $x = \lim_{s\to\infty} \sum_k \lambda_{k,i_s} x^{k,i_s} = \sum_k \lambda_k x^k \in \overline{X}$.
Finally, $\mathrm{Conv}(\bigcup_{k\leq K} X_k)$ is not necessarily polyhedral when X_k are nonempty polyhedral, but unbounded, sets. For example, by selecting $K = n = 2$, $X_1 := \{x \in \mathbf{R}^2 : x_1 \geq 0,\ x_2 = 0\}$ and $X_2 := \{[0;1]\}$, we see that the set $\mathrm{Conv}(X_1 \cup X_2)$ is not polyhedral, but its closure is. On inspection, the above reasoning demonstrates that when X_k are nonempty polyhedral sets given by polyhedral representations, then the polyhedral set $\widehat{X}$ defined as the right hand side set of $(*)$ is the closure of $\mathrm{Conv}(\bigcup_{k\leq K} X_k)$.

After two preliminary items above, let us pass to the essence of the matter. Consider the situation as follows. We are given n nonempty and bounded polyhedral sets $X_j \subset \mathbf{R}^r$, $j = 1, \ldots, n$. We will think of X_j as the "resource set" of the j-th production unit: entries in $x \in X_j$ are amounts of various resources, and X_j describes the set of vectors of resources available, in principle, for j-th unit. Each production unit j can possibly use any one of its $K_j < \infty$ different production plans. For each $j = 1, \ldots, n$, the vector $y_j \in \mathbf{R}^p$ representing the production of the j-th unit depends on the vector x_j of resources consumed by the unit and also on the production plan utilized in the unit. In particular, the production vector $y_j \in \mathbf{R}^p$ stemming from resources x_j under k-th plan can be picked by us, at our will, from the set

$$Y_j^k[x_j] := \left\{ y_j \in \mathbf{R}^p :\ z_j := [x_j; -y_j] \in V_j^k \right\},$$

where V_j^k, $k \leq K_j$, are given bounded polyhedral "technological sets" of the units with projections onto the x_j-plane equal to X_j, so that for every $k \leq K_j$ it holds

$$x_j \in X_j \quad \Longleftrightarrow \quad \exists y_j \text{ such that } [x_j; -y_j] \in V_j^k. \tag{5.3}$$

We assume that all the sets V_j^k are given by polyhedral representations, and we define

$$V_j := \bigcup_{k\leq K_j} V_j^k.$$

Let $R \in \mathbf{R}^r$ be the vector of total resources available to all n units and let $P \in \mathbf{R}^p$ be the vector of total demands for the products. For $j \leq n$, we want to select $x_j \in X_j$, $k_j \leq K_j$, and $y_j \in Y_j^{k_j}[x_j]$ in such a way that

$$\sum_j x_j \leq R \quad \text{and} \quad \sum_j y_j \geq P.$$

That is, we would like to find $z_j = [x_j; v_j] \in V_j$, $j \leq n$, in such a way that $\sum_j z_j \leq [R; -P]$. Note that the presence of "combinatorial part" in our decision – selection of production plans in finite sets – makes the problem difficult.

3. Apply Shapley-Folkman Theorem (Exercise I.19) to overcome, to some extent, the above difficulty and come up with a good and approximately feasible solution.

Solution: Let $s := [R; -P]$, and observe that our problem reads

$$\text{Find } z_j \in V_j \text{ such that } \sum_{j=1}^n z_j \leq s. \tag{P}$$

Note that given polyhedral representations of V_j^k, based on item 2, we can build explicit polyhedral representations of the convex hulls of the sets V_j, i.e., we can efficiently compute $\overline{V}_j$, where

$$\overline{V}_j := \mathrm{Conv}(V_j).$$

Let us relax the problem of interest (P) to the problem

$$\text{Find } z_j \in \overline{V}_j \text{ such that } \sum_{j=1}^n z_j \leq s. \tag{$\overline{P}$}$$

By calculus of polyhedral representations, $(\overline{P})$ is the problem of the form

Given polyhedral representation of nonempty polyhedral set $Z \subset \mathbf{R}^{r+p}$ and vector $s \in \mathbf{R}^{r+p}$, find $z \in Z$ such that $z \le s$.

Note that is an explicit Linear Programming feasibility problem. Thus, we can apply LP algorithms to check whether $(\overline{P})$ is solvable, and if it is the case – find a solution $\{z_j, j \le n\}$ to $(\overline{P})$. Applying Shapley-Folkman Theorem, we can convert, in a computationally efficient fashion, this solution into another feasible solution, $\{[x_j; v_j], j \le n\}$, for which for all but at most

$$d := \min\{r+p,\, n\}$$

components $[x_j; v_j]$ belong to V_j, that is, "are implementable" – for the corresponding j, one has $x_j \in X_j$ and $y_j = -v_j \in Y_j^{k_j}[x_j]$ with properly selected $k_j \le K_j$. Let J be the set of "bad" indices j, i.e., those for which $[x_j; v_j] \in \overline{V}_j \setminus V_j$. Note that for each $j \in J$ we still have $x_j \in X_j$. We can correct the corresponding y_j, passing from $[x_j; v_j]$ to $[x_j; \overline{v}_j]$ with $\overline{v}_j \in -Y_j^1[x_j]$, or, better, $\overline{v}_j$ defined as the optimal solution to the "best" – with the smallest optimal value – among the K_j convex optimization problems

$$\min_{u_k} \left\{ \|v_j - u_k\| :\ [x_j; u_k] \in V_j^k \right\}, \quad k \le K_j,$$

where $\|\cdot\|$ is some norm. As a result, we get "fully implementable" solution $\{[x_j; \overline{v}_j], j \le n\}$, where $\overline{v}_j = v_j$ for $j \notin J$, to problem (P). This solution, in general, may not be feasible when $J \neq \varnothing$. However, by selecting somehow norm $\|\cdot\|$, defining

$$D_j := \max_{x,v,x'v'} \left\{ \|v - v'\| :\ [x; v] \in V_j,\ [x'; v'] \in V_j \right\},\ \forall j \le n \quad \text{and} \quad D := \max_j D_j,$$

and taking into account that $\mathrm{Card}(J) \le d = \min\{r+p, n\}$, we have $\sum_j [x_j; \overline{v}_j] \le s + \delta$, $\|\delta\| \le dD$, and $\sum_j x_j \le R$. In the case of "mass production", when $\|P\|$ is large, the violation of the constraint $\sum_j \overline{v}_j \le -P$ as quantified by $\|\delta\|$ is a small fraction of the magnitude of P, and our implementable solution has chances to be a good, from a "practical perspective," surrogate of a feasible solution to (P).

5.5 Around Helly Theorem

Exercise I.23. [Alternative proof of Helly Theorem] The goal of this exercise is to build an alternative proof of Helly's Theorem, without using Radon's Theorem.

1. Consider a system $a_i^\top x \le b_i$, $i \le N$, of N linear inequalities in variables $x \in \mathbf{R}^n$. Helly's Theorem applied to the sets $A_i := \{x \in \mathbf{R}^n :\ a_i^\top x \le b_i\}$ gives us that

   (!) *If a system $a_i^\top x \le b_i$, $i \le N$, of linear inequalities in variables $x \in \mathbf{R}^n$ is infeasible, so is a properly selected sub-system composed of at most $n+1$ inequalities from the system.*

   Find an alternative proof of (!) without relying on Helly's or Radon's Theorems.

   Solution: Suppose that the system $a_i^\top x \le b_i$, $i \le N$, is infeasible. Then, by General Theorem of the Alternative there exist nonnegative weights λ_i and $\alpha < 0$ such that the vector $[0; \ldots; 0; \alpha]$ is a conic combination, with coefficients λ_i, of vectors $[a_i; b_i]$, $i \le N$. Note that $[a_i; b_i] \in \mathbf{R}^{n+1}$, and so by Caratheodory's Theorem in conic form it follows that the vector $[0; \ldots; 0; \alpha]$ is a conic combination of at most $n+1$ of the vectors $[a_i; b_i]$, let I be the set of their indexes. By GTA, the subsystem $a_i^\top x \le b_i$, $i \in I$, of the original system is infeasible, and the number of inequalities in it is at most $n+1$, as desired.

2. Extract from item 1 Helly's Theorem for polyhedral sets: *If $A_1, \ldots, A_N$, $N \ge n+1$, are polyhedral sets in $\mathbf{R}^n$ and every $n+1$ of these sets have a point in common, then all the sets have a point in common.*

   Solution: Let $A_i := \{x \in \mathbf{R}^n :\ P_i x \le p_i\}$ for $i \le N$. To justify the claim in question is the same as to prove that if $\cap_i A_i = \varnothing$, then the intersection of properly selected $k \le n+1$ sets from the collection is empty. Suppose that $\cap_i A_i = \varnothing$, that is, the system

   $$P_i x \le p_i, i \le N \qquad (*)$$

of linear inequalities in variables $x \in \mathbf{R}^n$ has no solutions. By item 1, we can select from $(*)$ $k \le n+1$ inequalities to get an infeasible subsystem of $(*)$. Denoting by I the set of indices i of the blocks $P_i x \le p_i$ of $(*)$ containing the k selected inequalities, we conclude that $k \le n+1$ sets A_i, $i \in I$, have no point in common.

3. Extract from item 2 Helly's Theorem (Theorem I.2.10).

   Solution: Let $A_1, \ldots, A_N$ be a collection of convex sets in $\mathbf{R}^n$, $N \ge n+1$, such that every $n+1$ sets from the collection have a point in common, and let us prove that all sets have a point in common. For a collection $\iota = \{\iota_1 < \iota_2 < \ldots < \iota_{n+1}\}$ of $n+1$ distinct from each other indices from $\{1, 2, \ldots, N\}$, let x_ι be a point from the (nonempty!) set $A_{\iota_1} \cap A_{\iota_2} \cap \ldots \cap A_{\iota_{n+1}}$, and let $\overline{A}_j := \mathrm{Conv}(\{x_\iota : j \in \iota\})$. Note that $\overline{A}_j$ is the convex hull of points from A_j (since $x_\iota \in A_j$ whenever $j \in \iota$), and thus $\overline{A}_j \subset A_j$ (as A_j is convex). The sets $\overline{A}_1, \ldots, \overline{A}_N$ are convex hulls of finite sets and as such are polyhedrally representable and therefore polyhedral. Every $n+1$ sets $\overline{A}_{\iota_1}, \overline{A}_{\iota_2}, \ldots, \overline{A}_{\iota_{n+1}}$, $\iota_1 < \iota_2 < \ldots < \iota_{n+1} \le N$, have a point in common, namely, $x_{\iota_1,\ldots,\iota_{n+1}}$. Then, by item 2, all the sets $\overline{A}_j$, $j \le N$, have a point in common, and this point is a common point of $A_1, \ldots, A_N$, since, as we already know, $\overline{A}_j \subset A_j$.

Exercise I.24. $A_0, A_1, \ldots, A_m$, $m = 2025$, are nonempty convex subsets of $\mathbf{R}^{2000}$, and A_0 is a triangle (convex hull of 3 affinely independent vectors). Which of the claims below are always (that is, for any $A_0, \ldots, A_m$ satisfying the above assumptions) true:

1. If every 3 among the sets $A_0, \ldots, A_m$ have a point in common, all $m+1$ sets have a point in common.
2. If every 4 among the sets $A_0, \ldots, A_m$ have a point in common, all $m+1$ sets have a point in common.
3. If every 2001 among the sets $A_0, \ldots, A_m$ have a point in common, all $m+1$ sets have a point in common.

Solution: The true statements are the second and the third ones. To see that the second statement is true, let us define $\overline{A}_i := A_i \cap A_0$. Then, we get $m+1$ convex sets such that every 3 of them intersect (since the intersection of a triple of $\overline{A}$-sets is the same as intersection of four of A-sets) and *all of them belong to the affine plane* Π *of dimension 2* (namely, the affine span of the triangle A_0). Applying Helly's theorem to the sets $\overline{A}_i$ (treated as the subsets of the 2-dimensional affine plane), we conclude that all of them have a point in common, and this point, of course, is a common point of $A_0, \ldots, A_m$. Since the second statement is true, so is the third (the third statement is true even without assumption that A_0 is a triangle).

To see that the first statement can be incorrect, consider the following 4 sets in $\mathbf{R}^3$: B_0 is a triangle in the plane $L := \{x \in \mathbf{R}^3 : x_3 = 0\}$, and $B_i := \{x \in \mathbf{R}^3 : [x_1, x_2] \in B_0,\ x_3 = \frac{1}{2} - \lambda_i(x_1, x_2)\}$, $1 \le i \le 3$, where $\lambda_i(x_1, x_2)$ are the barycentric coordinates of $[x_1; x_2] \in L$, that, is coefficients in the representation of $[x_1; x_2]$ as the linear combination of the 3 vertices of B_0 with sum of coefficients equal to 1. Note that we have $B_0 := \{[x_1; x_2; x_3] : \lambda_i(x_1, x_2) \ge 0,\ i \le 3,\ x_3 = 0\}$. Let us check that every 3 of our 4 sets B_0, B_1, B_2, B_3 have a point in common. Indeed, if the triple of sets in question does not contain B_0, the common point is $[\bar{x}_1; \bar{x}_2; \bar{x}_3]$, where $[\bar{x}_1; \bar{x}_2]$ is the barycenter of B_0 (the average of its vertices), so that $\lambda_i(\bar{x}_1, \bar{x}_2) = 1/3$, $1 \le i \le 3$, and $\bar{x}_3 = \frac{1}{2} - \frac{1}{3}$. Now let us verify that if triple of our sets includes B_0, the sets from the triple still have a point in common. By symmetry, it suffices to check this for the triple B_0, B_1, B_2, for which the common point is $[\tilde{x}_1; \tilde{x}_2; 0]$, with $\lambda_1([\tilde{x}_1; \tilde{x}_2]) = \lambda_2([\tilde{x}_1; \tilde{x}_2]) = \frac{1}{2}$, $\lambda_3([\tilde{x}_1; \tilde{x}_2]) = 0$ (that is, $[\tilde{x}_1; \tilde{x}_2]$ is the midpoint of a properly selected side of the triangle B_0). Thus, every 3 of the four sets B_0, B_1, B_2, B_3 have a point in common, while all four sets have no such a point: indeed, such a point x should have $x_3 = 0$ (since it belongs to B_0) and therefore $\frac{1}{2} - \lambda_i(x_1, x_2) = 0$, $i = 1, 2, 3$ (since this point belongs to B_1, B_2, B_3). Therefore, every 3 of the barycentric coordinates of $[x_1; x_2]$ should be equal to $1/2$, which is impossible, since their sum must be 1.

To show that the first statement is not always true, it suffices to place our 3D sets B_0, B_1, B_2, B_3 into 2000-dimensional space by augmenting 3 entries in point from R^3 by 1997 zero entries; as a result, we get 4 convex sets A_0, A_1, A_2, A_3 in $\mathbf{R}^{2000}$ such that the first of them is triangle, every 3 of the sets have a point in common, but all 4 sets have no such a point. Augmenting the 4 sets A_i we have built by 2021

copies of one of them, say, A_0, we get a family of 2025 convex sets in $\mathbf{R}^{2000}$ such that every 3 of them have a point in common, but the intersection of all sets is empty, which is a counterexample for the first statement.

Exercise I.25. Let $P_i := \{x \in \mathbf{R}^n : \ A_i x \leq b_i\}$ for $i \in \{1, \ldots, m\}$ and $C := \{x \in \mathbf{R}^n : \ Dx \geq d\}$ be nonempty polyhedral sets. Suppose that for any $n+1$ sets, $P_{i_1}, \ldots, P_{i_{n+1}}$, there is a translate of C, i.e., the set $C + u$ for some $u \in \mathbf{R}^n$, which is contained in all $P_{i_1}, \ldots, P_{i_{n+1}}$. Prove that there is a translate of C, which is contained in all of the sets $P_1, \ldots, P_m$.

Solution: For every $i = 1, \ldots, m$, we define the set $C_i := \{u \in \mathbf{R}^n : P_i \supseteq C + u\}$. Note that C_i is a convex set for every i. Indeed, if $u + c \in P_i$ and $v + c \in P_i$ for all $c \in C$, then for $\lambda \in [0, 1]$ and $c \in C$ one has $[\lambda u + (1-\lambda)v] + c = \lambda[u + c] + (1-\lambda)[v + c] \in P_i$ by convexity of P_i, implying that $\lambda u + (1-\lambda)v \in C_i$. From the statement of the problem we know that every $n + 1$ sets C_i have a non-empty intersection. From Helly's Theorem, we deduce that all of them have a non-empty intersection. In other words, there is a $u \in \mathbf{R}^n$ such that $P_i \supseteq C + u$ for every $i \in \{1, \ldots, m\}$.

Exercise I.26. A cake contains 300 g of raisins (you may think of every one of them as of a 3D ball of positive radius). John and Jill are about to divide the cake according to the following rules:

- first, Jill chooses a point a in the cake;
- second, John makes a *cut* through a, that is, chooses a 2D plane Π passing through a and takes the part of the cake on one side of the plane (both Π and the side are up to John, with the only restriction that the plane should pass through a); all the rest goes to Jill.

1. Prove that it may happen that Jill cannot guarantee herself 76 g of the raisins.

   Solution: Suppose there are 4 raisins, 75 g each, placed in the vertices of large tetrahedron; whatever point Jill chooses, John can cut off 3 of the four raisins.

2. Prove that Jill always can choose a in a way which guarantees her at least 74 g of the raisins.
3. Consider n-dimensional version of the problem, where the raisins are n-dimensional balls, the cake is a domain in $\mathbf{R}^n$, and "a cut" taken by John is defined as the part of the cake contained in the half-space

   $$\left\{x \in \mathbf{R}^n : \ e^\top (x - a) \geq 0\right\},$$

   where $e \neq 0$ is the vector ("inner normal to the cutting hyperplane") chosen by John. Prove that for every $\epsilon > 0$, Jill can guarantee to herself at least $\frac{300}{n+1} - \epsilon$ g of raisins, but in general cannot guarantee to herself $\frac{300}{n+1} + \epsilon$ g.

Solution:
(2-3): Let us consider the case of $n = 3$ (generalization to arbitrary n will be evident).
For every direction (that is, unit vector) $d \in \mathbf{R}^3$ consider the closed half-spaces

$$\left\{x \in \mathbf{R}^3 : \ d^\top x \leq \alpha\right\},$$

and let us look at the mass of raisins outside of such a half-space. This mass is clearly a continuous function of α (since the distribution of raisins' mass has density) which is close to 300 when α is very negative and close to 0 when α is very positive. It follows that there exists the largest $\alpha = \alpha(d)$ such that the mass of the raisins outside the half-space

$$H_d := \left\{x \in \mathbf{R}^3 : \ d^\top x \leq \alpha(d)\right\}$$

is exactly 74 g. Note that

> (!) *If John takes himself the part of the cake in the half-space* $\{x \in \mathbf{R}^3 : \ d^\top x \leq d^\top \bar{x}\}$ *with* $\bar{x} \in H_d$ (that is, d is exactly the outer normal to the cut chosen by John, and this cut passes through $\bar{x}$), *then Jill gets at least 74 g of raisins.*

In view of (!), it suffices to prove that the intersection of all sets H_d is nonempty. Indeed, in this case Jill can choose, as the point through which the cut should pass, a point in $\bigcap_d H_d$; then whatever John will do, his cut will be as explained in (!) with certain d, and therefore Jill will get at least 74 g of raisins.

To prove that $\bigcap_d H_d$ is nonempty, we can use Helly Theorem II. Let us check its assumptions: The sets H_d indeed are closed and convex sets in $\mathbf{R}^3$. The intersection of every 4 sets H_d indeed is nonempty, since, assuming the opposite, the complements of the 4 sets with empty intersection would together cover the entire space; but every one of these complements contains 74 g of raisins, and therefore the union of 4 of them can contain at most $4 \cdot 74 = 296$ g of raisins, while the entire space contains 300 g of raisins. It remains to verify that one can choose among the sets H_d finitely many sets with bounded intersection. This is evident, since the intersection of H_{e_i} and H_{-e_i} (e_1, e_2, e_3 are basic orth) is a stripe $a_i \le x_i \le b_i$ with finite a_i, b_i, so that the intersection of the 6 sets H_{e_i}, H_{-e_i}, $i = 1, 2, 3$, is a bounded box. Generalization to the n-dimensional case is evident.

Remarks:

1. With some minor effort, you can prove that Jill can find a point which guarantees her $\frac{300}{n+1}$ g of raisins, and not $\frac{300}{n+1} - \epsilon$ g.
2. If, instead of dividing raisins, John and Jill would divide in the same fashion *uniform and convex* cake (that is, a closed and bounded convex body X with a nonempty interior in $\mathbf{R}^n$, the reward being the n-dimensional volume of the part a person gets), the results would change dramatically: choosing as the point the center of masses of the cake

$$\bar{x} := \frac{\int\limits_X x dx}{\int\limits_X dx},$$

Jill would guarantee herself at least $\left(\frac{n}{n+1}\right)^n \approx \frac{1}{e}$ part of the cake. This is a not so easy corollary of the following extremely important and deep result:

> **Brunn-Minkowski Symmetrization Theorem:** *Let X be as above, and let $[a, b]$ be the projection of X on an axis ℓ, say, on the last coordinate axis. Consider the "symmetrization" Y of X, i.e., Y is the set with the same projection $[a, b]$ on ℓ and for every hyperplane orthogonal to the axis ℓ and crossing $[a, b]$, the intersection of Y with this hyperplane is an $(n-1)$-dimensional ball centered at the axis with precisely the same $(n-1)$-dimensional volume as the one of the intersection of X with the same hyperplane:*
>
> $$\left\{z \in \mathbf{R}^{n-1} : \ [z; c] \in Y\right\} = \left\{z \in \mathbf{R}^{n-1} : \ \|z\|_2 \le \rho(c)\right\}, \quad \forall c \in [a, b], \text{ and}$$
> $$\mathrm{Vol}_{n-1}\left(\left\{z \in \mathbf{R}^{n-1} : \ [z; c] \in Y\right\}\right) = \mathrm{Vol}_{n-1}\left(\left\{z \in \mathbf{R}^{n-1} : \ [z; c] \in X\right\}\right), \quad \forall c \in [a, b].$$
>
> *Then, Y is a closed convex set.*

5.6 Around Polyhedral Representations

Exercise I.27. Justify the calculus rules for polyhedral representations presented in Section 3.3.

Solution: This is straightforward.

Exercise I.28. Given two sets $U, V \subseteq \mathbf{R}^m$, we define

$$U + V = \{x \in \mathbf{R}^m : \exists u \in U, \exists v \in V \text{ such that } x = u + v\}.$$

Let $D := \{x \in \mathbf{R}^n : \ Ax + b + Q_s \subseteq P, \ \forall s \in S\}$ where the nonempty set $P \subset \mathbf{R}^m$ admits polyhedral representation, the nonempty set $S \subset \mathbf{R}^k$ is given but arbitrary, and the nonempty sets $Q_s \subset \mathbf{R}^m$ are indexed by $s \in S$.

1. Suppose that S is a finite set and for each $s \in S$ we have $Q_s = \{q_s\}$, i.e., is a single point. Then, will the set D be polyhedrally representable?
2. State sufficient conditions on the structure of sets Q_s and S that will guarantee that the resulting set D is polyhedral. Here, the goal is to have conditions as general as possible. Among your sufficient conditions, can you identify at least some of those that are necessary?

Solution:

1. This part follows immediately from the next one.
2. Note that $x \in D$ if and only if $Ax + b + q \in P$ holds for all $q \in \bigcup_{s\in S} Q_s$. Since P is polyhedrally representable, let $P = \{y \in \mathbf{R}^m : Gy \le g\}$. Then, $x \in D$ if and only if $G(Ax + b + q) \le g$ for all $q \in \bigcup_{s\in S} Q_s$, i.e., if and only if x satisfies

$$[GA]_i^\top x \le \inf_q \left\{ [(g - G(b + q))]_i : \ q \in \bigcup_{s\in S} Q_s \right\}, \qquad \text{for all rows } i.$$

   Clearly this is a polyhedral representation of D without making any assumptions on the structure of S or Q_s.

Exercise I.29. For $x \in \mathbf{R}^n$ and integer k, $1 \le k \le n$, let $s_k(x)$ be the sum of k largest entries in x. For example, $s_1(x) = \max_i\{x_i\}$, $s_n(x) = \sum_{i=1}^n x_i$, $s_3([3; 1; 2; 2]) = 3 + 2 + 2 = 7$. Now let $1 \le k \le n$ be two integers. For any integer $k = 1, \ldots, n$, define

$$X_{k,n} := \{[x; t] \in \mathbf{R}^n \times \mathbf{R} : \ s_k(x) \le t\}.$$

Observe that $X_{k,n}$ is a polyhedral set. Indeed, $s_k(x) \le t$ holds if and only if for every k indices $i_1 < i_2 < \ldots < i_k$ from $\{1, 2, \ldots, n\}$ we have $x_{i_1} + x_{i_2} + \ldots + x_{i_k} \le t$, which is nothing but a linear inequality in variables x, t. Since there are $\binom{n}{k}$ possible ways of selecting k indices from $\{1, 2, \ldots, n\}$, the number of linear inequalities describing $X_{k,n}$ is $\binom{n}{k}$, and these linear inequalities give the polyhedral description of $X_{k,n}$. The point of this exercise is to demonstrate that $X_{k,n}$ admits a "short" polyhedral representation, specifically,

$$X_{k,n} = \left\{ [x; t] \in \mathbf{R}^n \times \mathbf{R} : \ \exists z \in \mathbf{R}^n, \exists s \in \mathbf{R} \text{ s.t. } x_i \le z_i + s, \forall i, \ z \ge 0, \ \sum_{i=1}^n z_i + ks \le t \right\}. \quad (*)$$

Solution: Let $\overline{X}_{k,n}$ be the right hand side set in $(*)$, we will prove that (a) $X_{k,n} \subseteq \overline{X}_{k,n}$ and (b) $\overline{X}_{k,n} \subseteq X_{k,n}$.

(a): Observe first of all that both $X_{k,n}$ and $\overline{X}_{k,n}$ are "permutationally symmetric in x", meaning that when $[x; t] \in X_{k,n}$ and $\bar{x}$ is obtained from x by permuting entries, we have $[\bar{x}; t] \in X_{k,n}$, and similarly $[x; t] \in \overline{X}_{k,n}$ implies $[\bar{x}; t] \in \overline{X}_{k,n}$. It follows that in order to verify (a) it suffices to prove that if $[x; t] \in X_{k,n}$ and $x_1 \ge x_2 \ge \ldots \ge x_n$, then $[x; t] \in \overline{X}_{k,n}$. This is immediate: set $z_i := x_i - x_k$ for $i \le k$ and $z_i := 0$ for $i > k$, and $s := z_i$. Taking into account that the entries in x form a non-ascending sequence, we immediately see that $z \ge 0$, $x_i \le z_i + s$ for all i, and $s_k(x) = \sum_{i=1}^k x_i = \sum_{i=1}^k z_i + ks = \sum_{i=1}^n z_i + ks$. Recalling that $[x; t] \in X_{k,n}$, that is, $s_k(x) \le t$, we conclude that x, t, z, s satisfy all inequalities participating in the description of $\overline{X}_{k,n}$, that is, $[x; t] \in \overline{X}_{k,n}$. (a) is proved.

(b): let x, t, z, s satisfy all inequalities in the description of $\overline{X}_{k,n}$. When $i_1 < i_2 < \ldots < i_k$ is an ordered collection of k indices from $\{1, \ldots, n\}$, we have by the inequalities describing $\overline{X}_{k,n}$ that

$$x_{i_1} + x_{i_2} + \ldots + x_{i_k} \le ks + z_{i_1} + z_{i_2} + \ldots + z_{i_k} \le ks + \sum_{i=1}^n z_i \le t,$$

where the second inequality is due to $z \ge 0$. The resulting inequality holds true for all ordered collections of k indices $i_1, \ldots, i_k$, implying that $s_k(x) \le t$. Thus, $[x; t] \in \overline{X}_{l,n}$ implies $[x; t] \in X_{k,n}$, as claimed in (b).

Exercise I.30. [Computational study: Fourier-Motzkin elimination as an LP algorithm] It was mentioned in section 3.2.1 that Fourier-Motzkin elimination provides us with an algorithm for solving LP problems that terminates in finitely many steps. This algorithm, however, is of no computational value due to the potential rapid growth of the number of inequalities one may need to handle when eliminating more and more variables. The goal of this exercise is to get an impression of this phenomenon.

Our "guinea pig" will be transportation problem with n unit capacity suppliers and n unit demand customers:

$$\min_{x,t}\left\{t:\ t\geq\sum_{i=1}^{n}\sum_{i=1}^{n}c_{ij}x_{ij},\ \sum_{i}x_{ij}\geq 1,\ \forall j,\ \sum_{j}x_{ij}\leq 1,\ \forall i,\ x_{ij}\geq 0,\forall i,j\right\}.$$

This problem has n^2+1 variables and $(n+1)^2$ linear inequality constraints, and let us solve it by applying the Fourier-Motzkin elimination to project the feasible set of the problem onto the axis of the t-variable, that is, to build a finite system $\mathcal{S}$ of univariate linear inequalities specifying this projection.

How many inequalities do you think there will be in $\mathcal{S}$ when $n=1,2,3,4$? Check your intuition by implementing and running the F-M elimination, assuming, for the sake of definiteness, that $c_{ij}=1$ for all i,j.

Solution: Our results are as follows (your numbers could be different, since the outcome depends on the serial numbers assigned to the x-variables):

n	$m_{\rm ini}$	$m_{\rm fin}$
1	4	4
2	9	19
3	16	44,854
4	25	†

$m_{\rm ini}$ and $m_{\rm fin}$ are the number of inequalities in the initial and the final systems

†When $n=4$, we are supposed to eliminate $n^2=16$ of 17 variables in a system with 25 linear inequality constraints on these 17 variables. Eliminating the last 11 variables results in system of 974,236 constraints with 6 variables. Eliminating in this system the last – the sixth – of the variables would result in a system of 121,226,850 linear inequalities with 5 variables; building this system was terminated after the number of assembled so far inequalities reached our a priori limit $2^{23}=8,388,608$.

5.7 Around General Theorem on Alternative

Exercise I.31.

1. Prove Gordan's Theorem on Alternative:

   A system of strict homogeneous linear inequalities $Ax<0$ in variables x has a solution if and only if the system $A^T\lambda=0,\ \lambda\geq 0$ in variables λ has only the trivial solution $\lambda=0$.

   Solution: By GTA, the system has no solutions if and only if by using nonnegative aggregations with weights λ of the inequalities in the system we can derive a contradictory consequence inequality, i.e., $[A^T\lambda]^T x\,\Omega\,0$, where $\Omega=$ " $<$ " when $\lambda\neq 0$ and $\Omega=$ " $\leq$ " when $\lambda=0$. Note that the only two possible contradictory linear inequalities are of the form either $0^Tx<\epsilon$ with $\epsilon\leq 0$ or $0^Tx\leq\epsilon'$ with $\epsilon'<0$. In our case, when $\lambda=0$ and so $\Omega=$ " $\leq$ ", the right-hand side of the consequence inequality will be zero, so the second option cannot lead to any contradictory inequality. Thus, we deduce that when given the system $Ax<0$, we can derive a contradictory inequality if and only if $\Omega=$ " $<$ " and $A^T\lambda=0$ for some nonzero $\lambda\geq 0$. Thus, $Ax<0$ has no solutions if and only if there exists a nonzero vector $\lambda\geq 0$ such that $A^T\lambda=0$.

2. Prove Motzkin's Theorem on Alternative:

   A system $Ax<0$, $Bx\leq 0$ of strict and nonstrict homogeneous linear inequalities has a solution if and only if the system $A^T\lambda+B^T\mu=0,\ \lambda\geq 0,\ \mu\geq 0$ in variables λ,μ has no solution with $\lambda\neq 0$.

Solution: Same as above, the infeasibility of the system is equivalent to the existence of nonnegative weights λ, μ resulting in a contradictory consequence inequality $[A^\top\lambda + B^\top\mu]^\top x\,\Omega\,0$ with $\Omega =$ " $<$ " when $\lambda \neq 0$ and $\Omega =$ " $\leq$ " when $\lambda = 0$. Because the given system of constraints is homogeneous, the only one of these two options which can lead to a contradictory consequence inequality is that $A^\top\lambda + B^\top\mu = 0$ and $\lambda \neq 0$.

Exercise I.32. For the systems of constraints to follow, write them down equivalently in the standard form $Ax < b, Cx \leq d$ and point out their feasibility status ("feasible – infeasible") along with the corresponding certificates (certificate for feasibility is a feasible solution to the system; certificate for infeasibility is a collection of weights of constraints which leads to a contradictory consequence inequality, as explained in GTA).

1. $x \leq 0$ $(x \in \mathbf{R}^n)$

   Solution: already in the standard form, feasible, feasibility certificate $x = 0$.

2. $x \leq 0$, and $\sum_{i=1}^n x_i > 0$ $(x \in \mathbf{R}^n)$

   Solution: the standard form is given by $-\sum_i x_i < 0$, and $x \leq 0$, infeasible, infeasibility certificate $\lambda = [1; \ldots; 1] \in \mathbf{R}^{n+1}$

3. $-1 \leq x_i \leq 1$, $1 \leq i \leq n$, $\sum_{i=1}^n x_i \geq n$ $(x \in \mathbf{R}^n)$

   Solution: the standard form is given by $-\sum_i x_i \leq -n$, $x \leq [1; \ldots; 1]$, $-x \leq [1; \ldots; 1]$, feasible, feasibility certificate $x = [1; \ldots; 1]$.

4. $-1 \leq x_i \leq 1$, $1 \leq i \leq n$, $\sum_{i=1}^n x_i > n$ $(x \in \mathbf{R}^n)$

   Solution: the standard form is given by $-\sum_i x_i < -n$, $x \leq [1; \ldots; 1]$, $-x \leq [1; \ldots; 1]$, infeasible, infeasibility certificate is $\lambda = [1; 1; \ldots; 1; 0; \ldots; 0]$ (n zeros).

5. $-1 \leq x_i \leq 1$, $1 \leq i \leq n$, $\sum_{i=1}^n i x_i \geq \frac{n(n+1)}{2}$ $(x \in \mathbf{R}^n)$

   Solution: the standard form is given by $-\sum_i i x_i \leq -\frac{n(n+1)}{2}$, $x \leq [1; \ldots; 1]$, $-x \leq [1; \ldots; 1]$, feasible, feasibility certificate $x = [1; \ldots; 1]$.

6. $-1 \leq x_i \leq 1$, $1 \leq i \leq n$, $\sum_{i=1}^n i x_i > \frac{n(n+1)}{2}$ $(x \in \mathbf{R}^n)$

   Solution: the standard form is given by $-\sum_i i x_i < -\frac{n(n+1)}{2}$, $x \leq [1; \ldots; 1]$, $-x \leq [1; \ldots; 1]$, infeasible, infeasibility certificate is $\lambda = [1; 1; 2; 3; \ldots; n; 0; \ldots; 0]$ (n zeros).

7. $x \in \mathbf{R}^2$, $|x_1| + x_2 \leq 1$, $x_2 \geq 0$, $x_1 + x_2 = 1$

   Solution: the standard form is given by $-x_1 + x_2 \leq 1$, $x_1 + x_2 \leq 1$, $-x_2 \leq 0$, $x_1 + x_2 \leq 1$, $-x_1 - x_2 \leq -1$, feasible, feasibility certificate $x = [1; 0]$.

8. $x \in \mathbf{R}^2$, $|x_1| + x_2 \leq 1$, $x_2 \geq 0$, $x_1 + x_2 > 1$

   Solution: the standard form is given by $-x_1 + x_2 \leq 1$, $x_1 + x_2 \leq 1$, $-x_2 \leq 0$, $-x_1 - x_2 < -1$, infeasible, infeasibility certificate is $\lambda = [0; 1; 0; 1]$.

9. $x \in \mathbf{R}^4$, $x \geq 0$, the sum of two largest entries in x does not exceed 2, and $x_1 + x_2 + x_3 \geq 3$

   Solution: the standard form is given by $-x \leq 0$, $x_i + x_j \leq 2$, $1 \leq i < j \leq 4$, $-x_1 - x_2 - x_3 \leq -3$, feasible, feasibility certificate $x = [1; 1; 1; 0]$.

10. $x \in \mathbf{R}^4$, $x \geq 0$, the sum of two largest entries in x does not exceed 2, and $x_1 + x_2 + x_3 > 3$

    Solution: the standard form is given by $-x \leq 0$, $x_i + x_j \leq 2$, $1 \leq i < j \leq 4$, $-x_1 - x_2 - x_3 < -3$, infeasible, infeasibility certificate is as follows: sum up inequalities $x_1 + x_2 \leq 2$, $x_2 + x_3 \leq 2$, $x_1 + x_3 \leq 2$ with weights $1/2$ and add the inequality $-x_1 - x_2 - x_3 < -3$ with weight 1.

Exercise I.33. Let $(\mathcal{S})$ be the following system of linear inequalities in variables $x \in \mathbf{R}^3$

$$x_1 \leq 1,\ x_1 + x_2 \leq 1,\ x_1 + x_2 + x_3 \leq 1 \qquad (\mathcal{S})$$

In the following list, point out which inequalities are/are not consequences of this system, and certify your claims. To certify that a given inequality is a consequence of the given system, you need to provide nonnegative aggregation weights $\lambda \in \mathbf{R}^3_+$ for the inequalities in $(\mathcal{S})$ such that the

resulting consequence inequality implies the given inequality. To certify that a given inequality is not a consequence of the given system $(\mathcal{S})$, you need to find a point $x \in \mathbf{R}^3$ that satisfies the given system but violates the given inequality.

1. $3x_1 + 2x_2 + x_3 \leq 4$

   Solution: This is a consequence of the system with the certificate $\lambda = [1; 1; 1]$, i.e., when taking weighted sum of the inequalities from the system with weights $\lambda_1, \lambda_2, \lambda_3$, we get the inequality $3x_1 + 2x_2 + x_3 \leq 3$, which clearly implies the target inequality.

2. $3x_1 + 2x_2 + x_3 \leq 2$

   Solution: This is not a consequence of the system. A certificate for this is $x = [1; 0; 0]$ – this vector is feasible to the system but does not satisfy the inequality $3x_1 + 2x_2 + x_3 \leq 2$.

3. $3x_1 + 2x_2 \leq 3$

   Solution: This is a consequence of the system, the certificate being $\lambda = [1; 2; 0]$.

4. $3x_1 + 2x_2 \leq 2$

   Solution: This is not a consequence of the system, a certificate being $x = [1; 0; 0]$.

5. $3x_1 + 3x_2 + x_3 \leq 3$

   Solution: This is a consequence of the system, a certificate being $\lambda = [0; 2; 1]$.

6. $3x_1 + 3x_2 + x_3 \leq 2$

   Solution: This is not a consequence of the system, a certificate being $x = [1; 0; 0]$.

Make a generalization: prove that a linear inequality $px_1 + qx_2 + rx_3 \leq s$ is a consequence of $(\mathcal{S})$ if and only if $s \geq p \geq q \geq r \geq 0$.

Solution: By Inhomogeneous Farkas Lemma, an inequality is a consequence of the (feasible!) system $(\mathcal{S})$ if and only if there exists a nonnegative vector $\lambda \in \mathbf{R}^3_+$ such that $\lambda_1[1; 0; 0] + \lambda_2[1; 1; 0] + \lambda_3[1; 1; 1] = [p; q; r]$ and $\lambda_1 + \lambda_2 + \lambda_3 \leq s$, which is equivalent to $p = \lambda_1 + \lambda_2 + \lambda_3,\ q = \lambda_1 + \lambda_2,\ r = \lambda_3,\ p \leq s$, which in turn is equivalent to $s \geq p \geq q \geq r \geq 0$.

Exercise I.34. Is the inequality $x_1 + x_2 \leq 1$ a consequence of the system $x_1 \leq 1, x_1 \geq 2$? If yes, can it be obtained by taking a legitimate weighted sum of inequalities from the system and the identically true inequality $0^\top x \leq 1$, as it is suggested by the Inhomogeneous Farkas Lemma?

Solution: The given system is infeasible, and therefore *every* inequality is a consequence of the system. The consequence in question cannot be obtained by aggregating inequalities from the system and the identically true inequality $0^\top x \leq 1$, since in every aggregation of this type the coefficient at x_2 is zero. There is no contradiction with the Inhomogeneous Farkas Lemma, since the latter deals with *feasible* systems of inequalities and this is not applicable in our case.

Exercise I.35. Certify the correct statements in the following list:

1. The polyhedral set $X = \left\{x \in \mathbf{R}^3 :\ x \geq [1/3; 1/3; 1/3],\ \sum_{i=1}^3 x_i \leq 1\right\}$ is nonempty.

   Solution: A certificate is $x = [1/3; 1/3; 1/3] \in X$.

2. The polyhedral set $X = \left\{x \in \mathbf{R}^3 :\ x \geq [1/3; 1/3; 1/3],\ \sum_{i=1}^3 x_i \leq 0.99\right\}$ is empty.

   Solution: A certificate is $\lambda = [-1; -1; -1; 1]$: by taking weighted sum of the inequalities defining X using these weights λ is legitimate and leads to the contradictory inequality $0^\top x \leq -0.01$.

3. The linear inequality $x_1 + x_2 + x_3 \geq 2$ is violated somewhere on the polyhedral set $X = \left\{x \in \mathbf{R}^3 :\ x \geq [1/3; 1/3; 1/3],\ \sum_{i=1}^3 x_i \leq 1\right\}$.

   Solution: A certificate is $x = [1/3; 1/3; 1/3]$: this point belongs to X but does not satisfy the given inequality.

4. The linear inequality $x_1 + x_2 + x_3 \geq 2$ is violated somewhere on the polyhedral set $X = \left\{x \in \mathbf{R}^3 :\ x \geq [1/3; 1/3; 1/3],\ \sum_{i=1}^3 x_i \leq 0.99\right\}$.

   Solution: This statement is false: X is empty, and therefore every linear inequality is satisfied everywhere on X.

5. The linear inequality $x_1 + x_2 \le 3/4$ is satisfied everywhere on the polyhedral set $X = \{x \in \mathbf{R}^3 : \ x \ge [1/3; 1/3; 1/3], \ \sum_{i=1}^3 x_i \le 1.05\}$.

   Solution: A certificate is $\lambda = [0; 0; -1; 1]$ – taking weighted sum of the inequalities $x_1 \ge 1/3$, $x_2 \ge 1/3$, $x_3 \ge 1/3$, $x_1 + x_2 + x_3 \le 1.05$ with the weights $\lambda_1, \ldots, \lambda_4$, we get the inequality $x_1 + x_2 \le 1.05 - 1/3 < 3/4$.

6. The polyhedral set $Y = \{x \in \mathbf{R}^3 : \ x_1 \ge 1/3, \ x_2 \ge 1/3, \ x_3 \ge 1/3\}$ is not contained in the polyhedral set $X = \{x \in \mathbf{R}^3 : \ x \ge [1/3; 1/3; 1/3], \ \sum_{i=1}^3 x_i \le 1\}$.

   Solution: A certificate is $x = [1; 1; 1]$: this point is contained in Y but it is not contained in X.

7. The polyhedral set $Y = \{x \in \mathbf{R}^3 : \ x \ge [1/3; 1/3; 1/3], \ \sum_{i=1}^3 x_i \le 1\}$ is contained in the polyhedral set $X = \{x \in \mathbf{R}^3 : \ x_1 + x_2 \le 2/3, \ x_2 + x_3 \le 2/3, \ x_1 + x_3 \le 2/3\}$.

   Solution: It suffices to certify that every one of the constraints defining X is valid for Y, i.e., is a consequence of the constraints defining Y. The inequality $x_1 + x_2 \le 2/3$ can be obtained as the weighted sum of the inequalities $x_1 \ge 1/3, \ x_2 \ge 1/3, \ x_3 \ge 1/3, \ x_1 + x_2 + x_3 \le 1$ using the weights $0, 0, -1, 1$, and thus it is valid for Y. Similarly, we can certify that the inequalities $x_1 + x_3 \le 2/3$ and $x_2 + x_3 \le 2/3$ are valid for Y.

5.8 Around Linear Programming Duality

Exercise I.36. Let the polyhedral set $P = \{x \in \mathbf{R}^n : Ax \le b\}$, where $A = [a_1^\top; ...; a_m^\top]$, be nonempty. Prove that P is bounded if and only if every vector from $\mathbf{R}^n$ can be represented as a linear combination of the vectors a_i with nonnegative coefficients where at most n coefficients are positive. As a result, given A, all nonempty sets of the form $\{x \in \mathbf{R}^n : Ax \le b\}$ simultaneously are/are not bounded.

Solution: P is bounded if and only if for every $z \in \mathbf{R}^n$ the feasible LP program

$$\max_x \{z^\top x : \ Ax \le b\}$$

is bounded and thus is solvable, or, which is the same by LP Duality Theorem, if and only if the dual of this problem, i.e.,

$$\min_\lambda \{b^\top \lambda : \ \lambda \ge 0, A^\top \lambda = z\}$$

is solvable. Next, since the dual of this dual is feasible (P is nonempty!), the dual automatically is bounded, so that its solvability is the same as its feasibility. We conclude that P is bounded if and only if every $x \in \mathbf{R}^n$ is a conic combination of a_i's.
Applying the Caratheodory Theorem in conic form, the latter is the same as the possibility to represent every $z \in \mathbf{R}^n$ as a conic combination of at most n vectors from the collection $a_1, ..., a_m$.

Exercise I.37. Consider the linear program

$$\mathrm{Opt} = \max_{x \in \mathbf{R}^2} \{x_1 : \ x_1 \ge 0, \ x_2 \ge 0, \ ax_1 + bx_2 \le c\} \qquad (P)$$

where a, b, c are parameters. Answer the following questions:

1. Let $c = 1$. Is the problem feasible?

   Solution: The problem is feasible; a certificate is a feasible solution, e.g., $x = 0$.

2. Let $a = b = 1, \ c = -1$. Is the problem feasible?

   Solution: The problem is infeasible; a certificate for infeasibility is given by $\lambda = [-1; -1; 1]$: summing up the constraints $x_1 \ge 0, \ x_2 \ge 0, \ x_1 + x_2 \le -1$ with weights $-1, -1, 1$, we get the contradictory inequality $0^\top x \le -1$.

3. Let $a = b = 1,\ c = -1$. Is the problem bounded[5]?

   Solution: The problem is bounded since it is infeasible; certificate of infeasibility was given in the previous item.

4. Let $a = b = c = 1$. Is the problem bounded?

   Solution: The problem is bounded since the weighted sum of the constraints $x_1 \geq 0$, $x_2 \geq 0$, $x_1 + x_2 \leq 1$, the weights being $0, -1, 1$, gives us the consequence inequality $x_1 \leq 1$. Thus, the objective $\max x_1$ is bounded from above on the feasible set.

5. Let $a = 1,\ b = -1,\ c = 1$. Is the problem bounded?

   Solution: The problem is unbounded. Indeed, setting $d = [1; 1]$, we get $d_1 \geq 0$, $d_2 \geq 0$, $ad_1 + bd_2 = d_1 - d_2 \leq 0$, $[1; 0]^\top d = d_1 > 0$, implying that the points $x(t) := td$ are feasible when $t \geq 0$; it remains to note that the objective, as evaluated at $x(t)$, tends to $+\infty$ as $t \to \infty$.

6. Let $a = b = c = 1$. Is it true that $\mathrm{Opt} \geq 0.5$?

   Solution: A certificate is a feasible solution with objective value ≥ 0.5, e.g., the solution $x = [0.5; 0.5]$.

7. Let $a = b = 1,\ c = -1$. Is it true that $\mathrm{Opt} \leq 1$?

   Solution: The claim is true due to the fact that the problem is infeasible, for infeasibility certificate see item 2, and thus $\mathrm{Opt} = -\infty$.

8. Let $a = b = c = 1$. Is it true that $\mathrm{Opt} \leq 1$?

   Solution: The claim is true, a certificate is the collection of weights (Lagrange multipliers) $0, -1, 1$; taking the corresponding weighted sum of the constraints $x_1 \geq 0,\ x_2 \geq 0,\ x_1 + x_2 \leq 1$, we get the inequality $x_1 \leq 1$. Thus, the objective does not exceed 1 on the feasible set.

9. Let $a = b = c = 1$. Is it true that $x_* = [1; 1]$ is an optimal solution of (P)?

   Solution: The claim is false since x_* is infeasible (it violates the constraint $x_1 + x_2 \leq 1$).

10. Let $a = b = c = 1$. Is it true that $x_* = [1/2; 1/2]$ is an optimal solution of (P)?

    Solution: The claim is false since there exists a feasible solution $x = [1; 0]$ with larger objective value.

11. Let $a = b = c = 1$. Is it true that $x_* = [1; 0]$ is an optimal solution of (P)?

    Solution: The claim is true, the corresponding certificate is the collection of Lagrange multipliers $0, -1, 1$ associated with the constraints $x_1 \geq 0$, $x_2 \geq 0$, $x_1 + x_2 \leq 1$. Indeed, the multipliers are of proper signs, satisfy the complementary slackness condition and the KKT equation $0 \times [1; 0] - 1 \times [0; 1] + 1 \times [1; 1] = [1; 0]$.

Exercise I.38. Consider the LP program

$$\max_{x_1, x_2} \left\{ -x_2 : \begin{array}{c} x_1 \leq 0 \\ -x_1 \leq -1 \\ x_2 \leq 1 \end{array} \right\}$$

Write down the dual problem and check whether the optimal values are equal to each other.

Solution: The dual problem reads

$$\min_{\lambda_1, \lambda_2, \lambda_3} \left\{ -\lambda_2 + \lambda_3 : \begin{array}{c} \lambda_1 - \lambda_2 = 0 \\ \lambda_3 = -1 \\ \lambda_i \geq 0,\ 1 \leq i \leq 3 \end{array} \right\}$$

Both problems are clearly infeasible, and their optimal values ($-\infty$ and $+\infty$, respectively) differ from each other.

Exercise I.39. Write down the problems dual to the following linear programs:

[5] Recall that a maximization problem is called *bounded*, if the objective is bounded from above on the feasible set, which is the same as its optimal value being $< \infty$

1. $\max_{x\in\mathbf{R}^3}\left\{x_1+2x_2+3x_3:\ \begin{array}{l}x_1-x_2+x_3=0,\\ x_1+x_2-x_3\geq 100,\\ x_1\leq 0,\\ x_2\geq 0,\\ x_3\geq 0\end{array}\right\}$

   Solution: The dual problem is

$$\min_{\lambda\in\mathbf{R}^5}\left\{100\lambda_2:\ \begin{array}{l}\lambda_2\leq 0,\ \lambda_3\geq 0,\ \lambda_4\leq 0,\ \lambda_5\leq 0,\\ \lambda_1+\lambda_2+\lambda_3=1,\\ -\lambda_1+\lambda_2+\lambda_4=2,\\ \lambda_1-\lambda_2+\lambda_5=3\end{array}\right\}.$$

2. $\max_{x\in\mathbf{R}^n}\left\{c^\top x:\ Ax=b,\ x\geq 0\right\}$

   Solution: The dual problem is

$$\min_{\lambda=[\lambda_e;\lambda_g]}\left\{b^\top\lambda_e:\ \begin{array}{l}\lambda_g\leq 0,\\ A^\top\lambda_e+\lambda_g=c\end{array}\right\},$$

   or, after eliminating λ_g:

$$\min_{\lambda_e}\left\{b^\top\lambda_e:\ c\leq A^\top\lambda_e\right\}.$$

3. $\max_{x\in\mathbf{R}^n}\left\{c^\top x:\ Ax=b,\ \underline{u}\leq x\leq\overline{u}\right\}$

   Solution: The dual problem is

$$\min_{\lambda=[\lambda_e;\lambda_g;\lambda_\ell]}\left\{\overline{u}^\top\lambda_\ell+\underline{u}^\top\lambda_g+b^\top\lambda_e:\ \begin{array}{l}\lambda_\ell\geq 0,\ \lambda_g\leq 0,\\ \lambda_\ell+\lambda_g+A^\top\lambda_e=c\end{array}\right\}.$$

4. $\max_{x,y}\left\{c^\top x:\ Ax+By\leq b,\ x\leq 0,\ y\geq 0\right\}$

   Solution: The dual problem is

$$\min_{\lambda=[\lambda_{\ell,b};\lambda_{\ell,0},\lambda_g]}\left\{b^\top\lambda_{\ell,b}:\ \begin{array}{l}\lambda_{\ell,b}\geq 0,\ \lambda_{\ell,0}\geq 0,\ \lambda_g\leq 0,\\ A^\top\lambda_{\ell,b}+\lambda_{\ell,0}=c,\\ B^\top\lambda_{\ell,b}+\lambda_g=0\end{array}\right\},$$

   or, after eliminating $\lambda_{\ell,0}$ and λ_g,

$$\min_{\lambda_{\ell,b}}\left\{b^\top\lambda_{\ell,b}:\ \lambda_{\ell,b}\geq 0,\ A^\top\lambda_{\ell,b}\leq c,\ B^\top\lambda_{\ell,b}\geq 0\right\}.$$

Exercise I.40. Consider a primal-dual pair of linear programs given by

$$\mathrm{Opt}(P)=\min_x\left\{c^\top x:\ Ax\geq b\right\},\tag{P}$$

$$\mathrm{Opt}(D)=\max_y\left\{b^\top y:\ y\geq 0,\ A^\top y=c\right\}.\tag{D}$$

Suppose that both are feasible. Prove that the feasible set of at least one of these problems is unbounded.

Solution: See solution to Exercise IV.25.

Exercise I.41. Consider the following linear program

$$\mathrm{Opt}=\min_{\{x_{ij}\}_{1\leq i<j\leq 4}}\left\{2\sum_{1\leq i<j\leq 4}x_{ij}:\ x_{ij}\geq 0,\ 1\leq i<j\leq 4,\ \sum_{j>i}x_{ij}+\sum_{j<i}x_{ji}\geq i,\ 1\leq i\leq 4\right\}.$$

1. Show that the optimum objective value is at most 20.

Solution: The solution $x_{34} = 4$, $x_{23} = 3$, $x_{12} = 2$, and all other variables equal to 0 is a feasible solution to this LP and has the objective value equal to $2(2+3+4) = 18$. Since this is a minimization problem, we deduce that $\mathrm{Opt} \leq 18$.

2. Show that the optimum objective value is at least 10.

Solution: The dual of this LP is given by

$$\max_{y \in \mathbf{R}^4} \left\{ \sum_{i=1}^{4} i y_i : \ y_i \geq 0 \ \forall i, \ y_i + y_j \leq 2, \ 1 \leq i < j \leq 4 \right\}.$$

The solution $y_1 = y_2 = y_3 = y_4 = 1$ is feasible to the dual with an objective value of $1+2+3+4 = 10$. Therefore, by Weak LP Duality, we deduce that $\mathrm{Opt} \geq 10$.

Exercise I.42. We say that an $n \times n$ matrix P is *stochastic* if all of its entries are all nonnegative and the sum of the entries of each row is equal to 1. Show that if P is a stochastic matrix, then there is a nonzero vector $a \in \mathbf{R}^n$ such that $P^\top a = a$ and $a \geq 0$.

Solution: Consider the linear program

$$\min_{x \in \mathbf{R}^n} \left\{ 0^\top x : \ Px \geq x + e \right\},$$

where e is the all-ones vector in $\mathbf{R}^n$. Suppose that there is a feasible solution x to this LP, and let the index i be such that $x_i = \max_j \{x_j\}$. We have that $x_i + 1 \leq \sum_{k=1}^n P_{i,k} x_k \leq \sum_{k=1}^n P_{i,k} x_i = x_i \sum_{k=1}^n P_{i,k} = x_i$, which is a contradiction. Therefore, this LP is infeasible. This means that its dual is either infeasible or unbounded. The dual problem is given by $\max_{y \in \mathbf{R}^n} \left\{ e^\top y : \ y^\top P = y^\top, \ y \geq 0 \right\}$. Clearly, the solution $y = 0$ is feasible for the dual; thus the dual must be unbounded. Therefore, there is an $a \neq 0$, such that $a^\top P = a^\top$ and $a \geq 0$.

Exercise I.43. Let $A \in \mathbf{R}^{n \times n}$ be a symmetric matrix, i.e., $A^\top = A$. Consider the linear program

$$\min_x \left\{ c^\top x : \ Ax \geq c, \ x \geq 0 \right\}.$$

Prove that if $\bar{x}$ satisfies $A\bar{x} = c$ and $\bar{x} \geq 0$, then $\bar{x}$ is optimal.

Solution: Note that the dual of this optimization problem is given by

$$\max_{\lambda, \mu} \left\{ c^\top \lambda : \ A\lambda + \mu = c, \ \lambda \geq 0, \ \mu \geq 0 \right\},$$

where we used $A^\top = A$. The solution $\bar{x}$ along with $\bar{\mu} = 0$ such that $A\bar{x} = c$ and $\bar{x} \geq 0$ is thus feasible for the dual problem, and $\bar{x}$ is feasible for the primal one, with the same objective value. Therefore, by the "zero duality gap" LP optimality condition, Theorem I.4.11, we deduce that $\bar{x}$ is optimal for the primal problem.

Exercise I.44. Let $w \in \mathbf{R}^n$, and let $A \in \mathbf{R}^{n \times n}$ be a *skew-symmetric* matrix, i.e., $A^\top = -A$. Consider the following linear program

$$\mathrm{Opt}(P) = \min_{x \in \mathbf{R}^n} \left\{ w^\top x : \ Ax \geq -w, \ x \geq 0 \right\}.$$

Suppose that the problem is solvable. Provide a closed analytical expression for $\mathrm{Opt}(P)$.

Solution: The dual problem is given by

$$\begin{aligned}
\mathrm{Opt}(D) &= \max_{u,v} \left\{ -w^\top u : \ A^\top u + v = w, \ u \geq 0, \ v \geq 0 \right\} \\
&= \max_{u} \left\{ -w^\top u : \ A^\top u \leq w, \ u \geq 0 \right\} \\
&= \max_{u} \left\{ -w^\top u : \ -Au \leq w, \ u \geq 0 \right\} \qquad [\text{since } A^\top = -A] \\
&= \max_{u} \left\{ -w^\top u : \ Au \geq -w, \ u \geq 0 \right\} \\
&= -\min_{u} \left\{ w^\top u : Au \geq -w, \ u \geq 0 \right\} = -\mathrm{Opt}(P).
\end{aligned}$$

We are given that the primal problem is solvable, thus $\mathrm{Opt}(P) = \mathrm{Opt}(D)$ by LP Duality Theorem, and at the same time, as we just have seen, $\mathrm{Opt}(D) = -\mathrm{Opt}(P)$, implying that $\mathrm{Opt}(P) = 0$.

Exercise I.45. [Separation Theorem, polyhedral version] Let P and Q be two nonempty polyhedral sets in $\mathbf{R}^n$ such that $P \cap Q = \varnothing$. Suppose that the polyhedral descriptions of these sets are given as

$$P := \{x \in \mathbf{R}^n : \ Ax \leq b\} \quad \text{and} \quad Q := \{x \in \mathbf{R}^n : \ Dx \geq d\}.$$

Using LP duality show that there exists a vector $c \in \mathbf{R}^n$ such that

$$c^\top x < c^\top y, \quad \text{for all } x \in P \text{ and } y \in Q.$$

Solution: Consider the following linear program

$$\max_x \left\{0^\top x : \ Ax \leq b, \ Dx \geq d\right\},$$

together with its dual given by

$$\min_{p,q} \left\{b^\top p + d^\top q : \ A^\top p + D^\top q = 0, \ p \geq 0, \ q \leq 0\right\}.$$

Since $P \cap Q = \varnothing$, the primal problem is infeasible, therefore the dual problem can be either infeasible or unbounded. But $p = 0, q = 0$ is a feasible solution to the dual problem therefore, we conclude that the dual problem is unbounded, i.e., there exists, (z_p, z_q) such that $A^\top z_p + D^\top z_q = 0$, $z_p \geq 0$, $z_q \leq 0$ and $b^\top z_p + d^\top z_q < 0$. Let $c := A^\top z_p$. Then, for any $x \in P$ and any $y \in Q$, we have

$$\begin{aligned}
c^\top x = z_p^\top Ax &\leq z_p^\top b && [\text{as } z_p \geq 0 \text{ and } Ax \leq b] \\
&< -d^\top z_q && [\text{as } b^\top z_p + d^\top z_q < 0] \\
&\leq z_q^\top (-Dy) && [\text{as } z_q \leq 0 \text{ and } Dy \geq d] \\
&\leq c^\top y, && [\text{as } c = A^\top z_p = -D^\top z_q]
\end{aligned}$$

and thus we have proved the result.

Exercise I.46. Suppose we are given the linear program

$$\min_x \left\{c^\top x : \ Ax = b, \ x \geq 0\right\} \tag{P}$$

and its associated *Lagrangian* function

$$L(x, \lambda) := c^\top x + \lambda^\top (b - Ax).$$

The LP dual to (P) is (replacing $Ax = b$ with $Ax \geq b$, $-Ax \geq -b$)

$$\mathrm{Opt}(D) = \max_{\lambda_\pm, \mu} \left\{b^\top [\lambda_+ - \lambda_-] : A^\top [\lambda_+ - \lambda_-] + \mu = c, \lambda_\pm \geq 0, \mu \geq 0\right\},$$

or, after eliminating μ and setting $\lambda = \lambda_+ - \lambda_-$,

$$\mathrm{Opt}(D) = \max_\lambda \left\{b^\top \lambda : A^\top \lambda \leq b\right\}. \tag{D}$$

Now, let us consider the following game: Player 1 chooses some $x \geq 0$, and player 2 chooses some λ simultaneously; then, player 1 pays to player 2 the amount $L(x, \lambda)$. In this game, player 1 would like to minimize $L(x, \lambda)$ and player 2 would like to maximize $L(x, \lambda)$.

A pair (x^*, λ^*) with $x^* \geq 0$, is called an *equilibrium* point (or *saddle point* or *Nash equilibrium*) if

$$L(x^*, \lambda) \leq L(x^*, \lambda^*) \leq L(x, \lambda^*), \quad \forall x \geq 0 \ \text{ and } \ \forall \lambda. \tag{$*$}$$

(That is, in an equilibrium no player is able to improve his performance by unilaterally modifying his choice.)

Show that x^* and λ^* are optimal solutions to the problem (P) and to its dual, respectively, if and only if (x^*, λ^*) is an equilibrium point.

Solution: First, suppose that x^* and λ^* are optimal solutions of (P) and (D), respectively. We will show that they are in equilibrium. Since x^* is primal feasible, we have $Ax^* = b$, and so $L(x^*, \lambda) = c^\top x^* = L(x^*, \lambda^*)$ which proves the left inequality in $(*)$. Moreover, as λ^* is dual feasible, we have $c - A^\top \lambda^* \geq 0$. Hence, for every $x \geq 0$, we obtain

$$L(x, \lambda^*) = (c - A^\top \lambda^*)^\top x + b^\top \lambda^* \geq b^\top \lambda^* = c^\top x^* = L(x^*, \lambda^*),$$

where the second equality follows from the LP Duality Theorem which gives us $\mathrm{Opt}(P) = \mathrm{Opt}(D)$, i.e., $c^\top x^* = b^\top \lambda^*$. Justification of $(*)$ is complete.
We now prove the reverse. Suppose that $x^* \geq 0$ and λ^* are in equilibrium. The inequality $L(x^*, \lambda) \leq L(x^*, \lambda^*)$ yields $\lambda^\top (b - Ax^*) \leq (\lambda^*)^\top (b - Ax^*)$ for all λ. This can happen only if $Ax^* = b$, which establishes the primal feasibility of x^*. Furthermore, the inequality $L(x^*, \lambda^*) \leq L(x, \lambda^*)$ leads to $c^\top x^* \leq (c - A^\top \lambda^*)^\top x + b^\top \lambda^*$. Since this must be true for all $x \geq 0$, we get $c^\top x^* \leq b^\top \lambda^*$ (set $x = 0$) and $c - A^\top \lambda^* \geq 0$ and therefore λ^* is dual feasible. By weak LP duality, we conclude that $c^\top x^* = b^\top \lambda^*$ and it follows that x^* and λ^* are optimal solutions of the primal and the dual problems, respectively.

Exercise I.47. Given a polyhedral set $X = \left\{x \in \mathbf{R}^n : \ a_i^\top x \leq b_i, \ \forall i = 1, \ldots, m\right\}$, consider the associated optimization problem

$$\mathrm{Opt}(X) = \max_{x,t} \left\{t : \ B_\infty(x, t) \subseteq X\right\},$$

where $B_\infty(x, t) := \{y \in \mathbf{R}^n : \ \|y - x\|_\infty \leq t\}$. Is it possible to pose this optimization problem as a linear program with a polynomial in m, n number of variables and constraints? If it is possible, give such a representation explicitly. If not, argue why.

Solution: Note that in order for (x, t) to be feasible to the given optimization problem, for every $i = 1, \ldots, m$, we must have

$$b_i \geq \max_{y \in B_\infty(x,t)} \{a_i^\top y\} = a_i^\top x + t\|a_i\|_1,$$

where the last equality is evident. Hence, we arrive at

$$\mathrm{Opt}(X) = \max_{x,t} \left\{t : \ a_i^\top x + t\|a_i\|_1 \leq b_i, \ i = 1, \ldots, m\right\},$$

which clearly is a formulation with polynomially many variables and inequalities.

Exercise I.48. Consider the optimization problem

$$\min_{x \in \mathbf{R}^n} \left\{c^\top x : \ \tilde{a}_i^\top x \leq b_i \text{ for some } \tilde{a}_i \in A_i, \ i = 1, \ldots, m, \ x \geq 0\right\}, \qquad (*)$$

where $A_i = \{\bar{a}_i + \epsilon_i : \ \|\epsilon_i\|_\infty \leq \rho\}$ for $i = 1, \ldots, m$. In this problem, we basically mean that the constraint coefficient $\tilde{a}_{ij}$ (j-th component of the i-th constraint vector $\tilde{a}_i$) belongs to the interval uncertainty set $[\bar{a}_{ij} - \rho, \ \bar{a}_{ij} + \rho]$, where $\bar{a}_{ij}$ is its nominal value. That is, in $(*)$, we are seeking a solution x such that each constraint is satisfied for *some* coefficient vector from the corresponding uncertainty set.

Note that in its current form $(*)$, this problem is not a linear program (LP). Prove that it can be written as an *explicit* linear program and give the corresponding LP formulation.

Solution: This problem is equivalent to

$$\begin{aligned} \min \ & c^\top x \\ \text{s.t.} \ & \min_{\tilde{a}_i \in A_i} \{\tilde{a}_i^\top x\} \leq b_i \quad i = 1, \ldots, m \\ & x \geq 0. \end{aligned}$$

Note that when $x \geq 0$

$$\begin{aligned} \min\{\tilde{a}_i^\top x : \ \tilde{a}_i = \bar{a}_i + \epsilon_i, \ \|\epsilon_i\|_\infty \leq \rho\} &= \bar{a}_i^\top x + \min\{\epsilon_i^\top x : \ \|\epsilon_i\|_\infty \leq \rho\} \\ &= \bar{a}_i^\top x - \rho \sum_{j=1}^n x_j, \end{aligned}$$

where the last equality is evident. Then, the resulting LP formulation for problem in $(*)$ is given by

$$\begin{aligned} \min \;\; & c^\top x \\ \text{s.t.} \;\; & \overline{a}_i^\top x - \rho \sum_{j=1}^{n} x_j \le b_i \;\; i = 1, \ldots, m \\ & x \ge 0. \end{aligned}$$

Exercise I.49. Let $S = \{a_1, a_2, \ldots, a_n\}$ be a finite set composed of n distinct elements, and let f be a real-valued function defined on the set of all subsets of S. We say that f is *submodular* if, for every $X, Y \subseteq S$, the following inequality holds:

$$f(X) + f(Y) \ge f(X \cup Y) + f(X \cap Y).$$

1. Give an example of a submodular function f.

   Solution: A simple trivial example is the function $f(X) = 0, \forall X \subseteq S$. Here is another simple slightly less trivial example: given $a \in \mathbf{Z}^n_+$, consider the function $f(X) = \max\{a_i : \; i \in X\}$ for every $X \subseteq S$. There are quite a lot of other examples of submodular functions; interested readers can refer to books on submodularity and combinatorial optimization.

2. Let $f : 2^S \to \mathbf{Z}$ be an integer-valued submodular function such that $f(\varnothing) = 0$. Consider the polyhedron

   $$P_f := \left\{ x \in \mathbf{R}^{|S|} : \; \sum_{t \in T} x_t \le f(T), \; \forall T \subseteq S \right\},$$

   Consider

   $$\bar{x}_{a_k} := f(\{a_1, \ldots, a_k\}) - f(\{a_1, \ldots, a_{k-1}\}), \quad k = 1, \ldots, n.$$

   Show that $\bar{x}$ is feasible to P_f.

   Solution: To justify this claim, we need to show that $\sum_{t \in T} \bar{x}_t \le f(T)$ for all subsets T of S. If $T = \varnothing$, then by definition $\sum_{t \in T} \bar{x}_t = 0 = f(\varnothing)$. When $T \neq \varnothing$, we will show this by induction on $\mathrm{Card}(T)$. To see the base case, suppose T is a singleton, i.e., $T = \{a_k\}$ for some $k = 1, \ldots, n$. Then, since f is submodular, we always have

   $$f(\{a_1, \ldots, a_k\}) - f(\{a_1, \ldots, a_{k-1}\}) \le f(\{a_k\}), \quad \forall k = 1, \ldots, n.$$

   Thus, by its definition $\bar{x}_{a_k} \le f(\{a_k\})$ holds for all $k = 1, \ldots, n$. Now, for the inductive hypothesis suppose that $\sum_{t \in T'} \bar{x}_t \le f(T')$ for all subsets T' of S such that $\mathrm{Card}(T') < k$ for some $k \ge 2$, and let us show that the inequality holds for all subsets T of S of cardinality k as well to complete the induction. So, consider any $T = \{a_{\iota_1}, \ldots, a_{\iota_k}\}$, and define $T' := T \setminus \{a_{\iota_k}\}$. Thus, $\mathrm{Card}(T') = k - 1$, and by induction hypothesis we have $\sum_{t \in T'} \bar{x}_t \le f(T')$ and so

   $$\sum_{t \in T} \bar{x}_t = \bar{x}_{a_{\iota_k}} + \sum_{t \in T'} \bar{x}_t \le \big(f(\{a_1, \ldots, a_{\iota_k}\}) - f(\{a_1, \ldots, a_{\iota_k - 1}\}) \big) + f(T')$$

   Now, by defining the sets $X := T$ and $Y := \{a_1, \ldots, a_{\iota_k - 1}\}$, we see that $X \cup Y = \{a_1, \ldots, a_{\iota_k}\}$ and $X \cap Y = \{a_{\iota_1}, \ldots, a_{\iota_{k-1}}\} = T'$. Now, combining the previous inequality with the submodularity of f applied to the sets X and Y, we obtain

   $$\begin{aligned} \sum_{t \in T} \bar{x}_t &\le \big(f(\{a_1, \ldots, a_{\iota_k}\}) - f(\{a_1, \ldots, a_{\iota_k - 1}\}) \big) + f(T') \\ &= f(X \cup Y) - f(Y) + f(X \cap Y) \\ &\le f(X) = f(T), \end{aligned}$$

   as desired. This completes the induction and so $\bar{x}$ is in P_f.

3. Consider the following optimization problem associated with P_f:

$$\max_x \left\{ c^\top x : \ x \in P_f \right\}.$$

Write down the dual of this LP.

Solution: The dual of this LP is

$$\min_y \left\{ \sum_{T: T \subseteq S} f(T) y_T : \ \sum_{T \ni t} y_T = c_t, \, \forall t \in S, \ y_T \geq 0, \, \forall T \subseteq S \right\}.$$

4. Assume without loss of generality that $c_{a_1} \geq c_{a_2} \geq \ldots \geq c_{a_n}$. Identify a dual feasible solution and using the LP Duality Theorem show that the solution $\bar{x}$ specified in item 2 is optimal to the primal maximization problem associated with P_f.

Solution: The following is a feasible solution for the dual problem:

$$\bar{y}_T := \begin{cases} c_{a_k} - c_{a_{k+1}}, & \text{if } T = \{a_1, \ldots, a_k\} \text{ for some } k = 1, \ldots, n \\ 0, & \text{otherwise,} \end{cases}$$

where we define $c_{a_{n+1}} = 0$. Indeed, as $c_{a_1} \geq c_{a_2} \geq \ldots \geq c_{a_n}$, we immediately see that $\bar{y}_T \geq 0$ for all $T \subseteq S$. Now, consider any $t \in S$, and suppose i is such that $a_i = t$. Note that the only dual variables $\bar{y}_T$ that may take positive values are the ones corresponding to the sets $T = \{a_1, \ldots, a_k\}$ for some $k = 1, \ldots, n$. And among such sets the only ones that contain the given $t = a_i$ are the sets $\bar{T}_i := \{a_1, \ldots, a_i\}$, $\bar{T}_{i+1} := \{a_1, \ldots, a_i, a_{i+1}\}, \ldots, \bar{T}_n := \{a_1, \ldots, a_n\}$. Thus, for any $t \in S$, we have

$$\sum_{T \ni t} \bar{y}_T = \sum_{\ell=i}^{n} \bar{y}_{\bar{T}_\ell} = \sum_{\ell=i}^{n} \left(c_{a_\ell} - c_{a_{\ell+1}} \right) = c_{a_i} - c_{a_{n+1}} = c_{a_i} = c_t,$$

where we used the fact that $c_{a_{n+1}} = 0$.
Both of these solutions ($\bar{x}$ and $\bar{y}$) give the same objective value for their corresponding problems as

$$\begin{aligned} c^\top \bar{x} = \sum_{k=1}^{n} c_{a_k} \bar{x}_{a_k} &= \sum_{k=1}^{n} c_{a_k} \left(f(\{a_1, \ldots, a_k\}) - f(\{a_1, \ldots, a_{k-1}\}) \right) \\ &= \sum_{k=1}^{n} f(\{a_1, \ldots, a_k\}) \left(c_{a_k} - c_{a_{k+1}} \right) \\ &= \sum_{T: T \subseteq S} f(T) \bar{y}_T. \end{aligned}$$

Therefore, both solutions are optimal.

Remark. Note that when the submodular function f is integer-valued, we immediately see from the characterization of the optimal primal solution $\bar{x}$ that for all integer vectors $c \in \mathbf{Z}^n$ such that there exists an optimum solution to the primal problem, there exists an optimum solution (e.g. $\bar{x}$) where all variables take integer values. A system of linear inequalities $Ax \leq b$ with $b \in \mathbf{Z}^m$ and $A \in \mathbf{Q}^{m \times n}$ satisfying such a property (i.e., whenever $c \in \mathbf{Z}^n$ is such that there is an optimal solution to $\max_x \{c^\top x : Ax \leq b\}$ then there is an integer optimum solution) is called *totally dual integral* (TDI). Thus, we conclude that the polyhedron P_f associated with an integer-valued submodular function f is TDI. The TDI property is a well-known sufficient condition that guarantees that every extreme point (see section 6.4) of the associated polyhedron is integral. In particular, the TDI property generalizes *total unimodularity* (TU), i.e., the other well-known sufficient condition for the integrality of a polyhedron, which plays a key role in network-flow based optimization.

Exercises from Part II

8.1 Separation

Exercise II.1. Mark by "Y"/"N" those of the below listed cases where the linear form $f^\top x$ separates/does not separate the sets S and T:

- $S = \{0\} \subset \mathbf{R}$, $T = \{0\} \subset \mathbf{R}$, $f^\top x = x$

  Solution: N
- $S = \{0\} \subset \mathbf{R}$, $T = [0,1] \subset \mathbf{R}$, $f^\top x = x$

  Solution: Y
- $S = \{0\} \subset \mathbf{R}$, $T = [-1,1] \subset \mathbf{R}$, $f^\top x = x$

  Solution: N
- $S = \{x \in \mathbf{R}^3 : x_1 = x_2 = x_3\}$, $T = \{x \in \mathbf{R}^3 : x_3 \geq \sqrt{x_1^2 + x_2^2}\}$, $f^\top x = x_1 - x_2$

  Solution: N
- $S = \{x \in \mathbf{R}^3 : x_1 = x_2 = x_3\}$, $T = \{x \in \mathbf{R}^3 : x_3 \geq \sqrt{x_1^2 + x_2^2}\}$, $f^\top x = x_3 - x_2$
  Solution: Y
- $S = \{x \in \mathbf{R}^3 : -1 \leq x_1 \leq 1\}$, $T = \{x \in \mathbf{R}^3 : x_1^2 \geq 4\}$, $f^\top x = x_1$
  Solution: N
- $S = \{x \in \mathbf{R}^2 : x_2 \geq x_1^2, x_1 \geq 0\}$, $T = \{x \in \mathbf{R}^2 : x_2 = 0\}$, $f^\top x = -x_2$
  Solution: Y

Exercise II.2. Consider the set

$$M = \left\{ x \in \mathbf{R}^{2004} : \begin{array}{rcl} x_1 + x_2 + \ldots + x_{2004} & \geq & 1 \\ x_1 + 2x_2 + 3x_3 \ldots + 2004x_{2004} & \geq & 10 \\ x_1 + 2^2 x_2 + 3^2 x_3 \ldots + 2004^2 x_{2004} & \geq & 10^2 \\ \ldots\ldots\ldots\ldots & & \\ x_1 + 2^{2002} x_2 + 3^{2002} x_3 + \ldots + 2004^{2002} x_{2004} & \geq & 10^{2002} \end{array} \right\}$$

Is it possible to separate this set from the set $\{x_1 = x_2 = \ldots = x_{2004} \leq 0\}$? If yes, what could be a separating plane?

Solution: Separation is possible, and a separating plane is, e.g., $\{x : x_1 + \ldots + x_{2004} = 1/2\}$, since the linear form $\sum_i x_i$ is ≥ 1 on M and clearly is ≤ 0 on the set $\{x_1 = ... = x_{2004} \leq 0\}$.

Exercise II.3. Can the sets $S = \{x \in \mathbf{R}^2 : x_1 > 0, x_2 \geq 1/x_1\}$ and $T = \{x \in \mathbf{R}^2 : x_1 < 0, x_2 \geq -1/x_1\}$ be separated? Can they be strongly separated?

Solution: The sets are separated by the line $\{x \in \mathbf{R}^2 : x_1 = 0\}$ They cannot be strongly separated since the distance between the sets is zero (take large $t > 0$ and look at the points $[1/t; t] \in S$ and $[-1/t; t] \in T$).

Exercise II.4. Let $M \subset \mathbf{R}^n$ be a nonempty closed convex set. The *metric projection* $\mathrm{Proj}_M(x)$ of a point $x \in \mathbf{R}^n$ onto M is the $\|\cdot\|_2$-closest to x point of M, so that

$$\mathrm{Proj}_M(x) \in M \ \& \ \|x - \mathrm{Proj}_M(x)\|_2^2 = \min_{y \in M} \|x - y\|_2^2. \qquad (*)$$

1. Prove that for every $x \in \mathbf{R}^n$ the minimum in the right hand side of $(*)$ is achieved, and x_+ is a minimizer if and only if

$$x_+ \in M \ \& \ \forall y \in M : [x - x_+]^\top [x_+ - y] \geq 0. \tag{8.1}$$

   Derive from the latter fact that the minimum in $(*)$ is achieved at a unique point, the bottom line being that $\mathrm{Proj}_M(\cdot)$ is well defined
2. Prove that when passing from a point $x \in \mathbf{R}^n$ to its metric projection $x_+ = \mathrm{Proj}_M(x)$, the distance to any point of M does not increase, specifically,

$$\begin{array}{rcl} \forall y & \in & M : \|x_+ - y\|_2^2 \leq \|x - y\|_2^2 - \mathrm{dist}^2(x, M), \\ \mathrm{dist}(x, M) & := & \min_{u \in M} \|x - u\|_2 = \|x - x_+\|_2. \end{array} \tag{8.2}$$

3. Let $x \notin M$, so that, denoting $x_+ = \mathrm{Proj}_M(x)$, the vector $e = \frac{x - x_+}{\|x - x_+\|_2}$ is well defined. Prove that the linear form $e^\top z$ strongly separates $\{x\}$ and M, specifically,

$$\forall y \in M : e^\top y \leq e^\top x - \mathrm{dist}(x, M).$$

   Note: The fact just outlined underlies an alternative proof of Separation Theorem, where the first step is to prove that a point outside a nonempty closed convex set can be strongly separated from the set. In our proof, the first step was similar, but with M restricted to be polyhedral, rather than merely convex and closed.
4. Prove that the mapping $x \mapsto \mathrm{Proj}_M(x) : \mathbf{R}^n \to M$ is *contraction* in $\|\cdot\|_2$:

$$\forall u, u' \in \mathbf{R}^n : \|\mathrm{Proj}_M(u) - \mathrm{Proj}_M(u')\|_2 \leq \|u - u'\|_2.$$

5. Let M be the probabilistic simplex: $M = \{x \in \mathbf{R}^n : x \geq 0, \sum_i x_i = 1\}$, Justify the following recipe for computing $\mathrm{Proj}_M(x)$:

   > Let $\psi(t) = \sum_{i=1}^m [x_i - t]_+$. where $[s]_+ = \max[s, 0]$. ψ is piecewise linear, with breakpoints $x_1, x_2, \ldots, x_n$, continuous function of $t \in \mathbf{R}$. $\psi(t) \to +\infty$ as $t \to -\infty$, and $\psi(t) \to 0$ as $t \to +\infty$. Consequently, there exists (and can be easily computed due to piecewise linearity of ψ) $t \in \mathbf{R}$ such that $\sum_i [x_i - t]_+ = 1$. The metric projection of x onto M is nothing but the vector x_+ with coordinates $[x_i - t]_+$, $1 \leq i \leq n$.

   What is metric projection of the point $x = [1; 2; 2.5]$ on the 3-dimensional probabilistic simplex?

Solution:

1: Let $d = \inf_{y \in M} \|x - y\|_2$, so that there exists a sequence $\{y_i \in M\}_{i \geq 1}$ such that $\lim_{i \to \infty} \|x - y_i\|_2 = d$. Since $d < \infty$, the sequence $\{y_i\}$ is bounded, so that we can extract from it a converging subsequence $\{y_{i_s}, i_s < i_{s+1}\}_{s \geq 1}$. Since $y_{i_s} \in M$, the limit $\bar{y}$ of the subsequence belongs to M, and since $\|y_{i_s} - x\|_2 \to d$ as $s \to \infty$ and $\|\cdot\|_2$ is continuous, we conclude that $\|\bar{y} - x\|_2 = d$. Thus, the minimum in $(*)$ is achieved. Now let us prove that the closest to x points of M are exaclty the points satisfying (8.1), Note that when $x_+ \in M$ and $y \in M$, we have $x_+ + t(y - x) \in M$ when $0 \leq t \leq 1$ due to convexity of M. It follows that if x_+ is a minimizer of $\|z - y\|_2$ over $y \in M$, then the function $\phi(t) = \|x - [x_+ + t(y - x_+)]\|_2^2$ attains its minimum on the segment $0 \leq t \leq 1$ at $t = 0$. We have

$$\phi(t) = \|x - x_+\|_2^2 - 2t[x - x_+]^\top [y - x_+] + t^2 \|y - x_+\|_2^2;$$

since this smooth function achieves its minimum on $[0, 1]$ at the point $t = 0$, we have $\phi'(0) \geq 0$, which is the inequality in (8.1). As a byproduct, we see that $\|y - x\|_2^2 = \phi(1) \geq \phi(0) + \|y - x_+\|_2^2 = \|x - x_+\|_2^2 + \|y - x_+\|_2^2$, implying that if $y \in M$ and $y \neq x_+$, then $\|x - y\|_2 > \|x - x_+\|_2$, that is, x_+ is the unique minimizer of $\|x - y\|_2^2$ over $y \in M$. It remains to prove that if x_+ satisfies (8.1), then x_+ minimizes $\|y - x\|_2^2$ over $y \in M$. Indeed, assuming that x_+ satisfies (8.1) and given $y \in M$, the associated with this y function $\phi(t)$ is quadratic in t and satisfies $\phi(0) \geq 0$, $\phi'(0) \geq 0$, $\phi'' \geq 0$, implying that $\phi(0) \leq \phi(t)$ whenever $t \geq 0$; in particular, $\|x_+ - x\|_2^2 = \phi(0) \leq \phi(1) = \|y - x\|_2^2$. Thus, $\|x_+ - x\|_2^2 \leq \|y - x\|_2^2$ for all $y \in M$ and, in addition, $x_+ \in M$, implying that x_+ minimizes $\|y - x\|_2^2$ over $y \in M$. ■

2: Let $x \in \mathbf{R}^n$, $x_+ = \mathrm{Proj}_M(x)$, and $y \in M$. We have

$$\begin{array}{rcl} \|x - y\|_2^2 & = & \|[[x - x_+] + [x_+ - y]\|_2^2 = \|x - x_+\|_2^2 + \|x_+ - y\|_2^2 + 2[x - x_+]^\top [x_+ - y] \\ & \geq & \|x - x_+\|_2^2 + \|x_+ - y\|_2^2, \end{array}$$

where the concluding $\geq$ is due to (8.1).
3: Assuming $x \notin M$, for $y \in M$ we have

$$[x - x_+]^\top [x - y] = [x - x_+]^\top [x - x_+] + [x - x_+]^\top [x_+ - y] \geq \|x - x_+\|_2^2$$

with the inequality given by (8.1). Thus, $[x - x_+]^\top y \leq [x - x_+]^\top x - \|x - x_+\|_2^2$. Recalling what e is, we get $e^\top x \geq e^\top y + \|x - x_+\|_2 = e^\top y + \mathrm{dist}(x, M) \, \forall y \in M$. ■
4: Let $u_+ = \mathrm{Proj}_M(u)$, $u'_+ = \mathrm{Proj}_M(u')$. Let us set $e = u - u_+$, $f = u'_+ - u'$, so that $[u_+ - u'_+] + [e + f] = u - u'$ and $e^\top [u_+ - u'_+] \geq 0$, $f^\top [u_+ - u'_+] \geq 0$ by (8.1) as applied with $x = u$ and with $x = u'$. We conclude that $\|u - u'\|_2^2 = \|[u_+ - u'_+] + [e + f]\|_2^2 = \|u_+ - u'_+\|_2^2 + 2[e + f]^\top [u_+ - u'_+] + \|e + f\|_2^2 \geq \|u_+ - u'_+\|_2^2$. ■
5: Invoking item 1, all we need is to verify that with x_+ given by the construction in question, (8.1) holds true. Indeed, the inclusion $x_+ \in M$ is evident. Besides this,

$$\begin{array}{ll} \forall y \in M: & [x - x_+]^\top [x_+ - y] = -\sum_{i: x_i \leq t} x_i y_i + \sum_{i: x_i > t} t([x_i - t] - y_i) \\ & = -\sum_i \min[x_i, t] y_i + t \sum_{i: x_i > t} [x_i - t] = t - \sum_i \min[x_i, t] y_i \geq t - t \sum_i y_i = 0, \end{array}$$

where $\geq$ is due to nonnegativity of $y \in M$.
The metric projection of $[1; 2; 2.5]$ on 3-dimensional probabilistic simplex is the vector $[0; 0.25; 0.75] = [[1 - 1.75]_+; [2 - 1.75]_+; [2.5 - 1.75]_+]$. ■

Exercise II.5. [Follow-up to Exercise II.4] Let $p(z) = z^n + p_{n-1} z^{n-1} + ... + p_1 z + p_0$, $n \geq 1$ be a polynomial of complex variable z. By the Fundamental Theorem of Algebra, p has n roots $\lambda_1, ..., \lambda_n$. Treating complex numbers as 2D real vectors, prove that all roots of the derivative $p'(z) = n z^{n-1} + (n-1) p_{n-1} z^{n-2} + .. + p_1$ belong to the convex hull of $\lambda_1, ..., \lambda_n$.

Solution: Let $C = \mathrm{Conv}\{\lambda_1, ..., \lambda_n\}$, and let λ be a root of p'. Assuming that $\lambda \notin C$, let us lead this assumption to contradiction. Indeed, let $\overline{\lambda} = \mathrm{Proj}_C(\lambda)$ and $e = \lambda - \overline{\lambda}$, so that $e^\top [\lambda - \lambda_i] \geq e^\top e > 0$ by Exercise II.4,3. We have $p(z) = \prod_i (z - \lambda_i)$, whence, setting $f(z) = |p(z)|^2 = \prod_i \|z - \lambda_i\|_2^2 : \mathbf{R}^2 \to \mathbf{R}^2$, one has

$$\begin{array}{l} \frac{d}{dt}\big|_{t=0} f(\lambda + te) = 2 \Big[e^T [\lambda - \lambda_1] \|\lambda - \lambda_2\|^2 ... \|\lambda - \lambda_n\|^2 + \|\lambda - \lambda_1\|_2 e^T [\lambda - \lambda_1] \|\lambda - \lambda_3\|^2 ... \|\lambda - \lambda_n\|^2 \\ + ... + \|\lambda - \lambda_1^2\| ... \|\lambda - \lambda_{n-1}\|_2^2 e^\top [\lambda - \lambda_n] \Big] > 0. \end{array}$$

On the other hand, we have

$$0 = p'(\lambda) = \lim_{\delta \to 0} \frac{p(\lambda + \delta) - p(\lambda)}{\delta},$$

(why?). Denoting by $\imath$ the imaginary unit and setting $\lambda = a + \imath b$, $p(x + \imath y) = u(x, y) + \imath v(x, y)$ with real a, b, x, y, u, v, and looking what happens when $\delta \to 0$ stays (a) real, (b) purely imaginary, we get

$$\frac{\partial}{\partial x} u(a, b) = 0, \frac{\partial}{\partial x} v(a, b) = 0 \text{ and } \frac{\partial}{\partial y} u(a, b) = 0, \frac{\partial}{\partial y} v(a, b) = 0,$$

whence

$$\nabla\big|_{x=a, y=b} f(x + \imath y) = \nabla\big|_{x=a, y=b} [u^2(x, y) + v^2(x, y)] = 0,$$

so that $\frac{d}{dt}\big|_{t=0} f(\lambda + te) = 0$, which is a desired contradiction. ■

Exercise II.6. Derive the statement in Remark I.1.4 from the Separation Theorem.

Solution: We already know that the solution set of a whatever system of nonstrict linear inequalities is closed and convex, and all we need to prove is that a closed convex set $M \subset \mathbf{R}^n$ is a solution set of a sequence of nonstrict linear inequalities $a_i^\top x \leq b_i$, $i = 1, 2, ...$. There is nothing to prove when $M = \mathbf{R}^n$ (take empty system, or, if you want, single inequality $0^\top x \leq -1$). Similarly, there is nothing to prove when $M = \varnothing$ – take the system of inequalities $x_1 \leq -1, -x_1 \leq -1$. Now let M be nonempty and smaller than $\mathbf{R}^n$. The complement M^c of C is a nonempty open set; note that the set of all rational vectors from M^c can be arranged into sequence $c_1, c_2, ...$.

Indeed, let us look at the set T_N of all rational vectors from M^c with the total of magnitudes of numerators and denominators in representations of their (rational!) coordinates as fractions does not exceed a given integer N; for every N, this set is finite. We now can list all vectors from T_1, then list all unlisted yet vectors from T_2, then – all unlisted yet vectors from T_3, and so on; as a result, all rational vectors from M^c will be arranged into a sequence.

Now let $r(x) = \min_{y\in M} \|x-y\|_2$ be the distance from $x \in \mathbf{R}^n$ to M; since M is closed and nonempty, the minimum is achieved. Again invoking closedness of M, $r(x) > 0$ whenever $x \notin M$. Besides this, the function $r(x)$ clearly satisfies the relation $|r(x)-r(x')| \le \|x-x'\|_2$ and is therefore continuous. Note also that when $x \notin M$, the open ball $B(x)$ of radius $r(x)$ centered at x does not intersect M. By Separation Theorem, the balls $B(c_i)$ can be separated from M: for properly selected a_i we have $\sup_{x\in M} a_i^\top x \le \inf_{y\in B(c_i)} a_i^\top y$. We lose nothing by scaling a_i to become a unit vector, in which case the "separation inequality" becomes $\sup_{x\in M} a_i^\top x \le b_i := a_i^\top c_i - r(c_i)$. We claim that M is exactly the solution set of the resulting sequence of inequalities $a_i^\top x \le b_i$, $i = 1, 2, \ldots$. Indeed, by construction, every point from M solves this system. All we need to verify is that if $\bar{x}$ solves the system, then $\bar{x} \in M$. Assuming, on the contrary, that this is not the case, $\bar{x} \in M^c$ and therefore for some sequence $i_1 < i_2 < \ldots$ we have $c_{i_j} \to \bar{x}$ as $j \to \infty$, whence $a_{i_j}^\top (c_{i_j} - \bar{x}) \to 0$ as $j \to \infty$. Due to the origin of $\bar{x}$, $a_{i_j}^\top \bar{x} \le b_{i_j} = a_{i_j}^\top c_{i_j} - r(c_{i_j})$, whence $a_{i_j}^\top (c_{i_j} - \bar{x}) \ge r(c_{i_j})$, which combines with $a_{i_j}^\top (c_{i_j} - \bar{x}) \to 0$ as $j \to \infty$ to imply that $r(c_{i_j}) \to 0$ as $j \to \infty$. On the other hand, $r(\cdot)$ is continuous and $c_{i_j} \to \bar{x}$, $j \to \infty$, implying that $r(c_{i_j}) \to r(\bar{x})$ as $j \to \infty$. The bottom line is that $r(\bar{x}) = 0$, which is not the case, since $\bar{x} \notin M$ and M is closed. Thus, assuming that M is not the solution set of the system $a_i^\top x \le b_i$, $i = 1, 2, \ldots$, we arrive at contradiction. ■

8.2 Extreme points

Exercise II.7. Find extreme points of the following sets:

1. $X = \{x \in \mathbf{R}^3 : x_1 + x_2 \le 1, x_2 + x_3 \le 1, x_3 + x_1 \le 1\}$
2. $X = \{x \in \mathbf{R}^4 : x_1 + x_2 \le 1, x_2 + x_3 \le 1, x_3 + x_4 \le 1, x_4 + x_1 \le 1\}$

Solution: 1: The set is polyhedral; by algebraic characterization of extreme point of polyhedral sets, among the inequalities specifying the set, an extreme point w, if any, should make equalities 3 inequalities with linearly independent vectors of coefficients, that is, w should make equalities all 3 constraints specifying the set (their vectors of coefficients indeed are linearly independent). As a result, the only extreme point is $[0.5; 0.5; 0.5]$.

2: The same reasoning as in item 1 says that at an extreme point all constraints specifying the set should be satisfied as equalities, and the vectors of coefficients of these constraints should be linearly independent. The latter does *not* take place, so that there are no extreme points.

Explanation: when $n \ge 2$, the $n \times n$ matrix $A_n = \begin{bmatrix} 1 & 1 & & \\ & 1 & 1 & \\ & & \ddots & \ddots \\ & & & 1 \\ 1 & & & 1 \end{bmatrix}$ is nondegenerate when n is odd and is degenerate, with kernel spanned by the vector $[1; -1; 1; -1; ...; -1]$ when n is even. As a result, the set $X_n = \{x \in \mathbf{R}^n : A_n x \le [1; ...; 1]\}$ contains lines, and thus has no extreme point, when n is even, and has exactly one extreme point $[0.5; ...; 0.5]$ when n is odd. For odd n, $x \mapsto y = A_n x$ is a linear one-to-0one transformation of $\mathbf{R}^n$, and in y-variables X_n becomes the set $\{y \in \mathbf{R}^n : y \le [1; ...; 1]\}$. Thus, for odd n, X_n is just a translation of a polyhedral cone – the image of $\mathbf{R}^n_+$ under one-to-one linear transformation.

Exercise II.8. Let $M \subset \mathbf{R}^n$ be a nonempty closed convex set not containing lines, and $f^\top x$ be a linear function of $x \in \mathbf{R}^n$ achieving its maximum over X. Prove that among maximizers of this function on M there are extreme points of M.

Solution: Let $\overline{M} = \mathrm{Argmax}_{x\in M} f^\top x$ be the set of maximizers of $f^\top x$ over $x \in M$. By assumption, this set is nonempty; along with M, it is convex, closed, and does not contain lines. By item (i) of Krein-Milman Theorem $\overline{M}$ has an extreme point x_*; let us prove that x_* is an extreme point of M. Indeed, assuming $x \pm h \in M$, we should have $f^\top [x_* \pm h] \le f^\top x_*$ (since x_* is a maximizer of $f^\top x$ over

$x \in M$) which is possible only when $f^\top[x_* \pm h] = f^\top x_*$, Thus, $x_* \pm h \in \overline{M}$, implying that $h = 0$ (since $x_* \in \mathrm{Ext}(\overline{M})$). Thus, x_* is a desired extreme point maximizer of $f^\top x$ over $x \in M$. ■

Exercise II.9. Mark by **T** those of the below claims which always (i.e., for every data satisfying premise of the claim) are true:

1. If $\mathrm{Conv}(A) = \mathrm{Conv}(B)$, then $A = B$.

   Solution: evidently false – take $n = 1$, $A = \{0, 1, 2\}$, $B = \{0, 2\}$.

2. If $\mathrm{Conv}(A) = \mathrm{Conv}(B)$ is nonempty and $A, B, \mathrm{Conv}(A)$ are closed, then $A \cap B \neq \varnothing$.

   Solution: false – take $n = 1$, $A = \{2k+1\}_{k=-\infty}^{\infty}$, $B = \{2k\}_{k=-\infty}^{\infty}$.

3. If $\mathrm{Conv}(A) = \mathrm{Conv}(B)$ is nonempty and bounded, then $A \cap B \neq \varnothing$.

   Solution: false; take $A = \{\frac{1}{2k}\}_{k=1}^{\infty} \cup \{1 - \frac{1}{2k}\}_{k=1}^{\infty}$, $B = \{\frac{1}{2k+1}\}_{k=1}^{\infty} \cup \{1 - \frac{1}{2k+1}\}_{k=1}^{\infty}$, so that $A \cap B = \varnothing$ and $\mathrm{Conv}(A) = \mathrm{Conv}(B) = (0, 1)$.

4. If $\mathrm{Conv}(A) = \mathrm{Conv}(B)$ is nonempty, closed and bounded, then $A \cap B \neq \varnothing$.

   Solution: true. When $\mathrm{Conv}(A)$ is nonempty, closed, and bounded, by Krein-Milman Theorem, $\mathrm{Conv}(A)$ possesses an extreme point v, and by Fact II.6.10 $v \in A$. By the same token, $\mathrm{Conv}(A) = \mathrm{Conv}(B)$ implies that $v \in B$, so that $A \cap B$ is nonempty and, moreover, contains all extreme points of $\mathrm{Conv}(A) = \mathrm{Conv}(B)$. Applying Krein-Milman Theorem once more, we conclude that $A \cap B$ is not just nonempty, it is rich enough to ensure that $\mathrm{Conv}(A \cap B) = \mathrm{Conv}(A) = \mathrm{Conv}(B)$.

Exercise II.10. As is immediately seen, the only extreme point of the nonnegative orthant $\mathbf{R}^n_+ = \mathbf{R}_+ \times \mathbf{R}_+ \times \ldots \times \mathbf{R}_+$ is the origin, that is, the vector from $\{0\} \times \{0\} \times \ldots \times \{0\}$; as we know, the extreme points of n-dimensional unit box $\{x \in \mathbf{R}^n : 0 \leq x_i \leq 1, i \leq n\} = [0, 1] \times [0, 1] \times \ldots \times [0, 1]$ are zero/one vectors, that is, vectors from $\{0, 1\} \times \{0, 1\} \times \ldots \times \{0, 1\}$. Prove the following generalization of these observations:

> Let $X_i \subset \mathbf{R}^{n_i}$, $1 \leq i \leq K$, be closed convex sets. The set of extreme points of the direct product $X = X_1 \times \ldots \times X_K$ of these sets is the direct product of the sets of extreme points of X_i.

Solution: The vectors from X are the block vectors $x = [x_1; \ldots; x_K]$ with blocks $x_i \in X_i$. If such an x is an extreme point of X, that is, $x \pm h \in X$ implies $h = 0$, then for every i the relation $x_i \pm h_i \in X_i$ implies $h_i = 0$, since otherwise, setting $h = [0; ...; 0; h_i; 0; ...; 0]$ we would have $x \pm h \in X$ and $h \neq 0$, which is impossible; thus, $x \in \mathrm{Ext}(X_1) \times ... \times \mathrm{Ext}(X_K)$. Vice versa, if $x \in \mathrm{Ext}(X_1) \times ... \times \mathrm{Ext}(X_K)$ and $x \pm h \in X$, then $x_i \pm h_i \in X_i$ for all i, implying that $h_i = 0$ for all i, that is, $h = 0$; thus, $x \in \mathrm{Ext}(X)$. ■

Exercise II.11. Looking at the sets of extreme points of closed convex sets like the unit Euclidean ball, a polytope, the paraboloid $\{[x; t] : t \geq x^\top x\}$, etc., we see that these sets are closed. Do you think this always is the case? Is it true that the set $\mathrm{Ext}(M)$ of extreme points of a closed convex set M always is closed ?

Solution: The claim is not true. Indeed, consider the set X in 3D which is the union of the segment $\{[x_1; 0; 0] : -1 \leq x_1 \leq 1\}$ and the arc $\{[0; x_2; x_2^2], 0 \leq x_2 \leq 1\}$; this set is closed and bounded, and therefore so is its convex hull $M := \mathrm{Conv}(X)$ (Corollary I.2.5). We claim that when $t \in (0, 1)$, the point $x_t = [0; t; t^2]$ is an extreme point of M. Taking this claim for granted, we conclude that the point $[0; 0; 0]$ is the limit, as $t \to +0$, of extreme points x_t of M, but this limit clearly is not an extreme point – it is the midpoint of the segment with the endpoints $[\pm 1; 0; 0] \in M$.

It remains to prove that x_t is an extreme point of M. Observe first that x_t is an extreme point of the projection M_- of M onto the plane $L = \{x : x_1 = 0\} \ni x_t$. Indeed, M_- clearly belongs to the convex hull C of the projection of X onto L, that is, to the convex hull of the arc $\{[0; s; s^2] : 0 \leq s \leq 1\}$. We clearly have $C = \{x : x_1 = 0, 0 \leq x_2 \leq 1, x_2^2 \leq x_3 \leq x_2\}$, and x_t is an extreme point of C. Since this point belongs to $M_- \subset C$, it is extreme point of M_- as well. Now, to prove that $x_t \in \mathrm{Ext}(M)$ is the same as to prove that $x_t \pm h \in M$ implies $h = 0$. Indeed, let h be such that $x_t \pm h \in M$. Looking at the projections of $x_t \pm h$ onto L and taking into account that x_t is an extreme point of the projection of M onto L, we see that $h_2 = h_3 = 0$. Thus, we are in the situation $[\pm h_1; t; t^2] \in M$, and should prove that

$h_1 = 0$. Recalling what M is, we conclude that $[h_1; t; t^2]$ is convex combination of several points of the type $[s; 0; 0]$, $s \in [-1, 1]$, and several points of the type $[0; r; r^2]$ with $0 \le r \le 1$. If the total weight of the points of the first type in this combination is positive, then, projecting the combination onto L, we conclude that x_t is a convex combination of several points of the second type and the point $[0; 0; 0]$, the weight of the latter point being positive. This is impossible – all points participating in the latter convex combination belong to C, and, as we know, x_t is an extreme point of this set, implying that all points participating, with positive weights, in representation of x_t as a convex combination of points from C should be equal to x_t, see Fact II.6.9. The bottom line is that $[h_1; t; t^2]$ can be represented as a convex combination of points of the second type only, that is, $h_1 = 0$, that is, $h = 0$, as claimed. ■

Exercise II.12. Derive representation $(*)$ in Exercise I.29 from Example II.7.1 in section 7.1.1.

Solution: Given positive integers $k \le n$ and $x \in \mathbf{R}^n$, consider the LP program

$$\mathrm{Opt} = \max_u \left\{ \sum_i x_i u_i : 0 \le u_i \le 1, i \le n, \sum_i u_i = k \right\}$$

Example II.7.1 in section 7.1.1 says that extreme points of the bounded feasible set of the problem are 0/1 vectors with exactly k entries equal to 1, implying that $\mathrm{Opt} = s_k(x)$. We now have

$$\begin{array}{rcl} s_k(x) & = & \max_u \{ \sum_i x_i u_i : 0 \le u_i \le 1, i \le n, \sum_i u_i = k \} \\ & = & \min_{z^\pm, s} \left\{ \sum_i z_i^+ + ks : [z_i^+ - z_i^-] + s = x_i, i \le n, z^\pm \ge 0 \right\} \text{ [LP duality]} \\ & = & \min_{z^+, s} \left\{ \sum_i z_i^+ + ks : z^+ \ge 0, x_i \le z_i^+ + s, i \le n \right\}, \end{array}$$

or, equivalently,

$$t \ge s_k(x) \Longleftrightarrow \exists (z, s) : x_i \le z_i + s \, \forall i, z \ge 0, \sum_i z_i + ks \le t,$$

which is equivalent form of the representation we are justifying. ■

Exercise II.13. By Birkhoff Theorem, the extreme points of the polytope $\Pi_n = \{[x_{ij}] \in \mathbf{R}^{n \times n} : x_{ij} \ge 0, \sum_i x_{ij} = 1 \, \forall j, \sum_j x_{ij} = 1 \, \forall i\}$ are exactly the Boolean (i.e., with entries 0 and 1) matrices from this set. Prove that the same holds true for the "polytope of sub-doubly stochastic" matrices $\Pi_{m,n} = \{[x_{ij}] \in \mathbf{R}^{m \times n} : x_{ij} \ge 0, \sum_i x_{ij} \le 1 \, \forall j, \sum_j x_{ij} \le 1 \, \forall i\}$.

Solution: First, every Boolean matrix $[x_{ij}]$ from $\Pi_{m,n}$ is extreme point. Indeed, we know that every Boolean matrix is extreme point of the box $B_{m,n} = \{[x_{ij}] \in \mathbf{R}^{m \times n} : 0 \le x_{ij} \le 1 \, \forall i, j\}$, and it remains to refer to the evident fact: *When $Y \subset X$ is a nested pair of convex sets, then every extreme point v of X which happens to be in Y is an extreme point of Y.* Indeed, were v the midpoint of a nontrivial segment in Y, it would be the midpoint of a nontrivial segment in X, which is not the case.

Given that the set B of all Boolean matrices from $\Pi_{m,n}$ belongs to $\mathrm{Ext}(\Pi_{m,n})$, all we need to conclude that $B = \mathrm{Ext}(\Pi_{m,n})$ is to show that $\Pi_{m,n} = \mathrm{Conv}(B)$ (see Fact II.6.10). Our plan is as follows: given a matrix $x \in \Pi_{m,n}$, we will show that x can be made a North-Western $m \times n$ submatrix of $k \times k$ doubly stochastic matrix $\overline{x}$, with properly selected k. This is all we need: by Birkhoff Theorem, $\overline{x}$ is convex combination of $k \times k$ permutation matrices, implying that x is a convex combination of the $m \times n$ North-Western submatrices of these permutation matrices, and these submatrices clearly are Boolean matrices from $\Pi_{m,n}$.

Thus, let a matrix $x \in \Pi_{m,n}$ be given; we want to extend it by adding several rows and columns to a larger doubly stochastic matrix. First of all, by adding to x $n - m$ zero rows (if $n > m$) or $m - n$ zero columns (if $m > n$), we can reduce the situation to the one where $m = n$, which we assume from now on. Next, let $S = \sum_{i,j=1}^n x_{ij}$. Note that since the row sums in x are ≤ 1, we have $S \le n$, so that $\kappa := n - S$ is nonnegative; let d be the smallest integer which is $\ge \kappa$. This is how we can embed x, as the North-Western $n \times n$ submatrix, into $(n + d) \times (n + d)$ doubly stochastic matrix. Denote by r_i, $i \le n$, the sum of entries in i-th row of x, and by c_j, $j \le n$, the sum of entries in j-th column of x. Note that $0 \le r_i \le 1$, $0 \le c_j \le 1$ and $\sum_i r_i = \sum_j c_j = S$. Let also $\rho_i = 1 - r_i$, $\sigma_j = 1 - c_j$, so that $\rho_i \ge 0$, $\sigma_j \ge 0$, and $\sum_i \rho_i = \sum_j \sigma_j = \kappa$. Now let $\rho = [\rho_1/d; \rho_2/d; ...; \rho_n/d]$ and $\sigma = [\sigma_1/d; \sigma_2/d; ...; \sigma_n/d]$, so that ρ and σ are nonnegative vectors with sums of entries equal to $\kappa/d \le 1$. Setting $\theta = (1 - \kappa/d)/d$

and specifying $\overline{x}$ as the $(n+d)\times(n+d)$ matrix $\left[\begin{array}{c|c|c|c} x & \rho & \cdots & \rho \\ \hline \sigma^\top & \theta & \cdots & \theta \\ \hline \vdots & \vdots & \ddots & \\ \hline \sigma^\top & \theta & \cdots & \theta \end{array}\right]$, we, as is immediately seen, get a doubly stochastic matrix, and this is the desired doubly stochastic extension of x. ■

Exercise II.14. [Follow-up to Exercise II.13] Let m, n be two positive integers with $m \le n$, and $X_{m,n}$ be the set of $m \times n$ matrices $[x_{ij}]$ with $\sum_i |x_{ij}| \le 1$ for all $j \le n$ and $\sum_j |x_{ij}| \le 1$ for all $i \le m$. Describe the set $\mathrm{Ext}(X_{m,n})$. To get an educated guess, look at the matrices $\left[\begin{array}{ccc} 1 & 0 & 0 \\ 0 & 0 & -1 \end{array}\right]$, $\left[\begin{array}{ccc} 0 & 0 & 0 \\ 0 & 0 & -1 \end{array}\right]$, $\left[\begin{array}{ccc} 0.5 & -0.5 & 0 \\ -0.5 & 0.5 & 0 \end{array}\right]$ from $X_{2,3}$.

Solution: $\mathrm{Ext}(X_{m,n})$ is the set of all $m \times n$ matrices with entries $-1, 0, 1$ such that in every row there is exactly one nonzero entry, and in every column there is at most one nonzero entry.

In one direction: Let $x = [x_{ij}]$ be $m \times n$ matrix with entries $-1, 0, 1$, at most one nonzero entry per column, and exactly one nonzero entry per row; let us prove that $x \in \mathrm{Ext}(X_{m,n})$. First, x clearly belongs to $X_{m,n}$. It remains to prove that if $x \pm h \in X_{m,n}$, then $h = 0$. Indeed, in the situation in question, denoting $\sigma(i)$ the index j of the column with $x_{ij} \neq 0$ (for our x, such j exists for every $i \le m$), we should have $\sum_j |x_{ij} \pm h_{ij}| \le 1$. In particular, $|x_{i\sigma(i)} \pm h_{i\sigma(i)}| \le 1$, implying, in view of $|x_{i\sigma(i)}| = 1$ (all nonzero entries in our x are of magnitude 1!) that $h_{i\sigma(i)} = 0$. Therefore $1 \ge \sum_j |x_{ij} \pm h_{ij}| = \underbrace{|x_{i\sigma(i)}|}_{=1} + \sum_{j \neq \sigma(i)} |x_{ij} \pm h_{ij}|$, implying that $h_{ij} = 0$ for $j \neq \sigma(i)$. Thus, i-th row in h is zero; since $i \le m$ is arbitrary, $h = 0$, as required.

In the opposite direction: Let $x \in \mathrm{Ext}(X_{m,n})$, and let us prove that x has all entries in $\{-1, 0, 1\}$, with exactly one nonzero entry per row and at most one nonzero entry per column. Let $\xi_{ij} \in \{-1, 1\}$ be such that $\xi_{ij} x_{ij} = |x_{ij}|$ for all i, j, and let Ξ be the one-to-one linear transformation of the space $\mathbf{R}^{m \times n}$ of $m \times n$ matrices given by entrywise multiplication of a matrix by the matrix $[\xi_{ij}]$. Linear one-to-one transformation Ξ maps the polytope $X_{m,n}$ onto itself and thus maps onto itself the set $\mathrm{Ext}(X_{m,n})$. In particular, the matrix $\overline{x} = [|x_{ij}|]$ composed of magnitudes of entries in x (this is the image of x under the mapping Ξ) is an extreme point of $X_{m,n}$. Note that $\overline{x} \in \Pi_{m,n}$, where $\Pi_{m,n}$ is the polytope of entrywise nonnegative $m \times n$ matrices with all column and row sums not exceeding 1. Moreover, $\overline{x}$ is an extreme point of $\Pi_{m,n}$, since from $\overline{x} \pm h \in \Pi_{m,n}$ it clearly follows $\overline{x} \pm h \in X_{m,n}$, and the latter implies that $h = 0$ – $\overline{x}$ is an extreme point of $X_{m,n}$! By the result of Exercise II.13, $\overline{x}$ has entries 0 and 1 only, implying that all nonzero entries in x are ± 1. With this in mind, $\sum_i |x_{ij}| \le 1$, $j \le n$, implies that x has at most one nonzero entry per column, and $\sum_j |x_{ij}| \le 1$, $i \le m$, implies that x has at most one nonzero entry per row. It remains to verify that every row of x has a nonzero entry. Assume the opposite, say, that the first row of x is zero, and let us lead this assumption to a contradiction. In the case in question x has at most $m - 1$ nonzero entries (since, as we have already seen, there is at most one nonzero entry per row, and the first row is zero). Consequently, among $n > m - 1$ columns of x there is a zero column, w.l.o.g. let it be the first one. Thus, x has zero first column and zero first row, which combines with $x \in X_{m,n}$ to imply that when h is $m \times n$ matrix with the only nonzero entry, equal to 1, in the cell $1, 1$, we have $x \pm h \in X_{m,n}$, contradicting x being an extreme point of $X_{m,n}$. ■

Exercise II.15. [follow-up to Exercise II.13] Let x be an $n \times n$ entrywise nonnegative matrix with all row and all column sums ≤ 1. Is it true that for some doubly stochastic matrix $\overline{x}$, the matrix $\overline{x} - x$ is entrywise nonnegative?

Solution: Yes. By the result of Exercise II.13, x is a convex combination of Boolean matrices with column and row sums ≤ 1. Every matrix with the latter property clearly is obtained from appropriate permutation matrix by replacing with zeros some of the unit entries. Thus, every Boolean matrix with row and column sums ≤ 1 is entrywise $\le$ a permutation matrix, and therefore a convex combination of the matrices of the former class is entrywise $\le$ a convex combination of permutation matrices, which is a doubly stochastic matrix. ■

Exercise II.16. [Assignment problem] Consider the problem as follows:

There are n jobs and n workers. When worker j is assigned to job i, we get profit c_{ij}. We want to assign every worker to a job in such a way that every worker is assigned to exactly one job and every job is assigned to exactly one worker. Under this restriction, we want to maximize the total profit.

1. Pose the Assignment problem as the Boolean (i.e., with the decision variables restricted to be zeros and ones) Linear Programming problem.

   Solution: Encoding a candidate assignment by $n \times n$ matrix $x = [x_{ij}]$ with $x_{ij} = 1$ when job i is assigned to worker j and $x_{ij} = 0$ otherwise, we end up with the problem

   $$\max_x \left\{ \sum_{i,j} c_{ij} x_{ij} : x_{ij} \geq 0, \sum_j x_{ij} = 1 \, \forall i, \sum_i x_{ij} = 1 \, \forall j, x_{ij} \in \{0,1\} \right\} \qquad (!)$$

2. Think how to solve the problem from item 1 via plain Linear Programming

   Solution: Removing in (!) the Boolean constraints $x_{ij} \in \{0,1\}$, we arrive at the LP problem of maximizing a linear form over the polytope of doubly stochastic $n \times n$ matrices. The problem clearly is solvable, and among its optimal solutions there are extreme points of the polytope. By Birkhoff Theorem, these extreme points are permutation matrices. Thus, passing from (!) to the LP relaxation of the problem, we preserve the optimal value, and every LP algorithm which produces extreme point solutions will, as applied to relaxation, provide us with an optimal solution to (!).

3. [computational study] Consider the special case of Assignment problem where all profits c_{ij} are zeros or ones; you can interpret $c_{ij} = 1/0$ as the fact that worker j knows/does not know how to execute job j. In this situation Assignment problem requires from us to find an assignment which maximizes the total number of executed jobs. Assume now that the matrix $C = [c_{ij}]$ is generated at random, with entries taking, independently of each other, value 1 with probability $\epsilon \in (0,1)$ and value 0 with probability $1 - \epsilon$. For $n \in \{4, 8, 16, 32, 64, 128, 256\}$ and $\epsilon \in \{1/2, 1/4, 1/8, 1/16\}$, run 100 simulations per pair n, ϵ to find the empirical mean of the ratio "number of executed jobs in optimal assignment"$/n$ and look at the results.

   Solution: Our results are as follows:

	$n=4$	$n=8$	$n=16$	$n=32$	$n=64$	$n=128$	$n=256$
$\epsilon = 0.5000$	0.8800	0.9862	1.0000	1.0000	1.0000	1.0000	1.0000
$\epsilon = 0.2500$	0.6025	0.8187	0.9769	1.0000	1.0000	1.0000	1.0000
$\epsilon = 0.1250$	0.3325	0.5650	0.8094	0.9719	0.9995	1.0000	1.0000
$\epsilon = 0.0625$	0.2250	0.3688	0.5387	0.7906	0.9723	0.9993	1.0000

   The results allow to make an educated guess that with ϵ fixed and $n \to \infty$, the probability to get *all* n *jobs executed* in the optimal assignment goes to 1; this guess happens to be true.

Exercise II.17. Let $\nu = (\nu_1, ..., \nu_K)$ with positive integer ν_i, and let $\mathbf{S}^\nu = \mathbf{S}^{\nu_1} \times ... \times \mathbf{S}^{\nu_K}$ be the space of block-diagonal, with K diagonal blocks of sizes $\nu_i \times \nu_i$, $i \leq K$, symmetric matrices, let $\mathbf{S}^\nu_+$ be the cone composed of positive semidefinite matrices from $\mathbf{S}^\nu$, and let E be an m-dimensional affine plane in $\mathbf{S}^\nu$ which intersects $\mathbf{S}^\nu_+$. The intersection $X = E \cap \mathbf{S}^\nu_+$ is a closed nonempty convex set not containing lines and thus possessing extreme points. Let W be such a point, W^{ii} be the diagonal blocks of W, and r_i be the ranks of $\nu_i \times \nu_i$ matrices W^{ii}. Prove that

$$\sum_{i=1}^k r_i(r_i+1) \leq \sum_{i=1}^K \nu_i(\nu_i+1) - 2m.$$

What happens in the diagonal case $\nu_1 = ... = \nu_K = 1$?

Solution: Let $W^{ii} = U_i \Lambda_i U_i^\top$ be eigenvalue decompositions of W^{ii}; w.l.o.g. we can assume that the first r_i of eigenvalues of W^{ii} are positive, and the remaining eigenvalues are zero. For every collection of K symmetric $r_i \times r_i$ matrices D^i, denoting by $\overline{D}^i$ the $\nu_i \times \nu_i$ matrices obtained by augmenting D^i

with zero rows and columns, and setting $\overline{D} = \mathrm{Diag}\{U_1\overline{D}^1U_1^\top, ..., U_K\overline{D}^KU_K^\top\}$, we get $W \pm t\overline{D} \succeq 0$ for all small positive t. Now let us impose on the matrices D^i the requirement

$$\overline{D} \in L, \tag{!}$$

where L is the parallel to E linear subspace in $\mathbf{S}^\nu$. Assuming that

$$\mathrm{codim}\, L := \sum_i \nu_i(\nu_i+1)/2 - m < R := \sum_i r_i(r_i+1)/2,$$

relation (!), which is a system of $\mathrm{codim}\, L$ homogeneous linear equations on R variables $\{D^i_{pq}, p \le q \le r_i, i \le K\}$, has a nontrivial solution, implying that $W \pm tD \in X$ for some nonzero D and positive t, which is impossible. Thus, $\mathrm{codim}\, L \ge R$, as claimed. ■

In the diagonal case, the result becomes the following fact (perfectly well known to everybody who somehow dealt with the Simplex method in LP): *The number of nonzero entries in any extreme point of the feasible set of a feasible LP problem in the standard form* $\max_{x\in\mathbf{R}^k}\{c^\top x : Ax = b, x \ge 0\}$ *does not exceed the number of equality constraints (i.e., of rows in A).*

Exercise II.18. Let M be a closed convex set in $\mathbf{R}^n$ and $\bar{x}$ be a point of M.

1. Prove that if there exists a linear form $a^\top x$ such that $\bar{x}$ is the *unique* maximizer of the form on M, then $\bar{x}$ is an extreme point of M.

2. Is the inverse of 1) true, i.e., is it true that every extreme point $\bar{x}$ of M is the unique maximizer, over $x \in M$, of a properly selected linear form?

Solution: 1: the answer is positive, Indeed, let $\bar{x}$ be the unique maximizer over $x \in M$ of a linear form $f^\top x$. Assuming, on the contrary to what should be proved, that $\bar{x} \pm h \in M$ for some $h \ne 0$, the linear function $f^\top x$ attains its maximum on the segment $[\bar{x}-h, \bar{x}+h]$ in the midpoint of this segment, which for a linear function is possible only when the function is constant on the segment. Thus, all points on the segment maximize $f^\top x$ over $x \in M$, contradicting the fact that $\bar{x}$ is the unique maximizer of the function on M.

2: The inverse is not true in general. For example, consider the set

$$M = \{(x,y) \in \mathbf{R}^2 : y \ge \left\{ \begin{array}{ll} x^2 & , x \le 0 \\ 0 & , 0 \le x \le 1 \\ (x-1)^2 & , x \ge 1 \end{array} \right\}$$

(draw picture). The origin clearly is an extreme point of the set, but there are no linear forms on $\mathbf{R}^2$ attaining their maximum over M at the origin, and only at it.

Exercise II.19. Identify and justify the correct claims in the following list:

1. Let $X \subset \mathbf{R}^n$ be a nonempty closed convex set, P be an $m \times n$ matrix, and $Y = PX := \{Px : x \in X\} \subset \mathbf{R}^n$. Then
   - For every $x \in \mathrm{Ext}(X)$, $Px \in \mathrm{Ext}(Y)$

     Solution: Wrong – look at the orthogonal projection of the planar triangle with vertices $(0,0)$, $(1,1), (2,0)$ onto the first coordinate axis.
   - Every extreme point of Y is Px for some $x \in \mathrm{Ext}(X)$

     Solution: Wrong – look what happens when X is the stripe $0 \le x \le 1$ on the 2D plane and P is the same as in the solution of item 1.
   - When X does not contain lines, then every extreme point of Y is Px for some $x \in \mathrm{Ext}(X)$.

     Solution: Correct. Indeed, let $w \in \mathrm{Ext}(Y)$. Then the set $X_w = \{x \in X : Px = w\}$ is nonempty, convex, closed and does not contain lines, and thus has an extreme point $\overline{x}$. It suffices to verify that $\overline{x} \in \mathrm{Ext}(X)$. Indeed, let d be such that $\overline{x} \pm d \in X$, and let us prove that $d = 0$. We have $P[\overline{x} \pm d] \in Y$, and since $w = P\overline{x} \in \mathrm{Ext}(Y)$, we get $Pd = 0$, implying that $\overline{x} \pm d \in X_w$; since $\overline{x} \in \mathrm{Ext}(X_w)$, we conclude that $d = 0$.

2. Let X, Y be nonempty closed convex sets in $\mathbf{R}^n$, and let $Z = X + Y$. Then
   - If $w \in \mathrm{Ext}(Z)$, then $w = x + y$ for some $x \in \mathrm{Ext}(X)$ and $y \in \mathrm{Ext}(Y)$.

     Solution: Correct. Indeed, we have $w = x + y$ for some $x \in X, y \in Y$. If $x \notin \mathrm{Ext}(X)$, then $x \pm d \in X$ for some $d \neq 0$, whence $w \pm d = (x \pm d) + y \in Z$, contradicting $w \in \mathrm{Ext}(Z)$. By similar reasoning, $y \in \mathrm{Ext}(Y)$.
   - If $x \in \mathrm{Ext}(X)$, $y \in \mathrm{Ext}(Y)$, then $x + y \in \mathrm{Ext}(Z)$.

     Solution: Wrong – look what happens when $X = [0, 1] \subset \mathbf{R}$ and $Y = [2, 3] \subset \mathbf{R}$.

Exercise II.20. Let $X = \{x \in \mathbf{R}^n : a_i^\top x \leq b_i, i \leq m\}$ be a nonempty polyhedral set and $f^\top x$ be a linear form of $x \in \mathbf{R}^n$ which is bounded above on X:

$$\mathrm{Opt}(f) = \sup_{x \in X} f^\top x < \infty$$

Prove that

1. $\mathrm{Opt}(f)$ is achieved – the set $\operatorname*{Argmax}_{x \in X} f^\top x := \{x \in X : f^\top x = \mathrm{Opt}(f)\}$ is nonempty.

   Solution: This is nothing but the claim that bounded and feasible LP program has a solution (section 3.2.1 or Theorem II.7.12).
2. The set $\operatorname*{Argmax}_{x \in X} f^\top x$ is as follows: there exists an index set $I \subset \{1, 2, ..., m\}$, perhaps empty, such that

   $$\operatorname*{Argmax}_{x \in X} f^\top x = X_I := \{x : a_i^\top x \leq b_i \, \forall i, \, a_i^\top x = b_i \, \forall i \in I\}$$

   Solution: By Linear Programming Duality Theorem, the problem dual to primal problem $\mathrm{Opt}(f) = \max_x \{f^\top x : a_i^\top x \leq b_i, i \leq I\}$ reads $\max_\lambda \left\{\lambda^\top b : \lambda \geq 0, \sum_i \lambda_i a_i = f\right\}$ and is solvable with the optimal value $\mathrm{Opt}(f)$ – the same as the one of the primal problem. Let λ^* be an optimal solution to the dual problem, and let $I = \{i : \lambda_i^* > 0\}$. We claim that $X_* := \operatorname*{Argmax}_{x \in X} f^\top x = X_I$. In one direction: when $x \in X_*$, we have $x \in X$ and

   $$\mathrm{Opt}(f) = f^\top x = [\sum\nolimits_i \lambda_i^* a_i]^\top x = \sum\nolimits_{i \in I} \lambda_i^* [a_i^\top x] \leq \sum\nolimits_{i \in I} \lambda_i^* b_i = b^\top \lambda^* = \mathrm{Opt}(f),$$

   where the inequality is due to $\lambda_i^* \geq 0$, and the last equality – due to the fact that λ^* is optimal solution to the dual problem, and the dual optimal value is $\mathrm{Opt}(f)$. We conclude that the inequality in the chain is equality, so that $\sum_{i \in I} \lambda_i^* [b_i - a_i^\top x] = 0$. The latter relation implies $a_i^\top x = b_i$, $i \in I$ (since $a_i^\top x \leq b_i$ for all i and $\lambda_i^* > 0, i \in I$). In addition, $x \in X$, and we conclude that $x \in X_I$. Thus, $X_* \subset X_I$. Vice versa, if $x \in X_I$, then $x \in X$ and

   $$f^\top x = [\sum\nolimits_{i \in I} \lambda_i^* a_i]^\top x = \sum\nolimits_{i \in I} \lambda_i^* b_i = b^\top \lambda^* = \mathrm{Opt}(f),$$

   that is, $x \in X_*$ ■
3. Vice versa, if $I \subset \{1, ..., m\}$ is such that the set $X_I = \{x : a_i^\top x \leq b_i \, \forall i, a_i^\top x = b_i \, \forall i \in I\}$ is nonempty, then $X_I = X_* := \mathrm{Argmax}_{x \in X} f^\top x$ for properly selected f.
   Note: Nonempty sets of the form X_I, $I \subset \{1, ..., m\}$, are called *faces* of the polyhedral set X. This definition is not geometric – according to it, whether a given set Y is or is not a face in X, may depend not on X *per se*, but on its representation as the solution set of a finite system of linear inequalities. Items 2—3, taken together, state that in fact being a face of a polyhedral set is a geometric property – faces are exactly the sets $\operatorname*{Argmax}_{x \in X} f^\top x$ of all maximizers of linear forms bounded from above on X.

   Solution: Indeed, given $I \subset \{1, 2, ..., n\}$ such that X_I is nonempty, let us set $f = \sum_{i \in I} a_i$ [6], so that for $x \in X_I$ one has $f^\top x = \sum_{\in I} b_i$. On the other hand, for every $x \in X$ we have $f^\top x = \sum_{i \in} a_i^\top x \leq \sum_{i \in I} b_i$. We conclude that $\mathrm{Opt}(f) = \sum_{i \in I} b_i$ and $X_I \subset X_* := \operatorname*{Argmax}_{x \in X} f^\top x$,. The same reasoning

[6] recall that by our standard convention, $\sum_{i \in \varnothing} a_i = 0$.

as in the concluding part of the solution to the previous item (where λ_i^*, $i \in I$, should be set to 1) demonstrates the opposite inclusion $X_* \subset X_I$. Thus, $X_I = \mathrm{Argmax}_{x \in X} f^\top x$. ■

4. Extreme points of a face of X are extreme points of X.

   Solution: Assume that v is an extreme point of a face X_I of X; to prove that is an extreme point of X as well, we should show that whenever $v \pm h \in X$, it holds $h = 0$. To this end, note that if $v \pm h \in X$, then $a_i^\top [v \pm h] \le b_i$ for all i; when $i \in I$, the inequalities $a_i^\top [v \pm h] \le b_i$ imply that $a_i^\top [v \pm h] = b_i$ due to $a_i^\top v = b_i$. We see that in fact $v \pm h \in X_I$; since v is extreme point of X_I, we end up with the desired conclusion $h = 0$. ■

5. Extreme points of X, if any, are exactly the faces of X which are singletons.
   Note: As a corollary of 1—3, 5, we see that extreme points of polyhedral set X are exactly the maximizers of those linear forms which achieve their maximum on X at a unique point.

   Solution: In one direction: let v be an extreme point of X. By Theorem II.7.1, there exists n-element set $I \subset \{1, ..., m\}$ such that $a_i^\top v = b_i$ for $i \in I$ and the n vectors a_i, $i \in I$, are linearly independent. Since, in addition, $v \in X$, we conclude that $v \in X_I$, and the latter set is a singleton due to linear independence of a_i, $i \in I$. In the opposite direction: let $X_I = \{v\}$ for some I; then of course, v is an extreme point of X_I, which in view of item 4 implies that v is an extreme point of X.

Exercise II.21. [Follow-up to Exercise II.20]

1. Let $X \subset Y$ be nonempty closed convex sets in $\mathbf{R}^n$. Is it true that $\mathrm{Ext}(Y) \cap X \subset \mathrm{Ext}(X)$?

   Solution: The answer clearly is positive. Indeed, assuming that $w \in \mathrm{Ext}(Y) \cap X$ is not an extreme point of X, w is the midpoint of a nontrivial segment $\Delta \subset X$ and thus – a nontrivial segment $\Delta \subset Y$ (since $X \subset Y$), which is impossible.

2. Let X be a nonempty closed convex set contained in the polyhedral set $\{x : Ax \le b\}$. Assuming that the set $\overline{X} = X \cap \{x : Ax = b\}$ is nonempty, is it true that $\mathrm{Ext}(\overline{X}) = \mathrm{Ext}(X) \cap \overline{X}$?

   Solution: The answer is positive. Indeed, by item 1 it holds $\mathrm{Ext}(X) \cap \overline{X} \subset \mathrm{Ext}(X)$ due to $\overline{X} \subset X$. To prove the opposite inclusion, assume that an extreme point w of $\overline{X}$ is not extreme point of X, and let us lead this assumption to a contradiction. Since $w \in \overline{X} \subset X$, we have $w \in X$, and since w is not an extreme point of X, there exists a nontrivial segment $\Delta = [\underline{x}, \overline{x}] \subset X$ with w as the midpoint. By assumption, $A\overline{x} \le b$ and $A\underline{x} \le b$, which combines with $A\frac{1}{2}[\underline{x} + \overline{x}] = Aw = b$ to imply that $A\underline{x} = A\overline{x} = b$, that is, $\Delta \subset \overline{X}$. The bottom line is that w is the midpoint of a nontrivial segment in $\overline{X}$, which is he desired contradiction – w is an extreme point of $\overline{X}$!

3. By the result of Exercise II.13, the extreme points of the polytope $\Pi_{m,n} = \{[x_{ij}] \in \mathbf{R}^{m \times n} : x_{ij} \ge 0, \sum_i x_{ij} \le 1 \, \forall j, \sum_j x_{ij} \le 1 \, \forall i\}$ are exactly the Boolean matrices from this polytope. Now let $\widehat{\Pi}_{m,n}$ be the part of $\Pi_{m,n}$ cut off $\Pi_{m,n}$ by imposing on prescribed row and column sums of $m \times n$ matrix $x \in \Pi_{m,n}$ the requirement to be equal to 1, rather than to be ≤ 1. Assuming $\widehat{\Pi}_{m,n}$ nonempty, prove that the extreme points of this polytope are exactly the Boolean matrices contained in it.

   Solution: The fact that Boolean matrices contained in $\widehat{\Pi}_{m,n}$ are extreme points of this polytope is readily given by item 1 – we have already mentioned that these matrices are extreme points of the larger polytope $\Pi_{m,n}$. It remains to note that $\widehat{\Pi}_{m,n}$ is cut off $\Pi_{m,n}$ by converting into equalities several inequalities satisfied everywhere on $\Pi_{m,n}$, and thus by item 2 extreme points of $\widehat{\Pi}_{m,n}$ are extreme points of $\Pi_{m,n}$ and thus are Boolean matrices.

Exercise II.22. Let $X \subset \mathbf{R}^m$ be a nonempty polyhedral set, $x \mapsto Px + p : \mathbf{R}^n \to \mathbf{R}^m$ be an affine mapping, and Y be the image of X under this mapping. Mark by **T** the statements in the below list which are always (i.e., for all X, P, p compatible with the above assumptions) true:

1. Y is a nonempty polyhedral set.

   Solution: True, rule 4 in calculus of polyhedral representations, see section 3.3.

2. If X does not contain lines, so is Y.

   Solution: False – take $X = \left\{[x;y] \in \mathbf{R}^2 : \ y \geq |x|\right\}$ and consider the affine map $P[x;y] + p \equiv x$. Then, $Y = \mathbf{R}$ and is itself a straight line.

3. If X does contain lines, so does Y.

   Solution: False – take $X = \{[x;y] \in \mathbf{R}^2 : |x| \leq 1\}$ and $P[x;y] + p \equiv x$, resulting in $Y = [-1, 1]$.

4. If v is an extreme point of X, then $Pv + p$ is an extreme point of Y.

   Solution: False – take $X = \{[x;y] \in \mathbf{R}^2 : |x| + |y| \leq 1\}$ and $P[x;y] + p \equiv x$, resulting in $Y = [-1, 1]$. The image of the extreme point $[0; 1]$ of X under the affine mapping in question is not extreme for Y.

5. If z is an extreme point of Y, then $z = Pv + p$ for certain extreme point z of X.

   Solution: False– take $X = \{[x;y] \in \mathbf{R}^2 : |x| \leq 1\}$ and $P[x;y] + p \equiv x$, resulting in $Y = [-1, 1]$. Y has extreme points, and X does not.

6. If z is an extreme point of Y and X does not contain lines, then $z = Pv + p$ for certain extreme point z of X.

   Solution: True. By Exercise II.20, there is a linear form $f^\top y$ which attains its maximum over $y \in Y$ at z, and only at this point. It follows that the form $g^\top x$, $g = P^\top f$, attains its maximum over $x \in X$ exactly at the set $X^z = \{x \in X : Px + p = z\}$. By Exercise II.20, X^z is a face of X. Since X does not contain lines, so is X^z, implying that X^z has an extreme point, call it v. Since $v \in X^z$, we have $Pv + p = z$, and since v is an extreme point of face of X, it is extreme point of X by Exercise II.20.4. ■

Exercise II.23. Find extreme points of the following closed convex sets:

1. The set $\mathcal{S}_n = \{X \in \mathbf{S}^n : -I_n \preceq X \preceq I_n\}$

   Solution: $\mathrm{Ext}(\mathcal{S}_n)$ is the set of all matrices from $\mathbf{S}^n$ which are orthogonal, or, which is the same, symmetric $n \times n$ matrices with eigenvalues ± 1.
   In one direction: Let W be an orthogonal symmetric matrix; let us prove that this is an extreme point. Indeed, assuming that $W \pm D \in \mathcal{S}_n$ for some D, let us prove that $D = 0$. Otherwise there exists $x \in \mathbf{R}^n$ with $Dx \neq 0$; assuming w.l.o.g. that $\|x\|_2 = 1$, we have $\|Wx\|_2 = 1$, and $\|[W \pm D]x\|_2 \leq 1$ (since the spectral norm $\|V\|_{2,2}$ of a symmetric matrix $V \in \mathcal{S}_n$ is the maximum of magnitudes of eigenvalues of V and is therefore ≤ 1), On the other hand, assuming w.l.o.g. that $[Dx]^\top [Wx] \geq 0$, we have $\|[W + D]x\|_2^2 = \|Wx\|_2^2 + 2[Dx]^\top [Wx] + \|Dx\|_2^2 \geq \|Wx\|_2^2 + \|Dx\|_2^2 = 1 + \|Dx\|_2^2 > 1$, which is a desired contradiction.
   In the opposite direction: Let W be an extreme point of $\mathcal{S}_n$ and $W = U \,\mathrm{Diag}\{\lambda\} U^\top$ be the eigenvalue decomposition of W; we should verify that λ is a ± 1 vector. We clearly have $\|\lambda\|_\infty \leq 1$, and if $\|\lambda \pm d\|_\infty \leq 1$, then $W \pm U \,\mathrm{Diag}\{d\} U^\top \in \mathcal{S}_n$, implying that $d = 0$. Thus, λ is an extreme point of the unit box $\{x \in \mathbf{R}^n : \|x\|_\infty \leq 1\}$, and these points are the ± 1 vectors. ■

2. The set $\mathcal{S}_n^+ = \{X \in \mathbf{S}^n : 0 \preceq X \preceq I_n\}$

   Solution: The extreme points are exactly the orthogonal projectors – symmetric $n \times n$ matrices with eigenvalues 0 and 1. To see it, note that $\mathcal{S}_n^+$ is the image of $\mathcal{S}_n$ under the one-to-one affine mapping $X \mapsto \frac{1}{2}[X + I_n] : \mathbf{S}^n \to \mathbf{S}^n$.

3. The set $\mathcal{D}_{k,n} = \{X \in \mathbf{S}^n : I_n \succeq X \succeq 0, \mathrm{Tr}(X) = k\}$, where k is a positive integer $\leq n$.

   Solution: The extreme points are exactly the orthogonal rank k projectors, or, which is the same, the symmetric $n \times n$ matrices with k eigenvalues equal to 1 and the remaining eigenvalues equal to 0.
   In one direction: let $W \in \mathrm{Ext}(\mathcal{D}_{k,n})$, and let us prove that k eigenvalues of W are equal to 1, and the remaining – to 0. Indeed, let $W = U \,\mathrm{Diag}\{\lambda\} U^\top$ be the eigenvalue decomposition of W; since $W \in \mathcal{D}_{k,n}$, we should have $0 \leq \lambda_i \leq 1$ for $i \leq n$, and $\sum_i \lambda_i = k$. If now $d \in \mathbf{R}^n$ is such that $0 \leq \lambda_i \pm d_i \leq 1$ for $i \leq n$ and $\sum_i d_i = 0$, then $W \pm U \,\mathrm{Diag}\{d\} U^\top \in \mathcal{D}_{k,n}$, implying that $d = 0$ due to $W \in \mathrm{Ext}(\mathcal{D}_{k,n})$. Thus, λ should be an extreme point of the set $\{x \in \mathbf{R}^n : 0 \leq x_i \leq 1, i \leq n, \sum_i \lambda_i =$

$k\}$. As we know from Example II.7.1 in section 7.1.1, this implies that k entries in λ are equal to 1, and the remaining to 0.

In the opposite direction: Let W be symmetric $n \times n$ matrix with k eigenvalues equal to 1 and the remaining eigenvalues equal to 0, and let us prove that W is an extreme point of $\mathcal{D}_{k,n}$. Passing to representations of matrices in the eigenbasis of W, we lose nothing when assuming that $W = \sum_{i=1}^{k} e_i e_i^\top$, where $e_1, ..., e_n$ are the standard basic orths in $\mathbf{R}^n$. To see that $W \in \mathrm{Ext}(\mathcal{D}_{k,n})$, we should prove that if $\sum_{i=1}^{k} e_i e_i^\top \pm D \in \mathcal{D}_{k,n}$ and D is symmetric, then $D = 0$. Indeed, for D satisfying the premise of this claim, the diagonal entries of $\sum_{i=1}^{k} e_i e_i^\top \pm D$ should be between 0 and 1 and sum up to k, implying that $D_{ii} = 0$ for all i (the same Example II.7.1 we have already mentioned). In other words, the diagonals of positive semidefinite symmetric matrices $B^\pm = \sum_{i=1}^{k} e_i e_i^\top \pm D$ are $(\underbrace{1, ..., 1}_{k}, 0, ..., 0)$, implying that $D_{ij} = D_{ji} = B^\pm_{ij} = B^{ji} = 0$ whenever $\max[i, j] > k$ (since the 2×2 principal minors in $B^\pm$ should be nonnegative). Thus, all entries in D outside of the $k \times k$ angular submatrix $\overline{D}$ of D are zeros. Next, the matrices $I_k \pm \overline{D}$ are angular submatrices $E^\pm$ of symmetric matrices with eigenvalues between 0 and 1, implying by the Eigenvalue Interlacement Theorem that the eigenvalues of the symmetric $k \times k$ matrices $E^\pm$ are between 0 and 1, so that $E^\pm \in \mathcal{S}_k^+$. From item 2 we know that I_k is an extreme point of the latter set, implying that $\overline{D} = 0$. The bottom line is that $D = 0$, ■

4. The set $\mathcal{M}_n = \{X \in \mathbf{R}^{n \times n} : \|X\|_{2,2} \leq 1\}$ ($\|\cdot\|_{2,2}$ is the spectral norm)

Solution: The extreme points are exactly the orthogonal $n \times n$ matrices. To see that an orthogonal $n \times n$ matrix W is an extreme point of $\mathcal{M}_n$, you can use exactly the same reasoning as in the proof of the similar fact for $\mathcal{S}_n$, with $D \in \mathbf{R}^{n \times n}$ rather than $D \in \mathbf{S}_n$. To see that if W is an extreme point of $\mathcal{M}_n$, then W is orthogonal, or, which is the same, with all singular values equal to 1, look at the singular value decomposition $W = U\,\mathrm{Diag}\{\sigma\}V^\top$ of W, From $\|W\|_{2,2} \leq 1$ it follows that $\|\sigma\|_\infty \leq 1$, and if certain singular value σ_i is < 1, then the singular values of the matrices $W \pm tU[e_i e_i^\top]V^\top$ (e_i is i-th basic orth) for small positive t are ≤ 1, implying that $W \pm tU[e_i e_i^\top]V^\top \in \mathcal{M}_n$, which is impossible, ■

Exercise II.24. Prove the following fact (which can be considered as a matrix extension of Birkhoff Theorem):

For positive integers d, n, let $\Pi_{d,n}$ be the set of all $n \times n$ block matrices with $d \times d$ symmetric blocks X^{ij} satisfying

$$X^{ij} \succeq 0, \sum_j \mathrm{Tr}(X^{ij}) = 1 \forall i, \sum_i \mathrm{Tr}(X^{ij}) = 1 \forall j.$$

The extreme points of $\Pi_{d,n}$ are exactly the block matrices $[X^{ij}]_{i,j \leq n}$ as follows: for certain $n \times n$ permutation matrix P and unit vectors $e_{ij} \in \mathbf{R}^d$, one has

$$X^{ij} = P_{ij} e_{ij} e_{ij}^\top \, \forall i, j.$$

Solution: In one direction: Let $[W^{ij}]$ be an extreme point of $\Pi_{d,n}$ and $P_{ij} = \mathrm{Tr}(W_{ij})$, so that P is doubly stochastic. For every i, j, W^{ij} should be an extreme point of the set $\mathcal{D}_{ij} = \{X \in \mathbf{S}^d : X \succeq 0, \mathrm{Tr}(X) = P_{ij}\}$ (why?), whence, by item 3 of Exercise II.23, $W^{ij} = P_{ij} e_{ij} e_{ij}^\top$ for some unit e_{ij}. Besides this, P should be an extreme point of the polytope of doubly stochastic $n \times n$ matrices, since otherwise $P \pm D$ will be doubly stochastic for some nonzero D, implying that $W \pm \underbrace{[D_{ij} e_{ij} e_{ij}^\top]_{i,j \leq n}}_{\overline{D}} \in \Pi_{d,n}$ for nonzero block-matrix $\overline{D}$ with symmetric blocks, contradicting the fact that $W \in \mathrm{Ext}(\Pi_{d,n})$. Thus, by Birkhoff Theorem, P is a permutation matrix, and $W = [P_{ij} e_{ij} e_{ij}^\top]$ with unit $e_{ij} \in \mathbf{R}^d$. ■

In the opposite direction: Let $W = [P_{ij} e_{ij} e_{ij}^\top]$ with unit e_{ij} and permutation matrix P, and let $W \pm [D^{ij}] \in \Pi_{d,n}$ for some block-matrix with symmetric blocks D^{ij}; we should prove that $D^{ij} = 0$ for all i, j. If i, j are such that $P_{ij} = 1$, then the $d \times d$ matrices $e_{ij} e_{ij}^\top \pm D^{ij}$ are $\succeq 0$ with trace not exceeding 1 (as blocks in a matrix from $\Pi_{d,n}$), whence both matrices are $\succeq 0$, $\preceq I_d$, and with trace 1 (the latter – due to

$\mathrm{Tr}(e_{ij}e_{ij}^\top) = 1$). Thus, $e_{ij}e_{ij}^\top \pm D^{ij} \in \mathcal{D}_{1,d}$; applying item 3 of Exercise II.23, we conclude that $D^{ij} = 0$. And if $P_{ij} = 0$, then $W^{ij} \pm D^{ij}$ should be $\succeq 0$, again implying that $D^{ij} = 0$ due to $W^{ij} = P_{ij}e_{ij}e_{ij}^\top = 0$. ■

Exercise II.25. Let k, n be positive integers with $k \le n$, and let $s_k(\lambda)$ for $\lambda \in \mathbf{R}^n$ be the sum of k largest entries in λ. From the description of the extreme points of the polytope $X = \{x \in \mathbf{R}^n : 0 \le x_i \le 1, i \le n, \sum_{i=1}^n x_i \le k\}$, see Example II.7.2.2 in section 7.1.1, it follows that when $\lambda \in \mathbf{R}^n_+$, then

$$\max_{x \in X} \sum_{i=1}^n \lambda_i x_i = s_k(\lambda).$$

Prove the following matrix analogy of this fact:

> For k, n as above, let $\mathcal{X} = \{(X_1, ..., X_n) : X_i \in \mathbf{S}^d, 0 \preceq X_i \preceq I_d, i \le n, \sum_{i=1}^n X_i \preceq kI_d\}$. Then for $\lambda \in \mathbf{R}^n_+$ one has
>
> $$(X_1, ..., X_n) \in \mathcal{X} \implies \sum_{i=1}^n \lambda_i X_i \preceq s_k(\lambda) I_d,$$
>
> with the concluding $\preceq$ being $=$ for properly selected $(X_1, ..., X_n) \in \mathcal{X}$.

Solution: Assuming w.l.o.g. that $\lambda_1 \ge \lambda_2 \ge ... \ge \lambda_n$, for $X = (X_1, ..., X_n) \in \mathcal{X}$, setting $S_i = \sum_{j=1}^i X_j$, we have $S_i \preceq \min[i, k] I_d$. When $k = n$, we clearly have $\sum_{i=1}^n \lambda_i X_i \preceq \sum_{i=1}^n \lambda_i I_d = s_n(\lambda) I_d$, with $\preceq$ being $=$ when $X_i = I_d$, $i \le n$ Now let $k < n$, and let $\overline{X}_i = \begin{cases} I_d, & i \le k \\ 0, & i > k \end{cases}$, $\overline{S}_i = \sum_{j=1}^i \overline{X}_i$; note that $\overline{S}_i = \min[i, k] I_d \succeq S_i$, $i \le n$. We have

$$\begin{array}{l} \sum_{i=1}^n \lambda_i X_i = \sum_{i=1}^n \lambda_i [S_i - S_{i-1}] = \sum_{i=1}^{n-1} S_i \underbrace{[\lambda_i - \lambda_{i+1}]}_{\ge 0} + \underbrace{\lambda_n}_{\ge 0} S_n \\ \preceq \sum_{i=1}^{n-1} [\lambda_i - \lambda_{i+1}] \overline{S}_i + \lambda_n \overline{S}_n \\ = \sum_{i=1}^n \lambda_i [\overline{S}_i - \overline{S}_{i-1}] = s_k(\lambda) I_d, \end{array}$$

and the resulting inequality $\sum_i \lambda_i X_i \preceq s_k(\lambda) I_d$ is equality when $X_i = \overline{X}_i$, $i \le n$. ■

8.3 Cones and extreme rays

Exercise II.26. Let X be a nonempty closed and bounded set in $\mathbf{R}^n$. Which of the following statements are true?

1. $\mathrm{Conv}(X)$ is closed convex set.

   Solution: True – see Corollary I.2.5.5

2. $\mathrm{Cone}(X)$ is a closed cone.

   Solution: Wrong in general. When $X = \{x \in \mathbf{R}^2 : x_1^2 + (x_2 - 1)^2 \le 1\}$ (circle of unit radius in the upper half-plane touching the x_1-axis at the origin), $\mathrm{Cone}(X)$ is the open upper half-plane $\{x = [x_1; x_2] : x_2 > 0\}$ with origin added; this cone is not closed

3. When X is convex, $\mathrm{Cone}(X)$ is closed cone.

   Solution: Wrong in general, see example to item 2.

4. When $0 \notin X$, $\mathrm{Cone}(X)$ is a closed cone.

   Solution: Wrong in general. When $X = X^+ \cup X^-$ with $X^+ = \{[x_1; x_2; 1] \in \mathbf{R}^3 : x_1^2 + (x_2 - 1)^2 \le 1\}$, $X^- = \{[x_1; x_2; -1] \in \mathbf{R}^3 : x_1^2 + (x_2 - 1)^2 \le 1\}$, $\mathrm{Cone}(X)$ contains the circle $\{x_1^2 + (x_2 - 1)^2 \le 1\}$ in the plane $x_3 = 0$ and therefore contains the conic hull of this circle. As a result, $\mathrm{cl}\,\mathrm{Cone}(X)$ contains the tangent line $\{x_2 = 0, x_3 = 0\}$ to this circle, and this line clearly does not belong to $\mathrm{Cone}(X)$.

5. When $0 \notin X$ and X is convex, $\mathrm{Cone}(X)$ is closed cone.

Solution: True. The fact that $\mathrm{Cone}(X)$ is a cone holds true for every X; all we need is to prove that under the circumstances this cone is closed. Since X is nonempty closed convex set and $0 \notin X$, Separation Theorem applied to $\{0\}$ and X says that these two sets can be strongly separated, so that for properly selected e it holds $0 = e^\top 0 < \alpha := \inf_{x\in X} e^\top x$. Now let $y = \lim_{t\to\infty} y_t$ with $y_t \in \mathrm{Cone}(X)$; we want to prove that $y \in \mathrm{Cone}(X)$. We have $y_t = \sum_{i\le I_t} \lambda_{ti} x_{ti}$ with $\lambda_{ti} \ge 0$ and $x_{ti} \in X$. Setting $\lambda_t = \sum_i \lambda_{ti}$, we have

$$e^\top y = \lim_{t\to\infty} e^\top y_t = \lim_{t\to\infty} \sum_i \lambda_{it} \underbrace{e^\top x_{ti}}_{\ge\alpha>0}.$$

implying that the sequence of nonnegative reals λ_t is bounded. Therefore, passing to a subsequence, we may assume that $\lambda_t \to \bar\lambda$ as $t \to \infty$. Taking into account that $\|y_t\|_2 \le C\lambda_t$ with $C = \max_{x\in X} \|x\|_2 < \infty$, we see that when $\bar\lambda = 0$, one has $y = 0$, whence $y \in \mathrm{Cone}(X)$. And when $\bar\lambda > 0$, $y = \lim_{t\to\infty} y_t$ implies that $y = \bar\lambda \lim_{t\to\infty} \overline{x}_t$ with $\overline{x}_t = \lambda_t^{-1} \sum_i \lambda_{ti} x_{ti}$ (these points are well defined for large enough t's). Since X is convex, the points $\overline{x}_t$ belong to X, and since X is closed, the point $\overline{x} := \lim_{t\to\infty} \overline{x}_t$ belongs to X as well. Thus, y is a positive multiple of a point from X, so that $y \in \mathrm{Cone}(X)$. ■

6. When X is polyhedral, $\mathrm{Cone}(X)$ is a closed cone.

   Solution: True. By Krein-Milman Theorem, nonempty bounded polyhedral set is $\mathrm{Conv}\{v_1, ..., v_N\}$ for a finite nonempty set $\{v_1, ..., v_N\}$, whence clearly $\mathrm{Cone}(X) = \mathrm{Cone}(\{v_1, ..., v_N\}) = \{y = \sum_i \lambda_i v_i : \lambda_i \ge 0, i \le N\}$. Thus, $\mathrm{Cone}(X)$ admits polyhedral representation and is therefore polyhedral, and thus closed, set.

Exercise II.27. Let $X \subset \mathbf{R}^n$ be a nonempty polyhedral set given by polyhedral representation:

$$X = \{x : \exists u : Ax + Bu \le r\}$$

and let $K = \mathrm{Cone}(X)$ be the conic hull of X.

1. Is it true that K is a closed cone?

   Solution: Wrong in general. As every conic hull, K is a cone, but this cone not necessarily is closed. For example, when $X = \{[x_1; 1] : x_1 \in \mathbf{R}\} \subset \mathbf{R}^2$, $\mathrm{Cone}(X)$ is the union of the interior of the upper half-plane and of the origin, and this cone is not closed.

2. Prove that $\overline{K} := \mathrm{cl}\, K$ is a polyhedral cone and find polyhedral representation of $\overline{K}$.

   Solution: We claim that K admits polyhedral representation

   $$\overline{K} = \{x : \exists \lambda, u : \lambda \ge 0, A + Bu - \lambda r \le 0\}$$

   and is therefore a polyhedral cone. To justify the claim, denote the right hand side in the latter relation by K^+, so that K^+ is polyhedral (and thus closed) cone. To prove that $\overline{K} = K^+$ is the same as to check that, first, $K \subset K^+$ and, second, K is dense in K^+.
   To justify the first claim, note that $x \in K$ is of the form $\sum_i \lambda_i x_i$ with $\lambda_i \ge 0$ and $x_i \in X$; the latter means that for properly selected u_i it holds $Ax_i + Bu_i \le r$. Consequently, $A[\lambda_i x_i] + B[\lambda_i u_i] - \lambda_i r_i \le 0$. Summing up these vector inequalities, we get

   $$A\underbrace{\sum\nolimits_i \lambda_i x_i}_{x} + B\underbrace{\sum\nolimits_i \lambda_i u_i}_{=:u} + \underbrace{[\sum\nolimits_i \lambda_i]}_{=:\lambda} r \le 0;$$

   implying that $x \in K^+$ due to $\lambda = \sum_i \lambda_i \ge 0$.
   To justify the second claim, let us fix $\bar x \in X$ (X is nonempty!), so that $A\bar x + B\bar u - r \le 0$ for some $\bar u$. Now let $x \in K^+$, and let us prove that $x \in \mathrm{cl}\, K$. Indeed, $x \in K^+$ means that $Ax + Bu - \lambda r \le 0$ for some u and some $\lambda \ge 0$, implying that for $\epsilon > 0$ one has

   $$A\underbrace{[x + \epsilon\bar x]}_{=:x_\epsilon} + B[u + \epsilon\bar u] - \underbrace{[\lambda + \epsilon]}_{>0} r \le 0.$$

Dividing both sides by $[\lambda+\epsilon]$, we see that $x_\epsilon = [\lambda+\epsilon]x^\epsilon$ with $x^\epsilon = [\lambda+\epsilon]^{-1}x_\epsilon \in X$ Thus, $x_\epsilon \in K = \mathrm{Cone}(X)$; since $x_\epsilon \to x$ as $\epsilon \to +0$, we conclude that $x \in \mathrm{cl}\, K$, as claimed.

3. Assume that X is given by plain – no extra variables – polyhedral representation: $X = \{x : Ax \le b\}$. Build plain polyhedral representation of $\overline{K} := \mathrm{cl}\,\mathrm{Cone}(X)$.

   Solution: By the previous item,

   $$\overline{K} = \{x : \exists \lambda : \lambda \ge 0, Ax - b\lambda \le 0\},$$

   and to get plain polyhedral representation of K, it suffices to subject the above polyhedral representation to one step of Fourier-Motzkin elimination. To this end, let us set $\overline{A} = \left[\begin{array}{c|c} & -1 \\ A & -b \end{array}\right]$, so that $\overline{K} = \{x : \exists \lambda : \overline{A}[x;\lambda] \le 0\}$. Denoting the transposes of the rows of $\overline{A}$ by $[\alpha_i;\beta_i]$ with $\alpha_i \in \mathbf{R}^n$ and $\beta_i \in \mathbf{R}$ and denoting by I_0, I_+, I_- the sets of i's with $\beta_i = 0$, $\beta_i > 0$, $\beta_i < 0$, respectively, we have

   $$\begin{array}{rcl} \overline{K} & = & \left\{x : \exists \lambda : \overline{A}[x;\lambda] \le 0\right\} \\ & = & \left\{x : \begin{array}{l} \alpha_i^\top x \le 0 \, \forall i \in I_0 \\ [\beta_i^{-1}\alpha_i - \beta_j^{-1}\alpha_j]^\top x \le 0 \ \forall (i \in I_+, j \in I_-) \end{array}\right\} \end{array}$$

   and we end up with plain polyhedral representation of $\overline{K}$.

Exercise II.28. As we know, the extreme directions of the nonnegative orthant $\mathbf{R}^n_+ = \mathbf{R}_+ \times \mathbf{R}_+ \times ... \times \mathbf{R}_+$ are the vectors with single positive entry and remaining entries equal to 0. Prove the following generalization of this observation:

> Let $X_i \subset \mathbf{R}^{n_i}$, $1 \le i \le K$, be closed, nontrivial, and pointed cones. The extreme directions of the direct product $X = X_1 \times ... \times X_K$ of these cones, if any, are the block-vectors $d = [d_1; ...; d_K]$ with $d_i \in \mathbf{R}^{n_i}$ of the following structure: all but one blocks in d are zero, and the only nonzero block is an extreme direction of the corresponding factor X_i.

Solution: In one direction: if $d = [0; ...; 0; d_i; 0; ...; 0]$ with d_i being extreme direction of X_i and $d = d^1 + d^2$ with $d^1, d^2 \in X$, then d is nonzero and $d^1_j = d^2_j = 0$ for $j \ne i$; indeed, for j in question $d^1_j, d^2_j \in X_j$ and $d^1_j + d^2_j = d_j = 0$, and X_j is pointed. From $d_i = d^1_i + d^2_i$ with d_i being extreme direction of X_i both d^1_i and d^2_i are nonnegative multiples of d_i (indeed, d^1_i and d^2_i belong to X_i and sum up to the extreme direction d_i of X_i). Combining our observations, we conclude that d^1 and d^2 are nonnegative multiples of d, and we conclude that d is an extreme direction of X. In the opposite direction: let $d = [d_1; ...; d_K]$ be an extreme direction of X implying, in particular, that $d_i \in X_i$ for all i, and $d \ne 0$, so that d has a nonzero block, say, d_1. Since $d = \underbrace{[d_1; 0; ...; 0]}_{\in X} + \underbrace{[0; d_2; d_3; ...; d_K]}_{\in X}$ and d is extreme direction of X, the vector $[d_1; 0; ...; 0]$ must be nonnegative multiple of d; since $d_1 \ne 0$, $[d_1; 0; 0; ...; 0]$ in fact is a positive multiple of d, implying that $d_i = 0$, $i \ge 2$. It remains to verify that the only nonzero block, d_1, in d is an extreme direction of X_1. Assuming the opposite, we can represent d_1 as $d^1_1 + d^2_1$ with vectors d^χ_1, $\chi = 1, 2$, belonging to X_1 and not both being nonnegative multiples of d_1. But then $d = [d_1; 0; ...; 0] = d^1 + d^2$, with $d^\chi = [d^\chi_1; 0; ...; 0] \in X$, $\chi = 1, 2$, and at least one of d^1, d^2 not being a nonnegative multiple of d, which contradicts the fact that d is an extreme direction of X. ■

Exercise II.29. Describe all extreme rays of

1. positive semidefinite cone $\mathbf{S}^n_+$
2. Lorentz cone $\mathbf{L}^n$
3. Lorentz cone $\mathbf{L}^n$, $n \ge 2$, is the special case of the following construction: given a norm $\|\cdot\|$ on $\mathbf{R}^{n-1}$ $(n \ge 2)$, we associate with it the set

   $$\mathbf{K}^n_{\|\cdot\|} = \{[x;t] \in \mathbf{R}^n : t \ge \|x\|\},$$

   which is a pointed nontrivial cone with a nonempty interior (why?); note that $\mathbf{L}^n = \mathbf{K}^n_{\|\cdot\|_2}$. Describe the extreme directions of $\mathbf{K}^n_{\|\cdot\|}$.

Solution:

1. Extreme rays of $\mathbf{S}^n_+$ are nonnegative multiples $\mathbf{R}_+ \times ee^\top$, $e \in \mathbf{R}^n \setminus \{0\}$, of positive semidefinite rank 1 matrices.
   In one direction: When $e \in \mathbf{R}^n \setminus \{0\}$, in every representation $ee^\top = d^1 + d^2$ with $d^1 \succeq 0$, $d^2 \succeq 0$, for every x orthogonal to e we should have $0 = x^\top[ee^\top]x = \underbrace{x^\top d^1 x}_{\geq 0} + \underbrace{x^\top d^2 x}_{\geq 0}$, that is, $[\mathbf{R} \cdot e]^\perp$ is in the kernel of both d^1 and d^2, implying that the only eigenvector of d^i with nonzero eigenvalue, if any, is proportional to e. In other words, eigenvalue decomposition of d^i is $\lambda_i ee^\top$ with nonnegative λ_i (since λ_i is an eigenvalue of $d^i \succeq 0$). Thus, d^i, $i = 1, 2,$, are nonnegative multiples of $ee^\top$, so that $ee^\top$ is extreme direction of $\mathbf{S}^n_+$. In the opposite direction: let $E \in \mathbf{S}^n_+$ and $E = \sum_i \lambda_i e_i e_i^\top$ be eigenvalue decomposition of E. When the number of nonzero eigenvalues λ_i is > 1, say, $\lambda_1 > 0$ and $\lambda_2 > 0$, then $E = \lambda_1 e_1 e_1^\top + \sum_{i \geq 2} \lambda_i e_i e_i^\top$ is decomposition of E into sum of two positive semidefinite matrices which are not proportional to E, that is, positive semidefinite matrix of rank > 1 is not an extreme direction of $\mathbf{S}^n_+$. Since an extreme direction should be nonzero and positive semidefinite, it must be of the form $ee^\top$ with nonzero vector e. ■
2. The extreme directions of $\mathbf{L}^1 = \mathbf{R}_+$ are positive reals. When $n > 1$, the extreme directions of $\mathbf{L}^n$ are exactly positive multiples of vectors $[e; 1]$ with $e \in \mathbf{R}^{n-1}$, $\|e\|_2 = 1$, see solution to item 3.
3. Denoting by $B = \{x \in \mathbf{R}^{n-1} : \|x\| \leq 1\}$ the unit ball of norm $\|\cdot\|$, the extreme directions of $\mathbf{K}^n_{\|\cdot\|}$ are positive multiples of vectors $[x; 1]$ with $x \in \mathrm{Ext}(B)$. Indeed, the set
   $$Y := \{[x; 1] \in \mathbf{K}^n_{\|\cdot\|}\} = B \times \{1\}$$
   clearly is a base of the cone $\mathbf{K}^n_{\|\cdot\|}$, see Definition II.6.37. By Fact II.6.38.(iv), the extreme directions of $\mathbf{K}^n_{\|\cdot\|}$ are positive multiples of the vectors from $\mathrm{Ext}(Y)$, and
   $$\mathrm{Ext}(Y) = \mathrm{Ext}(B \times \{1\}) = \mathrm{Ext}(B) \times \{1\} = \{[x; 1] : x \in \mathrm{Ext}(B)\},$$
   where the second equality is due to Exercise II.10. ■

8.4 Recessive cone

Exercise II.30. Let M be a convex set, and let $\bar{x}$ and h be such that $R_{\bar{x}} := \{\bar{x} + th : t \geq 0\} \subset M$.

1. Is it always true that whenever $x \in M$, the set $R_x = \{x + th, t \geq 0\}$ is contained in M ?

Solution: The answer is no, example being $M = \{[x_1; x_2] \in \mathbf{R}^2 : x_1 \geq 0, x_2 > 0\} \cup \{[0; 0]\}$. This set clearly is convex and contains the ray $\{[x_1; 1] : x_1 \geq 0\}$ (that is, the ray $R_{\bar{x}}$ corresponding to $\bar{x} = [0; 1]$ and $h = [1; 0]$), but does not contain the parallel ray $R_{[0;0]} = \{[x_1; 0] : x_1 \geq 0\}$ emanating from $[0; 0] \in M$. ■

2. Let h be a recessive direction of $\overline{M} = \mathrm{cl}\, M$, and let $\bar{x}$ be a point from the relative interior of M. Is it always true that the set $R_{\bar{x}} = \{\bar{x} + th : t \geq 0\}$ is contained in M ?

Solution: The answer is yes. Indeed, by Lemma I.1.30 the ray $R = \{\bar{x} + th : t \geq 0\}$ is contained in $\overline{M}$, and since every point $x = \bar{x} + th$ on this ray is of the form $\frac{1}{2}\bar{x} + \frac{1}{2}x'$ with $x' \in R \subset \mathrm{cl}\, M$ (you can take $x' = \bar{x} + 2th$), $x \in M$ by Lemma I.1.30.

Exercise II.31. Let $M \subset \mathbf{R}^n$ be a cone, not necessary closed; recall that pointedness of a cone M means that the only vector x such that $x \in M$ and $-x \in M$ is the zero vector. Which of the following statements are always true:

1. M is pointed if an only if the only representation of 0 as the sum of $k \geq 1$ vectors $x_i \in M$ is the representation with $x_i = 0$, $i \leq k$.

Solution: This is true. Indeed, if M is not pointed, so that $\pm x \in M$ for some $x \neq 0$, then setting $k = 2$, $x_1 = x$, $x_2 = -x$, we get a representation of 0 as the sum of two nonzero vectors from M. On the other hand, when M is pointed and $0 = x_1 + \ldots + x_k$ with $x_i \in M$, then either $k = 1$ and $x_1 = 0$, or $k > 1$,

and then for every $i \le k$ we have $0 = x_i + \underbrace{\sum_{j \ne i} x_j}_{\in M}$ implying that $\pm x_i \in M$, whence, by pointedness, $x_i = 0$, $i \le k$. ■

2. M is pointed if and only if M does not contain straight lines (one-dimensional affine planes) passing through the origin.

Solution: True. In one direction: if M contains straight line passing through the origin, that is, the set $\{th : t \in \mathbf{R}\}$ with some $h \ne 0$ is contained in M, then $\pm h \in M$ and $h \ne 0$, contradicting pointedness of M. In the opposite direction: if M is not pointed, that is, $\pm h \in M$ for some $h \ne 0$, then M, being conic, contains the straight line $\{th : t \in \mathbf{R}\}$ passing through the origin. ■

3. M is pointed if and only if M does not contain straight lines.

Solution: Wrong – take $M = \{[x_1, x_2] \in \mathbf{R}^2 : x_2 > 0\} \cup \{[0;0]\}$. This cone is pointed (since all nonzero vectors from M have the second coordinate positive) and contains the line $\{[x_1; 1] : x_1 \in \mathbf{R}\}$. ■

4. Assuming M closed, M is pointed if and only if M does not contain straight lines.

Solution: True. By Lemma II.6.13, if a closed convex set contains a line, it contains all parallel lines intersecting the set, so that a closed cone M contains lines if and only if it contains lines passing through the origin, and it remains to use item 2. ■

5. M is pointed cone if and only if the closure of M is so.

Solution: Wrong, the counter-example being the pointed cone $M = \{[x_1; x_2] \in \mathbf{R}^2 : x_2 > 0\} \cup \{[0;0]\}$. ■

6. The closure of M is a pointed cone if and only if M does not contain straight lines.

Solution: True. If M contains a line, then this line is contained in the closed cone $\mathrm{cl}\, M$, so that $\mathrm{cl}\, M$ is not pointed by item 4. Vice versa, if $\mathrm{cl}\, M$ is not pointed, it contains a line $\{th : t \in \mathbf{R}\}$ $(h \ne 0)$ passing through the origin by item 2, and therefore by the result stated in Exercise II.30 M contains all lines of the form $\{x + th : t \in \mathbf{R}\}$ with $x \in \mathrm{rint}\, M$ $[\ne \varnothing]$. ■

Exercise II.32. Literal interpretation of the words "polyhedral cone" is: a polyhedral set $\{x : Ax \le b\}$ which is a cone. An immediate example is the solution set $\{x : Ax \le 0\}$ of *homogeneous* system of linear inequalities. Prove that this example is generic: whenever a polyhedral set $K = \{x : Ax \le b\}$ is a cone, one has $K = \{x : Ax \le 0\}$.

Solution: One way to prove the claim is to note that when the set $K = \{x : Ax \le b\}$ is a cone, this (clearly closed) set, as every closed cone, coincides with its recessive cone: $K = \mathrm{Rec}(K)$, and Fact II.6.20 states that for a nonempty polyhedral set $M = \{x : Ax \le b\}$ one has $\mathrm{Rec}(M) = \{x : Ax \le 0\}$.
A "bare hands" proof of the claim in question can be found in solution to Exercise I.4.

Exercise II.33. Prove the following modification of Proposition II.6.23:

> (!) *Let $X \subset \mathbf{R}^N$ be a nonempty closed convex set such that $X \subset V + \mathrm{Rec}(X)$ for some bounded and closed set V, let $x \mapsto \mathcal{A}(x) = Ax + b : \mathbf{R}^N \to \mathbf{R}^n$ be an affine mapping, and let $Y = \mathcal{A}(X) := \{y : \exists x \in X : y = \mathcal{A}(x)\}$ be the image of X under this mapping. Let also*
>
> $$K = \{h \in \mathbf{R}^n : \exists g \in \mathrm{Rec}(X) : h = Ag\}.$$
>
> *Then the recessive cone of the closure $\overline{Y}$ of Y is the closure $\overline{K}$ of K. In particular, when K is closed (as definitely is the case when $\mathrm{Rec}(X)$ is polyhedral), it holds $\mathrm{Rec}(\overline{Y}) = K$.*

Solution: If $y \in Y$ and $h \in K$, so that $y = \mathcal{A}(x)$ and $h = Ag$ for some $x \in X$ and $g \in \mathrm{Rec}(X)$, then $x + tg \in X$ for all $t \ge 0$, so that $y + th = \mathcal{A}(x + tg) \in Y \subset \overline{Y}$ whenever $t \ge 0$. Thus, h is a recessive direction of $\overline{Y}$, so that $K \subset \mathrm{Rec}(\overline{Y})$, and since the cone $\mathrm{Rec}(\overline{Y})$ is closed, $\overline{K}$ belongs to $\mathrm{Rec}(\overline{Y})$ along with K.
Vice versa, under the premise of Proposition, let $h \in \mathrm{Rec}(\overline{Y})$; we want to prove that $h \in \overline{K}$. Indeed,

selecting somehow $y \in Y$, we have $y + ih \in \overline{Y}$, $i = 1, 2, ...$ Next, from $X \subset V + \mathrm{Rec}(X)$ is follows that $Y \subset \widehat{Y} := \mathcal{A}(V) + A\mathrm{Rec}(X) = \mathcal{A}(V) + K$, and therefore $\overline{Y} \subset \mathrm{cl}(\mathcal{A}(V) + K) = \mathcal{A}(V) + \overline{K}$, where the concluding equality is due to the fact that $\mathcal{A}(V)$ is a compact set along with V [7]. Thus, $y + ih \in \overline{Y} \subset \mathcal{A}(V) + \overline{K}$, implying that for every i there exists $\delta_i \in \mathbf{R}^n$, $v_i \in V$ and $g_i \in \mathrm{Rec}(X)$ such that $y + ih = \mathcal{A}(v_i) + Ag_i + \delta_i$ and $\delta_i \to 0$ as $i \to \infty$. Setting $h_i = i^{-1}Ag_i$, we have $h = h_i + i^{-1}[\mathcal{A}(v_i) + \delta_i - y]$, and the second term in the right hand side of this equality tends to 0 as $i \to \infty$ due to the boundedness of V and of the sequence $\{\delta_i\}$. We conclude that $h = \lim_{i\to\infty} h_i$ with $h_i \in K$ for all i (due to $g_i \in \mathrm{Rec}(X)$), so that $h \in \overline{K}$. Recalling what h is, we conclude that $\mathrm{Rec}(\overline{Y}) \subset \overline{K}$. The opposite inclusion has already been verified, and we arrive at $\mathrm{Rec}(\overline{Y}) = \overline{K}$. ■

Exercise II.34. [follow-up to Exercise II.33]

1. Let $K_1 \subset \mathbf{R}^n, K_2 \subset \mathbf{R}^n$ be closed cones, and let $K = K_1 + K_2$.
   - Is it always true that K is a cone?

     Solution: K clearly is a cone.
   - Is it always true that K is closed?

     Solution: The answer is negative, as is shown by the following example: $K_1 = \{[x; y; z] \in \mathbf{R}^3 : y, z \geq 0, yz \geq x^2\}$ (this, up to one-to-one linear substitution of variables, is the 3D Lorentz cone), $K_2 = \{[x; y; z] : x = y = 0, z \leq 0\}$ (just a ray). In this case K contains all lines $\ell_a = \{[x; y; z] : y = a, z = 0\}$ with $a > 0$; indeed, given $a > 0$ and x, the vector $[x; a; x^2/a]$ belongs to K_1, and the vector $[0; 0; -x^2/a]$ belongs to K_2, so that the sum $[x; a; 0]$ of these vectors belongs to K, implying that $\ell_a \subset K$. On the other hand, the only vector of the form $[x; 0; 0]$ belonging to K clearly is the sum of some vector from K_1 with the y-coordinate equal to 0 and a vector from K_2; the only option for the first vector is to be of the form $[0; 0; z]$ with $z \geq 0$, and in this case, x must be zero. We see that K contains all lines ℓ_a with $a > 0$, but does not contain the line $\ell_0 \subset \mathrm{cl} \cup_{a>0} \ell_a$.
   - Let K_2 be polyhedral. Is it always true that K is closed?

     Solution: The answer is negative, as is shown by the example of the previous item, where K_2 is a ray.
   - Let both K_1 and K_2 be polyhedral. Is it always true that K is closed?

     Solution: The answer is positive: by evident reasons, K admits polyhedral representation and therefore is polyhedral.
2. Let X_i, $i = 1, ..., I$, be closed convex sets in $\mathbf{R}^n$ with nonempty intersection. Is it true that $\cap_i \mathrm{Rec}(X_i) = \mathrm{Rec}(\cap_i X_i)$?

   Solution: The answer is positive: selecting $x \in \cap_i X_i$, we have $h \in \mathrm{Rec}(\cap_i X_i)$ iff $x + th \in \cap_i X_i$ for all $t \geq 0$, or, which is the same, iff $h \in \mathrm{Rec}(X_i)$ for every i.
3. Let X_1, X_2 be nonempty closed convex sets in $\mathbf{R}^n$, let $K_1 = \mathrm{Rec}(X_1)$, $K_2 = \mathrm{Rec}(X_2)$, $\overline{X} = \mathrm{cl}(X_1 + X_2)$, $\overline{K} = \mathrm{cl}(K_1 + K_2)$.
   - Is it always true that $\overline{K} \subset \mathrm{Rec}(\overline{X})$?

     Solution: The answer is positive: selecting $x_i \in X_i$ and $h_i \in \mathrm{Rec}(X_i)$, $i = 1, 2$, we have $x_1 + x_2 + t(h_1 + h_2) \in X_1 + X_2$ for all $t \geq 0$, implying that $h_1 + h_2 \in \mathrm{Rec}(\overline{X})$. Thus, the cone $K_1 + K_2$ belongs to the cone $\mathrm{Rec}(\overline{X})$, and since the latter cone is closed, $\overline{K}$ belongs to this cone as well.
   - Is is always true that $\overline{K} = \mathrm{Rec}(\overline{X})$?

[7] We have used a nearly evident statement (prove it!): *if A, B are nonempty sets in $\mathbf{R}^m$ and A is bounded, then* $\mathrm{cl}(A + B) = \mathrm{cl}(A) + \mathrm{cl}(B)$.

Solution: The answer is negative: take $X_1 = \{[x;t] : x^2 \le t\}$, $X_2 = -X_1 = \{[x;t] : x^2 \le -t\}$. Then $K_1 = \{[0;t] : t \ge 0\}$, $K_2 = \{[0;t] : t \le 0\}$, so that $\overline{K} = \{[0;t], t \in \mathbf{R}\}$. At the same time, we clearly have $X_1 + X_2 = \mathbf{R}^2$, that is, $\mathrm{Rec}(\overline{X}) = \mathbf{R}^2$.

- Assume that $X_i \subset V_i + K_i$ for properly selected closed and bounded set V_i, $i = 1, 2$, Is it true that $\overline{K} = \mathrm{Rec}(\overline{X})$?

  Solution: The answer is positive. Indeed, let $Y = X_1 \times X_2$, $L = K_1 \times K_2$, $V = V_1 \times V_2$. Then clearly Y is a nonempty closed convex set, $L = \mathrm{Rec}(Y)$, and V is a bounded and closed set such that $Y \subset V + L$. Setting $\mathcal{A}(x_1, x_2) = x_1 + x_2$, we get a linear mapping acting from $\mathbf{R}^n \times \mathbf{R}^n$ to $\mathbf{R}^n$ such that $\overline{X} = \mathrm{cl}\,\mathcal{A}(Y)$ and $\overline{K} = \mathrm{cl}\,\mathcal{A}(L)$, so that $\overline{K} = \mathrm{Rec}(\overline{X})$ by the result of Exercise II.33.

Exercise II.35. Let $f(x) = x^\top Cx - c^\top x + \sigma$ be quadratic form with $C \succeq 0$. By Exercise I.15, the set $E = \{x : f(x) \le 0\}$ is convex (and of course closed). Assuming $E \neq \varnothing$, describe $\mathrm{Rec}(E)$.

Solution: Let $\bar{x} \in E$. Ray $\{\bar{x} + th : t \ge 0\}$ is contained in E if and only if

$$\forall t \ge 0 : t^2 \underbrace{h^\top Ch}_{\ge 0} + 2t\bar{x}^\top Ch - tc^\top h \le \underbrace{-[\bar{x}^\top C\bar{x} - c^\top \bar{x} + \sigma]}_{\ge 0},$$

which is possible if and only if $h^\top Ch = 0$ and $c^\top h \ge 0$. Recalling that for $C \succeq 0$ relation $h^\top Ch = 0$ is equivalent to $h \in \mathrm{Ker}C$, we get

$$\mathrm{Rec}(E) = \{h \in \mathrm{Ker}C : c^\top h \ge 0\}.$$

8.5 Around majorization

Exercise II.36. Let $x \in \mathbf{R}^m$, let $X[x]$ be the convex hull of all permutations of x, and let $X_+[x]$ be the set of all vectors x' dominated by a vector form $X[x]$:

$$X_+[x] = \{y \mid \exists z \in X[x] : y \le z\}.$$

1) Prove that $X_+[x]$ is a polyhedral set.

2) Prove the following characterization of $X_+[x]$: $X_+[x]$ is exactly the set of solutions of the system of inequalities $s_j(y) \le s_j(x)$, $j = 1, \ldots, m$, in variables y, where, as always $s_j(z)$ is the sum of the j largest entries in vector z.

Solution: 1) The set $X_+[x]$ is the sum of the polyhedral set $X[x]$ and the polyhedral cone $-\mathbf{R}^m_+$ and therefore admits immediate polyhedral representation: denoting by Σ the set of all $m!$ permutation matrices of size $m \times m$, we have

$$X_+[x] = \{y : \exists\{\lambda_\sigma, \sigma \in \Sigma\}, z \in \mathbf{R}^m : \lambda \ge 0, \sum_\sigma \lambda_\sigma = 1, z \ge 0, y = \sum_{\sigma\in\Sigma} \lambda_\sigma[\sigma x] - z\}$$

and is therefore polyhedral. ■

2) To justify the claim, let us fix x, and let $X^+[x]$ be the set of all solutions to the system of constraints $s_j(y) \le s_j(x)$, $1 \le j \le m$. We want to prove that $X^+[x] = X_+[x]$. First, if $y \in X_+[x]$, then $y \le \overline{y}$ for some $\overline{y} \in X[x]$. By Majorization Principle, we have $s_j(\overline{y}) \le s_j(x)$, $j \le m$ (in fact, the last of these inequalities is equality, but this does not matter now). And since $y \le \overline{y}$ and $s_j(z)$ is monotonically nondecreasing in z, we have $s_j(y) \le s_j(\overline{y}) \le s_j(x)$, $j \le m$, so that $y \in X^+[x]$. We conclude that $X_+[x] \subset X^+[x]$. To prove the inverse inclusion, let $y \in X^+[x]$, that is, $s_j(y) \le s_j(x)$, $j \le m$. Setting $\Delta = s_m(x) - s_m(y)$, we get $\Delta \ge 0$. Keeping all but the smallest entry in x intact and decreasing the smallest entry by Δ, we get a vector $\overline{x}$ such that $s_j(x) = s_j(\overline{x})$ for $j < m$ and $s_m(\overline{x}) = s_m(y)$. Thus, $s_j(y) \le s_j(\overline{x})$ for all j, the inequality being equality when $j = m$. By Majorization Principle, $y = D\overline{x}$ for some doubly stochastic matrix D, and since by construction $\overline{x} \le x$, we have $D\overline{x} \le Dx$, whence $y \le Dx$. Since Dx, by Birkhoff theorem, belongs to $X[x]$, we conclude that y is dominated by some point from $X[x]$, that is, $y \in X_+[x]$. Thus, $X^+[x] \subset X_+[x]$. ■

8.6 Around polars

Exercise II.37. Justify the last three claims in Example II.6.12.

Solution: 5: We have $\sup_{z\in DX} y^\top z = \sup_{x\in X} y^\top Dx = \sup_{x\in X}[D^\top y]^\top x$. Thus, $y \in \mathrm{Polar}\,(DX)$ if and only if $D^\top y \in \mathrm{Polar}\,(X)$. ■

6: We have $E = \{x = C^{-1/2}u : u^\top u \le 1\}$, whence $\max\limits_{x\in E} y^\top x = \max\limits_{u:u^\top u\le 1} [C^{-1/2}y]^\top u = \|C^{-1/2}y\|_2$, Thus, $\mathrm{Polar}\,(E) = \{y : \|C^{-1/2}y\|_2 \le 1\} = \{y : [C^{-1/2}y]^\top[C^{-1/2}y] \le 1\} = \{y : y^\top C^{-1} y \le 1\}$. ■

7: This is evident.

Exercise II.38. [more on polars]

1. Recall that for $U \subset \mathbf{R}^n$, $\mathrm{Vol}(U)$ stands for the ratio of the n-dimensional volume of U and the volume of the n-dimensional unit Euclidean ball. Check that for a centered at the origin ellipsoid $E = \{x : x^\top Cx \le 1\}$ ($C \succ 0$) we have $\mathrm{Vol}(E)\mathrm{Vol}(\mathrm{Polar}\,(E)) = 1$.
2. Let $C \succ 0$ and let ellipsoid $E = \{x : (x-c)^\top C(x-c) \le 1\}$ contain the origin. Compute $\mathrm{Polar}\,(E)$.
3. Let X_k, $k \le K$, be closed convex sets in $\mathbf{R}^n$ containing the origin. Prove that

$$\begin{array}{rcll} \mathrm{Polar}\,(\mathrm{Conv}(\cup_k X_k)) & = & \cap_k \mathrm{Polar}\,(X_k) & (a) \\ \mathrm{Polar}\,(\cap_k X_k) & = & \mathrm{cl}\,\mathrm{Conv}(\cup_k \mathrm{Polar}\,(X_k)) & (b) \end{array}$$

Solution: 1: By Example II.6.12.4, $\mathrm{Polar}\,(E) = \{x : x^\top C^{-1} x\}$, so that by the results of Exercise I.14 one has $\mathrm{Vol}(E) = \mathrm{Det}^{-1/2}(C)$, $\mathrm{Vol}(\mathrm{Polar}\,(E) = \mathrm{Det}^{-1/2}(C^{-1}) = \mathrm{Det}^{1/2}(C)$.

2: We have

$$\begin{array}{l} \mathrm{Polar}\,(E) = \{y : \max_{x=c+C^{-1/2}u:u^\top u\le 1} y^\top x \le 1\} = \{y : c^\top y + \max_{u:\|u\|_2\le 1} u^\top[C^{-1/2}y] \le 1\} \\ = \{y : \sqrt{yC^{-1}y} \le 1 - c^\top y\} \subseteq Q := \{y : y^\top C^{-1} y - [1 - c^\top y]^2 \le 0\} \end{array}$$

Let us prove that the $\subseteq$ above is in fact equality. To this end note that Q is a sublevel set of inhomogeneous quadratic form with the matrix

$$\Theta := C^{-1} - cc^\top = C^{-1/2}[I - dd^\top]C^{-1/2},$$

where $d, = C^{1/2}c$, so that $d^\top d = c^\top Cc \le 1$ due to $0 \in E$. We conclude that $\Theta \succeq 0$, implying that Q is convex (Exercise I.15). Now, to prove that the $\subseteq$ in question is in fact equality is the same as to prove that the linear function $1 - c^\top y$ is nonnegative everywhere on Q. Assuming that the latter is not the case and observing that $0 \in Q$, among the values taken on Q by the linear function in question there are both positive and negative, and since Q is convex, there should be $y \in Q$ with $c^\top y = 1$, and the latter clearly is forbidden by the definition of Q.

Thus, the polar of E is

$$Q = \{y : y^\top \Theta y + 2c^\top y \le 1\}.$$

Geometrically, this is

— either ellipsoid – this is the case when $\Theta \succ 0$, or, which is the same, $0 \in \mathrm{int}\, E$,

— or hyperparaboloid/elliptic cylinder – the set which in coordinates $t = Dx$ with properly selected nonsingular D is given by $\gamma t_1 \ge \alpha + \sum_{i\ge 2}[t_i - \beta_i]^2$ – this is what happens when $0 \in \mathrm{bd}\, E$.

3: A linear form does not exceed a real a on the convex hull of the union of K nonempty convex sets if and only if it does not exceed a on every one of these sets, resulting in (a). Setting $Y_k = \mathrm{Polar}\,(X_k)$, $k \le K$, so that $X_k = \mathrm{Polar}\,(Y_k)$ by Proposition II.6.42 (recall that X_k are closed, convex, and contain the origin) and applying (a) to the sets Y_k in the role of X_k, we get $\mathrm{Polar}\,(\mathrm{Conv}(\cup_k Y_k)) = \cap_k \mathrm{Polar}\,(Y_k) = \cap_k X_k$, whence also $\mathrm{Polar}\,(\mathrm{cl}\,\mathrm{Conv}(\cup_k Y_k)) = \cap_k X_k$. Since the set $\mathrm{cl}\,\mathrm{Conv}(\cup_k Y_k)$ is closed, convex, and contains the origin, it is the polar of its polar (Proposition II.6.42), that is, $\mathrm{cl}\,\mathrm{Conv}(\cup_k Y_k) = \mathrm{Polar}\,(\cap_k X_k)$. Recalling what Y_k are, we arrive at (b). ■

Exercise II.39. Let $X \subset \mathbf{R}^n$ be a cone given by polyhedral representation

$$X = \{x \in \mathbf{R}^n : \exists u : Ax + Bu \le r\}$$

Is the dual to X cone X_* polyhedral? If yes, build a polyhedral representation of X_*.

Solution: The fact that the cone dual to a polyhedral cone is polyhedral as well was explained in Remark II.6.27 and Exercise I.4. An independent reasoning (which as a byproduct yields polyhedral representation of X_* and as such can be considered as an addition to the calculus of polyhedral representations, see section 3.3), is as follows. We have

$$\begin{array}{rcl} y \in X_* & \Longleftrightarrow & y^\top x \geq 0 \, \forall x \in X \iff 0 \leq \min_{x,u} \left\{ y^\top x : Ax + Bu \leq r \right\} \\ & \Longleftrightarrow & 0 \leq \max_\lambda \left\{ -r^\top \lambda : \lambda \geq 0, A^\top \lambda + y = 0, B^\top \lambda = 0 \right\} \text{ [LP Duality]} \\ & \Longleftrightarrow & \exists \lambda : r^\top \lambda \leq 0, \lambda \geq 0, A^\top \lambda + y = 0, B^\top \lambda = 0, \end{array}$$

and we end up with polyhedral representation of X_*.

Exercise II.40.

1. Let $X \subset \mathbf{R}^n$ be a nonempty polyhedral set given by polyhedral representation

$$X = \{x \in \mathbf{R}^n : \exists u : Ax + Bu \leq r\}$$

Is the polar $\mathrm{Polar}\,(X)$ of X polyhedral? If yes, point out a polyhedral representation of $\mathrm{Polar}\,(X)$. For non-polyhedral extension, see Exercise IV.36.

Solution: We have

$$\begin{array}{rcl} \mathrm{Polar}\,(X) & = & \{y : \mathrm{Opt}(P) := \max_{x,u}\{y^\top x : Ax + Bu \leq r\} \leq 1\} \\ & = & \{y : \mathrm{Opt}(D) := \min_\lambda \left\{ r^\top \lambda : \lambda \geq 0, A^\top \lambda = y, B^\top \lambda = 0 \right\} \leq 1\} \\ & & \text{[by LP Duality; note that } (P) \text{ is feasible due to } X \neq \varnothing] \\ & = & \{y : \exists \lambda : r^\top \lambda \leq 1, \lambda \geq 0, y = A^\top \lambda, B^\top \lambda = 0\}, \\ & & \text{[since by the above, the dual problem is solvable when } y \in \mathrm{Polar}\,(X)] \end{array}$$

and we end up with polyhedral representation of $\mathrm{Polar}\,(X)$, implying polyhedrality of the polar.

2. Compute the polars of
   1. probabilistic simplex $\Delta = \{x \in \mathbf{R}^n : x \geq 0, \sum_i x_i = 1\}$

      Solution: $\mathrm{Polar}\,(\Delta) = \{y \in \mathbf{R}^n : y \leq [1; ...; 1]\}$

   2. convex hull of nonempty finite set of points $a_1, ..., a_N$ from $\mathbf{R}^n$

      Solution: $\mathrm{Polar}\,(\mathrm{Conv}\{a_1, ..., a_N\}) = \{y : a_i^\top y \leq 1, i \leq N\}$

   3. the set $\{x \in \mathbf{R}^n : x \leq b\}$

      Solution: $\mathrm{Polar}\,(\{x : x \leq b\}) = \{y : y \geq 0, y^\top b \leq 1\}$

8.7 Miscellaneous exercises

Exercise II.41. Let $X = \{x \in \mathbf{R}^n : Ax \leq b\}$ be a nonempty polyhedral set.

1. Prove that X is bounded if and only if every one of the vectors $\pm e_i$, (e_i, $1 \leq i \leq n$, are the standard basic orths) can be represented as conic combination of columns of $A^\top$.

   Solution: A nonempty polyhedral set $\{x : Ax \leq b\} \subset \mathbf{R}^n$ is bounded if and only if the optimal values in the $2n$ optimization problems $\max_x\{\pm e_i^\top x : Ax \leq b\}$ are finite, and this, by LP Duality Theorem, boils down to feasibility of their duals, the latter being exactly the possibility to represent $\pm e_i$ as conic combination of the columns of $A^\top$.

2. Certify the statements:
   - The polyhedral set $X = \{x \in \mathbf{R}^3 : x \geq [1/3; 1/3; 1/3], \sum_{i=1}^3 x_i \leq 1\}$ is bounded.

Solution: The set is

$$\{x : Ax \le b\},\ A = \begin{bmatrix} -1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & -1 \\ 1 & 1 & 1 \end{bmatrix},\ b = {}_q\begin{bmatrix} -1/3 \\ -1/3 \\ -1/3 \\ 1 \end{bmatrix}$$

It suffices to verify that every one of the vectors $\pm e_i$, $i = 1, ..., 3$, is a conic combination of the columns of $A^\top$. The vectors $-e_i$ are among the columns of $A^\top$; to get e_1, sum up all columns of $A^\top$ but the first one, and similarly for e_2 and e_3.

- The polyhedral set $X = \{x \in \mathbf{R}^3 : x_1 \ge 1/3, x_2 \ge 1/3, \sum_{i=1}^3 x_i \le 1\}$ is unbounded.

  Solution: By Lemma II.6.13 a polyhedral set is unbounded if and only if it is nonempty and its recessive cone is nontrivial. For the set in question, certificate of nonemptiness is, e.g., $x = [1/3; 1/3; 1/3]$, and a nonzero vector in $\mathrm{Rec}(X) = \{x \in \mathbf{R}^3 : x_1 \ge 0, x_2 \ge 0, x_1 + x_2 + x_3 \le 0\}$ is, e.g., $x = [0; 0; -1]$.

Exercise II.42. Prove the easy part of Theorem II.7.7, specifically, that every $n \times n$ permutation matrix is an extreme point of the polytope Π_n of $n \times n$ doubly stochastic matrices.

Solution: Let $\overline{\Pi}_n$ be the set of all $n \times n$ matrices with entries from $[0, 1]$. As we know, the extreme points of $\overline{\Pi}_n$ are exactly the $n \times n$ matrices with zero and one entries. In view of this, the claim to be proved is readily given by the following statement (evident due to geometric characterization of extreme points): *If $X \subset Y$ are convex sets, then every extreme point of Y which happens to belong to X is an extreme point of X.*

Exercise II.43. [robust LP] Consider an *uncertain* Linear Programming problem – a family

$$\left\{\min_{x \in \mathbf{R}^n} \{c^\top x : [A + \sum\nolimits_{\nu=1}^N \zeta_\nu \Delta_\nu]x \le b + \sum\nolimits_{\nu=1}^N \zeta_\nu \delta_\nu\} : \zeta \in \mathcal{Z}\right\} \tag{8.3}$$

of LP instances of common sizes (n variables, m constraints). The associated story is as follows: we want to solve an LP program with the data not known exactly when the problem is being solved; what we know at this time, is that the "true problem" belongs to the parametric family given, according to (8.3), by the "nominal data" c, A, b, the "basic perturbations Δ_ν, δ_ν" and the *perturbation set* $\mathcal{Z}$ through which run the data perturbations ζ specifying particular instances in the family. In this situation (quite typical for real life applications of LP, where partial data uncertainty is the rule rather than the exception), one way to "immunize" decisions against data uncertainty is to look for *robust solutions* – those remaining feasible for all perturbations of the data from the perturbation set – by solving the *Robust Counterpart* (RC) of our uncertain problem – the optimization problem

$$\min_x \left\{c^\top x : [A + \sum\nolimits_{\nu=1}^N \zeta_\nu \Delta_\nu]x \le b + \sum\nolimits_{\nu=1}^N \zeta_\nu \delta_\nu \ \forall (\zeta \in \mathcal{Z})\right\} \tag{RC}$$

(RC) is *not* an LP program – it has finitely many decision variables and infinite (when $\mathcal{Z}$ is "massive") system of linear constraints on these variables. Optimization problems of this type are called *semi-infinite* and are, in general, difficult to solve. However, the RC of an uncertain LP is easy, provided that $\mathcal{Z}$ is a "computation-friendly" set, for example, nonempty set given by polyhedral representation:

$$\mathcal{Z} = \{\zeta : \exists u : P\zeta + Qu \le r\} \tag{8.4}$$

Now goes the exercise *per se*:
Use LP duality to reformulate (RC), (8.4) as an explicit LP program.

Solution: The constraints of (RC) are of the form

$$\max_{\zeta \in \mathcal{Z}} \sum\nolimits_\nu \zeta_\nu \left[\Delta_\nu x - \delta_\nu\right]_j \le [b - Ax]_j,\ 1 \le j \le m,$$

or, which is the same,

$$\max_{\zeta, u} \left\{\sum\nolimits_\nu \left[\Delta_\nu x - \delta_\nu\right]_j \zeta_\nu : P\zeta + Qu \le r\right\} \le [b - Ax]_j,\ 1 \le j \le m.$$

Applying LP Duality, the constraints can be rewritten as

$$\min_{\lambda^j}\left\{r^\top\lambda^j:\ \begin{array}{l}[P^\top\lambda^j]_\nu=\Delta_\nu x-\delta_\nu,\ 1\le\nu\le N\\ Q^\top\lambda^j=0,\ \lambda^j\ge 0\end{array}\right\}\le[b-Ax]_j,\ 1\le j\le m. \qquad (*)$$

We see that x is robust feasible (i.e., feasible for (RC)) if and only if x can be augmented by properly selected $\lambda^1,...,\lambda^m$ to satisfy system $(*)$ of linear constraints on x and λ^j's. As a result, (RC) is equivalent to the explicit LP program

$$\min_{x,\lambda^1,...,\lambda^m}\left\{\left.\begin{array}{ll}&\Delta_\nu x-[P^\top\lambda^j]_\nu=\delta_\nu,1\le\nu\le N\\ c^\top x:&Q^\top\lambda^j=0,\ \lambda^j\ge 0\\ &r^\top\lambda^j+[Ax]_j\le b_j\end{array}\right\},\ j=1,...,m\right\}$$

Exercise II.44. Consider scalar linear constraint

$$a^\top x\le b \qquad (1)$$

with uncertain data $a\in\mathbf{R}^n$ (b is certain) varying in the set

$$\mathcal{U}=\{a:|a_i-a_i^*|/\delta_i\le 1,1\le i\le n,\sum\nolimits_{i=1}^n|a_i-a_i^*|/\delta_i\le k\} \qquad (2)$$

where a_i^* are given "nominal data," $\delta_i>0$ are given quantities, and $k\le n$ is an integer (in literature, this is called "budgeted uncertainty"). Rewrite the Robust Counterpart

$$a^\top x\le b\ \forall a\in\mathcal{U} \qquad \text{(RC)}$$

in a tractable LO form (that is, write down an explicit system (S) of linear inequalities in variables x and additional variables such that x satisfies (RC) if and only if x can be extended to a feasible solution of (S)).

Solution: Let D be diagonal $n\times n$ matrix with diagonal entries δ_i, and let $a_i-a_i^*=\delta_i\epsilon_i$, so that

$$\begin{array}{rcl}\mathcal{U}&=&\{a=a^*+D\epsilon:-1\le\epsilon_i\le 1\,\forall i,\sum_i|\epsilon_i|\le k\}\\ &=&\{a=a^*+D\epsilon:-u\le\epsilon\le u,u_i\le 1\,\forall i,\sum_i u_i\le k\}.\end{array}$$

x is robust feasible iff

$$\begin{array}{rcl}b&\ge&\max\limits_a\left\{x^\top a:a\in\mathcal{U}\right\}=\max\limits_{\epsilon,u}\left\{x^\top[a^*+D\epsilon]:-u\le\epsilon\le u,\ u\le[1;...;1],\sum_i u_i\le k\right\}\\ &=&x^\top a^*+\max\limits_{\epsilon,u}\left\{[Dx]^\top\epsilon:-u\le\epsilon\le u,u\le[1;...;1],\sum_i u_i\le k\right\}\\ &=&x^\top a^*+\min\limits_{\lambda_{\ell,u},\lambda_{g,u},\lambda_{\ell,1},\lambda_{\ell,k}}\left\{[1;...;1]^\top\lambda_{\ell,1}+k\lambda_{\ell,k}:\right.\\ &&\qquad\qquad\left.\begin{array}{r}\lambda_{\ell,u}\ge 0,\lambda_{\ell,1}\ge 0,\lambda_{\ell,k}\ge 0,\lambda_{g,u}\le 0\\ \lambda_{\ell,u}+\lambda_{g,u}=Dx\\ -\lambda_{\ell,u}+\lambda_{g,u}+\lambda_{\ell,1}+\lambda_{\ell,k}[1;...;1]=0\end{array}\right\}\\ &&\text{[LO duality]}\\ &=&x^\top a^*+\min\limits_{\lambda_{\ell,1},\lambda_{\ell,k}}\left\{[1;...;1]^\top\lambda_{\ell,1}+k\lambda_{\ell,k}:\ \begin{array}{l}\lambda_{\ell,1}\ge 0,\lambda_{\ell,k}\ge 0\\ -\lambda_{\ell,1}-\lambda_{\ell,k}[1;...;1]\le Dx\le\lambda_{\ell,1}+\lambda_{\ell,k}[1;...;1]\end{array}\right\}\\ &&\text{[eliminating }\lambda_{\ell,u},\lambda_{g,u}]\end{array}$$

Thus, (RC) can be represented as the system of linear constraints

$$\begin{array}{l}\lambda_{\ell,1}\ge 0,\lambda_{\ell,k}\ge 0,[a^*]^\top x+[1;...;1]^\top\lambda_{\ell,1}+k\lambda_{\ell,k}\le b,\\ -\lambda_{\ell,1}-\lambda_{\ell,k}[1;...;1]\le Dx\le\lambda_{\ell,1}+\lambda_{\ell,k}[1;...;1],\end{array}$$

in variables $x,\lambda_{\ell,1},\lambda_{\ell,k}$.

Exercise II.45. [computational study, follow-up to Exercise II.43]
Preliminaries. Consider oscillator transmitting harmonic wave with unit wavelength and placed at

some point P in 3D. Physics says that the electric field generated by the oscillator, when measured at a remote point A, is

$$e_A(t) \approx r^{-1} \underbrace{\alpha \cos\left(\omega t - 2\pi r + \theta + 2\pi d \cos(\phi)\right)}_{E_A(t)} \qquad (*)$$

where

- t is time, ω is the frequency,
- r is the distance from A to the origin O, d is the distance from P to the origin, $\phi \in [0, \pi]$ is the angle between the directions $\overrightarrow{OP}$ and $\overrightarrow{OA}$,
- α and θ are responsible for how the oscillator is actuated.

The difference between the left and the right hand sides in $(*)$ of order of r^{-2} and in all our subsequent considerations can be completely ignored.

It is convenient to assemble α and θ into the *actuation weight* – the complex number $w = \alpha \mathrm{e}^{\imath\theta}$ ($\imath$ is the imaginary unit); with this convention, we have

$$E_A(t) = \Re\left[w D_P(\phi) \mathrm{e}^{\imath\omega t - 2\pi r}\right], \; D_P(\phi) = \mathrm{e}^{2\pi\imath d\cos(\phi)}.$$

where $\Re[\cdot]$ stands for the real part of a complex number. The complex-valued function $D_P(\phi) : [0, \pi] \to \mathbf{C}$, called *the diagram* of the oscillator, is responsible for the directional density of the energy emitted by the oscillator: when evaluated at certain 3D direction $\vec{e}$, this density is proportional to $|D_p(\phi)|^2$, where ϕ is the angle between the direction $\vec{e}$ and the direction $\overrightarrow{OP}$. Physics says that when our transmitting antenna is composed of K harmonic oscillators located at points $P_1, ..., P_K$ and actuated with weights $w_1, ..., w_K$, the directional density of energy emitted by the resulting *antenna array*, as evaluated at a direction $\vec{e}$, is proportional to $|\sum_k w_k D_k(\phi_k(\vec{e}))|^2$, where $\phi_k(\vec{e})$ is the angle between the directions $\vec{e}$ and $\overrightarrow{OP_k}$.

Consider the design problem as follows. We are given linear array of K oscillators placed at the points $P_k = (k-1)\delta\mathbf{e}$, $k \le K$, where $\mathbf{e}$ is the first basic orth (that is, the unit vector "looking" along the positive direction of the x-axis), and $\delta > 0$ is a given distance between consecutive oscillators. Our goal is to specify actuation weights w_k, $k \le K$, in order to send as much of total energy as possible along the directions which make at most a given angle γ with $\mathbf{e}$. To this end, we intend to act as follows:

> We want to select actuation weights w_k, $k \le K$, in such a way that the magnitude $|D^w(\phi)|$ of the complex-valued function
>
> $$D^w(\phi) = \sum_{k=1}^{K} w_k \mathrm{e}^{2\pi\imath(k-1)\delta\cos(\phi))}$$
>
> of $\phi \in [0, \pi]$ is "concentrated" on the segment $0 \le \phi \le \gamma$. Let us normalize the weights by the requirement
>
> $$D^w(0) = 1$$
>
> and minimize under this restriction the "sidelobe level"
>
> $$\max_{\gamma \le \phi \le \pi} |D^w(\phi)|$$
>
> over w.

To get a computation-friendly version of this problem, we replace the full range $[0, \pi]$ of values of ϕ with M-point equidistant grid

$$\Gamma = \{\phi_\ell = \frac{\ell\pi}{M-1} : 0 \le \ell \le M-1\},$$

thus converting our design problem into the optimization problem

$$\mathrm{Opt} = \min_{t,w} \left\{ t : \begin{array}{l} |\sum_{k=1}^{K} w_k \mathrm{e}^{2\pi\imath(k-1)\delta\cos(\phi_\ell)}| \le t \, \forall(\ell : \phi_\ell > \gamma) \\ \sum_{k=1}^{K} w_k \mathrm{e}^{2\pi\imath(k-1)\delta} = 1 \end{array}, w_k \in \mathbf{C}, k \le K \right\} \qquad (P)$$

which is a convex problem in $2k$ real variables – real and imaginary parts of $w_1, ..., w_K$.
Your tasks are as follows:

1. Process problem (P) numerically and find the optimal design $w^{\rm n} = \{w_k^{\rm n}, k \le K\}$ along with the optimal value $\mathrm{Opt}^{\rm n}$. Here and in what follows, recommended setup is
   - number of oscillators $K = 24$, distance between consecutive oscillators $\delta = 0.125$
   - $\gamma = \pi/12$
   - cardinality M of the equidistant grid Γ is 512

   Draw the plot of the modulus of the resulting diagram

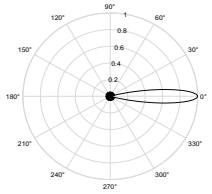
   $$D^{\rm n}(\phi) = \sum_{k=1}^{K} w_k^{\rm n} \mathrm{e}^{2\pi\imath(k-1)\delta\cos(\phi)}$$

   and compute the corresponding "energy concentration" $\mathcal{C}^{\rm n}$, with concentration of a diagram $D(\cdot)$ defined as

   $$\mathcal{C} = \frac{\sum_{\ell:\phi_\ell\le\gamma} \sin(\phi_\ell)|D(\phi_\ell)|^2}{\sum_{\ell=1}^{M} \sin(\phi_\ell)|D(\phi_\ell)|^2}$$

   – up to discretization of ϕ, this is the ratio of the energy emitted in the "cone of interest" (i.e., along the directions making angle at most γ with $\mathbf{e}$) to the total emitted energy. Factors $\sin(\phi_\ell)$ reflect the fact that when computing the energy emitted in a spatial cone, we should integrate $|D(\cdot)|^2$ over the part of the unit sphere in $3D$ cut off the sphere by the cone.

   Solution: Our computation yielded diagram with modulus as shown on Figure S2II.1.

   $\mathrm{Opt}^{\rm n} = 0.053$, $\mathcal{C}^{\rm n} = 74.8\%$

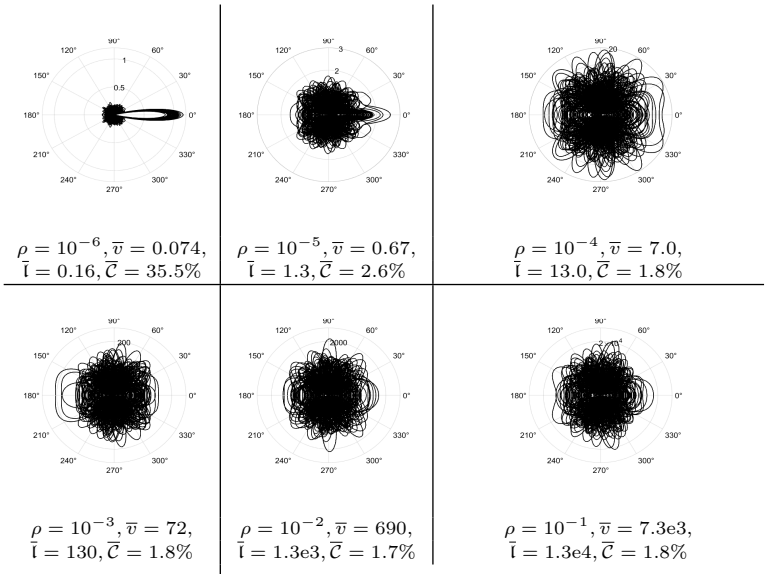
   Figure S2II.1. Optimal diagram, dream – no actuation errors.

2. Now note that "in reality" the optimal weights $w_k^{\rm n}$, $k \le K$ are used to actuate physical devices and as such cannot be implemented with the same 16-digit accuracy with which they are computed; they definitely will be subject to small implementation errors. We can model these errors by assuming that the "real life" diagram is

   $$D(\phi) = \sum\nolimits_{k=1}^{K} w_k^n (1 + \rho\xi_k) \mathrm{e}^{2\pi\imath(k-1)\delta\cos(\phi)}$$

   where $\rho \ge 0$ is some (perhaps small) perturbation level and $\xi_k \in \mathbf{C}$ are "primitive" perturbations responsible for the implementation errors and running through the unit disk $\{\xi : |\xi| \le 1\}$. It is not a great sin to assume that ξ_k are independent across k random variables uniformly distributed on the unit circumference in $\mathbf{C}$. Now the diagram becomes random and can violate the constraints of (P) , unless $\rho = 0$; in the latter case, the diagram is the "nominal" one given by the optimal weights $w^{\rm n}$, so that it satisfies the constraints of (P) with t set to $\mathrm{Opt}^{\rm n}$.
   Now, what happens when $\rho > 0$? In this case, the diagram $D(\cdot)$ and its deviation v from the prescribed value 1 at the origin, its sidelobe level $\mathfrak{l} = \max_{\ell:\phi_\ell>\gamma} |D(\phi_\ell)|$, and energy concentration become random. A crucial "real life" question is how large are "typical values" of these quantities. To get impression of what happens, you are asked to carry out the numerical experiment as follows:
   - select perturbation level $\rho \in \{10^{-\ell}, 1 \le \ell \le 6\}$
   - for selected ρ, simulate and plot 100 realizations of the modulus of the actual diagram, and find empirical averages $\overline{v}$ of v, $\overline{\mathfrak{l}}$ of $\mathfrak{l}$, and $\overline{\mathcal{C}}$ of $\mathcal{C}$.

   Solution: Our experimental results are shown on Figure S2II.2. To put the above concentration numerics into proper perspective, note that with our setup, the surface of the "spherical hat" cut off the unit sphere by our cone of interest is 1.7% of the total surface of the sphere, so that energy concentration 1.7% we can get without any trouble by placing just one oscillator at the origin.



$\rho = 10^{-6}, \overline{v} = 0.074,$ $\overline{\mathsf{i}} = 0.16, \overline{\mathcal{C}} = 35.5\%$	$\rho = 10^{-5}, \overline{v} = 0.67,$ $\overline{\mathsf{i}} = 1.3, \overline{\mathcal{C}} = 2.6\%$	$\rho = 10^{-4}, \overline{v} = 7.0,$ $\overline{\mathsf{i}} = 13.0, \overline{\mathcal{C}} = 1.8\%$
$\rho = 10^{-3}, \overline{v} = 72,$ $\overline{\mathsf{i}} = 130, \overline{\mathcal{C}} = 1.8\%$	$\rho = 10^{-2}, \overline{v} = 690,$ $\overline{\mathsf{i}} = 1.3\text{e}3, \overline{\mathcal{C}} = 1.7\%$	$\rho = 10^{-1}, \overline{v} = 7.3\text{e}3,$ $\overline{\mathsf{i}} = 1.3\text{e}4, \overline{\mathcal{C}} = 1.8\%$

Figure S2II.2. Nominal diagram – reality,
Magnitudes of 100 actual diagrams stemming from the optimal solution to (P)

Taking into account that in "real life" implementation errors in antenna weights hardly could be less than 0.1% (corresponding to $\rho = 10^{-3}$), we would qualify the *nominal* design yielded by the optimal solution to the nominal problem (P), same as the nominal optimal value, as wishful thinking completely meaningless for actual antenna design.

3. Apply Robust Optimization methodology from Exercise II.43 to build "immunized against implementation errors" solution to (P), compute these solutions for perturbation levels $10^{-\ell}$, $1 \leq \ell \leq 6$, and subject the resulting designs to numerical study similar to the one outlined in the previous item.
Note: (P) is *not* a Linear Programming program, so that you cannot formally apply the results stated in Exercise II.43; what you can apply, is the Robust Optimization "philosophy."

Solution:

- With our model of implementation errors, the effect of these errors on the value of the actual diagram $D(\cdot)$ as evaluated at a point $\phi \in \Gamma$ is in adding to the value

$$D_w(\phi) = \sum_{k=1}^{k} w_k \mathrm{e}^{2\pi\imath(k-1)\delta\cos(\phi)}$$

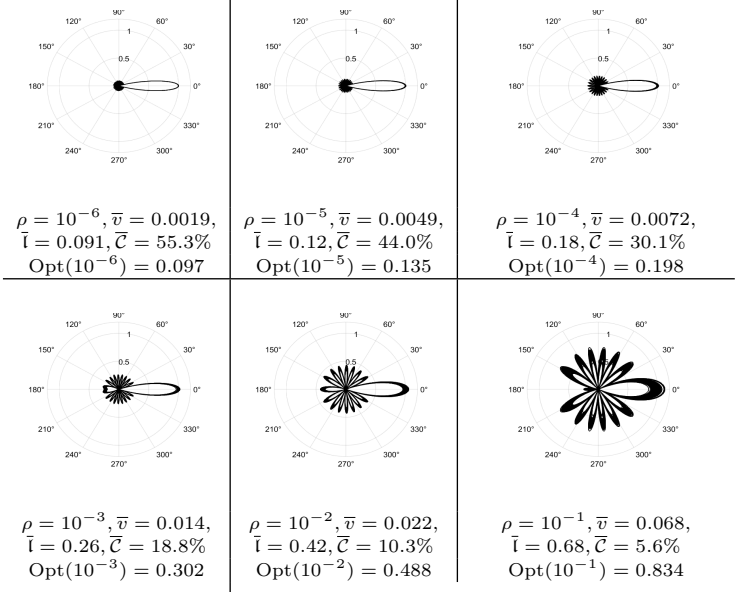
of the "no-errors" diagram corresponding to candidate weights w a perturbation which can be whatever complex number of the modulus not exceeding $\rho\sum_k |w_k|$. Thus, the "robust" – worst-case w.r.t. implementation errors, the perturbation level being ρ – sidelobe level corresponding to candidate weights w is

$$\max_{\ell:\phi_\ell>\gamma}|D_w(\phi_\ell)| + \rho\sum\nolimits_k |w_k|,$$

and the robust counterpart of the system of inequality constraints in (P) is the constraint

$$t \geq \max_{\ell:\phi_\ell>\gamma}|D_w(\phi_\ell)| + \rho\sum\nolimits_k |w_k|. \qquad (C)$$

As about the normalizing equality constraint in (P), formally its robust counterpart is contradictory – we cannot select w_k such that the actual diagram as evaluated at $\phi = 0$ will be exactly one, whatever be the perturbations. *It makes full sense to keep this constraint as is* – in the presence of implementation errors, it will be violated by at most $\rho\sum_k |w_k|$, the same quantity which we



$\rho = 10^{-6}, \overline{v} = 0.0019,$ $\overline{\mathfrak{l}} = 0.091, \overline{\mathcal{C}} = 55.3\%$ $\mathrm{Opt}(10^{-6}) = 0.097$	$\rho = 10^{-5}, \overline{v} = 0.0049,$ $\overline{\mathfrak{l}} = 0.12, \overline{\mathcal{C}} = 44.0\%$ $\mathrm{Opt}(10^{-5}) = 0.135$	$\rho = 10^{-4}, \overline{v} = 0.0072,$ $\overline{\mathfrak{l}} = 0.18, \overline{\mathcal{C}} = 30.1\%$ $\mathrm{Opt}(10^{-4}) = 0.198$
$\rho = 10^{-3}, \overline{v} = 0.014,$ $\overline{\mathfrak{l}} = 0.26, \overline{\mathcal{C}} = 18.8\%$ $\mathrm{Opt}(10^{-3}) = 0.302$	$\rho = 10^{-2}, \overline{v} = 0.022,$ $\overline{\mathfrak{l}} = 0.42, \overline{\mathcal{C}} = 10.3\%$ $\mathrm{Opt}(10^{-2}) = 0.488$	$\rho = 10^{-1}, \overline{v} = 0.068,$ $\overline{\mathfrak{l}} = 0.68, \overline{\mathcal{C}} = 5.6\%$ $\mathrm{Opt}(10^{-1}) = 0.834$

Figure S2II.3. Robust design.
Magnitudes of 100 actual diagrams stemming from optimal solutions to (R)

see in (C). Hopefully, this quantity will be made small by minimization over w of the right hand side in (C).

With the outlined approach the robust w.r.t. implementation errors counterpart of (P) is the convex optimization problem

$$\mathrm{Opt}(\rho) = \min_{t,w} \left\{ t : \begin{array}{l} |\sum_{k=1}^K w_k \mathrm{e}^{2\pi\imath(k-1)\delta\cos(\phi_\ell)}| + \rho\sum_k |w_k| \le t \,\forall(\ell : \phi_\ell > \gamma) \\ \sum_{k=1}^K w_k \mathrm{e}^{2\pi\imath(k-1)\delta} = 1 \\ w_k \in \mathbf{C},\ k \le K \end{array} \right\} \qquad (R)$$

- The results of our experiments with robust designs yielded by (R) are shown in Figure S2II.3. Comparison with similar results for the nominal design speaks for itself loud and clear.

Exercise II.46. Prove the statement "symmetric" the Dubovitski-Milutin Lemma:

> The cone M_* dual to the arithmetic sum of k (closed or not) cones $M^i \subset \mathbf{R}^n$, $i \le k$, is the intersection of the k cones M_*^i dual to M^i.

Solution: By evident reasons, a linear function $f^\top x$ can be nonnegative everywhere on the arithmetic sum of k nonempty sets $M^1 + ... + M^k$ if and only if it is nonnegative on every one of these sets.

Exercise II.47. Prove the following polyhedral version of the Dubovitski-Milutin Lemma:

> Let $M^1, ..., M^k$ be polyhedral cones in $\mathbf{R}^n$, and let $M = \cap_i M^i$. The cone M_* dual to M is the sum of cones M_*^i, $i \le k$, dual to M^i, so that a linear form $e^\top x$ is nonnegative on M if and only it can be represented as the sum of linear forms $e_i^\top x$ nonnegative on the respective cones M_i.

Solution: This is immediate consequence of Proposition II.6.30 combined with the fact that by calculus of polyhedral representations from section 3.3 and the result of Exercise II.39, intersections, sums, and duals of polyhedral cones are polyhedral cones and therefore are closed.

Exercise II.48. [follow-up to Exercise II.47] Let $A \in \mathbf{R}^{m\times n}$ be a matrix with trivial kernel, $e \in \mathbf{R}^n$, and let the set

$$X = \{x : Ax \geq 0, e^\top x = 1\} \qquad (*)$$

be nonempty and bounded. Prove that there exists $\lambda \in \mathbf{R}^m$ such that $\lambda > 0$ and $A^\top\lambda = e$.

Prove "partial inverse" of this statement: *if* $\mathrm{Ker}A = \{0\}$ *and* $e = A^\top\lambda$ *for some* $\lambda > 0$, *the set* $(*)$ *is bounded.*

Solution: Let E be the image space of A, and P be the orthogonal projector of $\mathbf{R}^m$ onto E. Since $\mathrm{Ker}A = \{0\}$, there exists $g \in \mathbf{R}^m$ such that $A^\top g = e$, so that $y \in \mathbf{R}^m$ is representable as Ax with $e^\top x = 1$ if and only if $y \in E$ and $g^\top y = 1$. Therefore

$$Y := \{y = Ax : x \in X\} = \{y \in E : y \geq 0, g^\top y = 1\}$$

Y is cut off the cone $M = \mathbf{R}^m_+ \cap E$ by the linear equality constraint $g^\top y = 1$ and is nonempty and bounded. Clearly M is pointed along with $\mathbf{R}^m_+$ and is nontrivial (since Y is nonempty). Treating M as a cone in Euclidean space E equipped with the inner product $\langle\cdot,\cdot\rangle$ inherited from the standard inner product on $\mathbf{R}^m$, and denoting by f the orthoprojection of g onto E, we have

$$Y = \{y \in M : \langle f, y\rangle = 1\};$$

since Y is nonempty and bounded, and M is closed, nontrivial and pointed, Fact II.6.28.iii states that $f \in \mathrm{int}\, M_*$, where M_* is the dual of the cone $M \subset E$. In other words, for some $r > 0$ and all $f' \in E$, $\|f' - f\|_2 \leq r$, we have $f' \in M_*$. Now let $\overline{\lambda} \in \mathrm{int}\,\mathbf{R}^m_+$ be such that $\|\overline{\lambda}\|_2 \leq r$, and let $\overline{g} = g - \overline{\lambda}$ and $\overline{f} = P\overline{g} \in E$, so that $\|f - \overline{f}\|_2 \leq \|g - \overline{g}\|_2 \leq r$. The latter inequality, due to the origin of r, implies that

$$\langle \overline{f}, y\rangle \geq 0 \ \forall y \in M,$$

or, which is the same,

$$\overline{g}^\top y \geq 0 \ \forall y \in \mathbf{R}^m_+ \cap E.$$

By polyhedral version of Dubovitski-Milutin Lemma (Exercise II.47), there exists $\widetilde{\lambda} \in (\mathbf{R}^m_+)_* = \mathbf{R}^m_+$ and

$$\mu \in E_* := \{\mu : \mu^\top u \geq 0 \,\forall u \in E\} = E^\perp = [\mathrm{Im}A]^\perp = \mathrm{Ker}A^\top$$

such that $\overline{g} = \lambda + \mu$, implying that

$$g = \overline{g} + \overline{\lambda} = \underbrace{[\widetilde{\lambda} + \overline{\lambda}]}_{\lambda>0} + \mu,$$

whence

$$e = A^\top g = A^\top\lambda + A^\top\mu = A^\top\lambda.$$

We have found $\lambda > 0$ with $A^\top\lambda = e$, as required.

To prove "partial inverse", note that if $\lambda > 0$, then the set

$$Z = \{y \in \mathbf{R}^m_+ : \lambda^\top y = 1\}$$

is bounded; when $A^\top\lambda = e$, X is the inverse linear image of Z under the linear embedding $x \mapsto Ax : \mathbf{R}^n \to \mathbf{R}^m$, and therefore X is bounded along with Z.

Exercise II.49. Let E be a linear subspace in $\mathbf{R}^n$, K be a closed cone in $\mathbf{R}^n$, and $\ell(x) : E \to \mathbf{R}$ be a linear (linear, not affine!) function which is nonnegative on $K \cap E$. Which of the following claims are always true:

1. $\ell(\cdot)$ can be extended from E onto the entire $\mathbf{R}^n$ to yield a linear function which is nonnegative on K
2. Assuming $\mathrm{int}\, K \cap E \neq \varnothing$, $\ell(\cdot)$ can be extended from E onto the entire $\mathbf{R}^n$ to yield a linear function which is nonnegative on K.

3. Assuming, in addition to $\ell(x) \geq 0$ for $x \in K \cap E$, that $K = \{x : Px \leq 0\}$ is a polyhedral cone, $\ell(\cdot)$ can be extended from E onto the entire $\mathbf{R}^n$ to yield a linear function which is nonnegative on K.

Solution: The first claim is wrong in general. The simplest counterexample is $K = \mathbf{L}^3$: take a generator of K - an emanating from the origin ray on the boundary of the cone, say, $R = \{x \in \mathbf{R}^3 : x_3 = x_1, x_2 = 0\}$, and let $E = \{x \in \mathbf{R}^3 : x_1 = x_3\}$ be the 2D plane tangent to the surface of the cone along the ray R. Linear function $\ell(x) = x_2 : E \to \mathbf{R}$ is nonnegative on $K \cap E = R$, but cannot be extended to a linear function $f(x) = e^\top x$ on $\mathbf{R}^3$ nonnegative on K; indeed, points $x_\delta = [1; -\delta; 1]$ with $\delta > 0$ are in E, so that we should have $f(x_\delta) = \ell(x_\delta) = -\delta$; at the same time, x_δ belongs to the plane E and $\|x_0 - x_\delta\|_2 = \delta$; since E is tangent to the boundary of K at x_0, there are points $x_\delta^+ \in K$ with $\|x_\delta - x_\delta^+\|_2 \leq O(1)\delta^2$ (you can take $x_\delta^+ = [1; -\delta; \sqrt{1+\delta^2}]$), so that $f(x_\delta^+) \leq f(x_\delta) + \|e\|_2\|x_\delta^+ - x_\delta\|_2 \leq -\delta + O(1)\|e\|_2\delta^2$, that is, $f(x_\delta^+) < 0$ for all small $\delta > 0$, which is a desired contradiction, since $x_\delta^+ \in K$, and $f(x)$ on K is nonnegative.

The second claim is true by Dubovitski-Milutin Lemma (DML). Indeed, in the notation of this Lemma, set $M_1 = K$ and $M_2 = E$, thus satisfying the premise of Lemma. Extend $\ell(\cdot)$ to a whatever linear form $e^\top x$ of $x \in \mathbf{R}^n$; this form is nonnegative on $M_1 \cap M_2$ and therefore, by DML, can be represented as $g^\top x + h^\top x$ with $g^\top x$ nonnegative for $x \in M_1 = K$ and $h^\top x$ nonnegative when $x \in M_2 = E$, that is, with $h^\top x = 0$ for $x \in E$ (E is a linear subspace!). The desired extension of $\ell(x)$ from E to a nonnegative on K linear form is $x \mapsto g^\top x$.

The third claim is true. Indeed, E is a linear subspace and thus is a polyhedral cone. We can find a vector $e \in \mathbf{R}^n$ such that $\ell(x) = e^\top x$ for $x \in E$. Applying the result of Exercise II.47 to the polyhedral cones $M^1 = K$ and $M^2 = E$, we conclude that under the premise of item 3 we have $e = e_1 + e_2$ with $e_1 \in M_*^1 = K_*$ and $e_2 \in M_*^2 = E^\perp$, implying that $e_1^\top x$ is the desired linear form nonnegative on K and equal to $\ell(x)$ on E. ■

Exercise II.50. Let $n > 1$. Is the unit $\|\cdot\|_2$-ball $B_n = \{x \in \mathbf{R}^n : \|x\|_2 \leq 1\}$ a polyhedral set? Justify your answer.

Solution: The answer, of course, is "no". Indeed, were B_n polyhedral, the set of extreme points of B_n would be finite (Corollary II.7.2), which contradicts the evident fact that $\mathrm{Ext}(B_n) = S_n := \{x \in \mathbf{R}^n : \|x\|_2 = 1\}$, and this set is infinite when $n > 1$. To show that $\mathrm{Ext}(B_n) = S_n$ is clearly the same as to show that every point $e \in S_n$ is extreme; to this end note that when $e \pm h \in B_n$, we have $2\|e\|_2^2 + 2\|h\|_2^2 = \|e+h\|_2^2 + \|e-h\|_2^2 \leq 2$, and the resulting inequality implies that $h = 0$ due to $\|e\|_2 = 1$.

It is worthy of mentioning that *"for all practical purposes," B_n is a simple polyhedral set.* Specifically, it is known (see [Nem24, section 1.4]) that for every $\epsilon \in (0, 1/2)$ and every n one can explicitly write down system of $O(1)n\ln(1/\epsilon)$ linear inequalities with $O(1)n\ln(1/\epsilon)$ variables such that the projection of the solution set of this system onto the plane of the first n variables is in-between B_n and $(1+\epsilon)B_n$. When $\epsilon = 1.0\text{e}{-17}$, usual computer does not distinguish between 1 and $1+\epsilon$, so that for all practical purposes B_n admits explicit polyhedral representation; to get $\epsilon = 1.\text{e}{-17}$, this representation should involve $\approx 79n$ linear inequalities on $\approx 28n$ variables.

Exercise II.51. The unit box $\{x \in \mathbf{R}^n : -1 \leq x_i \leq 1, i \leq n\}$ is cut off $\mathbf{R}^n$ by a system of $m = 2n$ linear inequalities and is a nonempty and bounded polyhedral set. However, when we eliminate any inequality from this system, the solution set of the resulting system becomes unbounded. To see that this situation is in a sense extreme, prove the following claim:

> *Consider the solution set of a system of m linear inequalities in n variables x, i.e., the set*
>
> $$X := \{x \in \mathbf{R}^n : \ Ax \leq b\},$$
>
> *where $A = [a_1^\top; a_2^\top; \ldots; a_m^\top]$. Suppose that X is nonempty and bounded. Then, whenever $m > 2n$, one can drop from this system a properly selected inequality in such a way that the solution set of the resulting subsystem remains bounded.*

A provocative follow-up: *Is it possible to cut off from $\mathbf{R}^{1000}$ a bounded set by using* only a *single linear inequality?*

Solution: Suppose X is nonempty and bounded, and let $m > 2n$. Recall from Fact II.6.18 that a nonempty closed convex set is bounded if and only if its recessive cone is trivial. Then, as X is closed (it is polyhedral!), we have $\mathrm{Rec}(X) = \{0\}$. Moreover, based on Fact II.6.20 we have $\mathrm{Rec}(X) = \{x : Ax \le 0\}$. Thus, the closed cone $K := \{x : Ax \le 0\}$ is trivial, or, which is the same by Fact II.6.28, the dual K_* of this cone is the entire $\mathbf{R}^n$. On the other hand, as is explained immediately after Proposition II.6.26, $K_* = \mathrm{Cone}(\{-a_1, \ldots, -a_m\})$. Thus, $\mathrm{Cone}(\{-a_1, \ldots, -a_m\}) = \mathbf{R}^n$, implying, in particular, that $\mathrm{rank}\, A = n$ (since the conic hull of $-a_1, \ldots, -a_m$ belongs to the linear span of the collection $a_1, \ldots, a_m$). Without loss of generality, we can assume that $a_1, \ldots, a_n$ are linearly independent. Hence, the vector $\overline{a} := \sum_{i=1}^n a_i$ belongs to $K_* = \mathbf{R}^n$ and therefore $\overline{a}$ is a conic combination of vectors $-a_1, \ldots, -a_m$. Then, by conic version of Caratheodory's Theorem (Fact I.2.7) we can select n vectors $a_{i_1}, \ldots, a_{i_n}$ from the given m vectors $a_1, \ldots, a_m$ in such a way that $\overline{a}$ is a conic combination of the vectors $-a_{i_1}, \ldots, -a_{i_n}$. As a result, all vectors of the form

$$\sum_{i=1}^n (1 - \mu_i) a_i, \quad \text{where } \mu \ge 0, \tag{$*$}$$

are conic combinations of vectors from the collection $\{-a_1, -a_2, \ldots, -a_n, -a_{i_1}, -a_{i_2}, \ldots, -a_{i_n}\}$. All vectors $\sum_{i=1}^n z_i a_i$ with $\|z\|_\infty \le 1$ admit a representation of the form $(*)$ and therefore, as we have just seen, they belong to the cone $\mathrm{Cone}(\{-a_1, -a_2, \ldots, -a_n, -a_{i_1}, -a_{i_2}, \ldots, -a_{i_n}\})$. Since $a_1, \ldots, a_n$ are linearly independent, the set of linear combinations of these vectors with coefficients of magnitude ≤ 1 contains a neighborhood of the origin. Thus, the cone

$$\mathrm{Cone}(\{-a_1, -a_2, \ldots, -a_n, -a_{i_1}, -a_{i_2}, \ldots, -a_{i_n}\})$$

contains a neighborhood of the origin and is therefore the entire $\mathbf{R}^n$. On the other hand, by the same argument as above, this cone is dual to the cone

$$\left\{x : \ a_i^\top x \le 0, \, i \le n, \ a_{i_j}^\top x \le 0, \, j \le n\right\},$$

so that the latter cone is trivial. Thus, the recessive cone of the set

$$X^+ := \left\{x : \ a_i^\top x \le b_i, \, i \le n, \ a_{i_j}^\top x \le b_{i_j}, \, j \le n\right\}$$

is trivial, and therefore this set is bounded. Thus, we conclude that we can extract a carefully selected set of $2n$ constraints from the constraints $a_i^\top x \le b_i$, $i \le m$, such that they still result in a bounded set in $\mathbf{R}^n$. ■

The answer to the follow-up question is positive: the insolvable linear inequality $0^\top x \le -1$ cuts off $\mathbf{R}^{1000}$ the empty set which of course is bounded.

Exercise II.52. [computational study] let $\omega^N = (\omega_1, ..., \omega_N)$ be an N-element i.i.d. sample drawn from the standard Gaussian distribution (zero mean, unit covariance) on $\mathbf{R}^d$. How many extreme points are there in the convex hull of the points from the sample?

1. Consider the planar case $d = 2$ and think how to list extreme points of $\mathrm{Conv}\{\omega_1, ..., \omega_N\}$. Fill the following table:

N	2	4	8	16	32	64	128
U							
M							
L							

   where U is the maximal, M is the mean, and L is the minimal # of extreme points observed when processing 100 samples ω^N of a given cardinality.

   Solution: The simplest way to check whether a point, say, ω_1, from N-element sample ω^N of 2D points is or is not extreme, is to look at $N - 1$ lines ℓ_j, $j = 2, ..., N$, linking ω_1 and ω_j. When no triple of points from the sample belong to a common line (which happens with probability 1), ω_1 is extreme point of $\mathrm{Conv}(\omega^N)$ if and only if all points of the sample are on one side of one of these lines.

Here are our results:

N	2	4	8	16	32	64	128
U	2	8	14	18	20	24	26
M	2.00	7.36	10.22	12.60	14.72	16.54	18.36
L	2	6	8	8	10	10	10
E	4	7.30	10.06	12.37	14.34	17.23	17.77

(the E-row is the answer to the question of item 2).

2. Think how to upper-bound the expected number of extreme points in the set $W = \mathrm{Conv}(\omega^N)$.

   Solution: Similarly to the previous item, ignoring "degenerate" samples with total probability mass 0, ω_1 is an extreme point of W if one can select $d-1$ points $\omega_{i_2}, ..., \omega_{i_d}$ with $2 \le i_2 < i_3 < ...$ in such a way that the entire sample is on one side of the hyperplane passing through $\omega_1, \omega_{i_2}, \omega_{i_3}, ..., \omega_{i_d}$. Probability π of this outcome clearly is independent of what is the collection $i_2 < i_3 < ... < i_d$ and can be reliably estimated via simulation, namely, as follows: we simulate $M \gg 1$ times d-element sample ω^d, measure the Euclidean distance from the hyperplane containing this sample to the origin, and recover the distribution P of this distance. We clearly have

$$\pi = \int_0^\infty [\psi^{N-d}(s) + (1-\psi(s))^{N-d}] dP(s),$$

   where ψ is the cumulative distribution function of $\mathcal{N}(0,1)$ random variable, and we can estimate π by substituting the expectation w.r.t. P with expectation w.r.t. the empirical approximation of P. After π is estimated, we can upper-bound the probability for ω_1 to be an extreme point of W by the quantity $\theta = \binom{N-1}{d-1}\pi$, resulting in the upper bound θN on the expected number of extreme points of W.

Exercise II.53. [computational study] Given positive integers m, n, with $n \ge 2$, consider randomly generated system $Ax \le b$ of m linear inequalities with n variables. We assume that A, b are generated by drawing the entries, independently of each other, from $\mathcal{N}(0,1)$.

1. Consider the planar case $n = 2$. For $m = 2, 4, 8, 16$, generate 100 samples of $m \times 2$ systems and fill the following table:

m	2	4	8	16
F				
B				

   where F is the number of feasible systems, and U is the number of feasible systems with bounded solution sets.

   Solution: see item 3 below.

Intermezzo: related theoretical results originating from [Nem24, Exercise 2.23] are as follows. Given positive integers m, n with $n \ge 2$, consider homogenous system $Ax \le 0$ of m inequalities with n variables. We call this system *regular*, if its matrix A is regular, regularity of a matrix B meaning that all square submatrices of B are nonsingular. Clearly, the entries of a regular matrix are nonzero, and when a $p \times q$ matrix B is drawn at random from a probability distribution on $\mathbf{R}^{p\times q}$ which has a density w.r.t the Lebesgue measure, B is regular with probability 1.

Given regular $m \times n$ homogeneous system of inequalities $Ax \le 0$, let $g_i(x) = \sum_{j=1}^n A_{ij}x_j$, $i \le m$, so that g_j are nonconstant linear functions. Setting $\Pi_i = \{x : g_i(x) = 0\}$, we get a collection of m hyperplanes in $\mathbf{R}^n$ passing through the origin. For a point $x \in \mathbf{R}^n$, the *signature* of x is, by definition, the m-dimensional vector $\sigma(x)$ of signs of the reals $g_i(x)$, $1 \le i \le m$. Denoting by Σ the set of all m-dimensional vectors with entries ± 1, for $\sigma \in \Sigma$ the set $\mathcal{C}_\sigma = \{x : \sigma(x) = \sigma\}$ is either empty, or is a nonempty open convex set; when it is nonempty, let us call it a *cell* associated with A, and the corresponding σ – an A-feasible signature. Clearly, for regular system, $\mathbf{R}^n$ is the union of all hyperplanes Π_i and all cells associated with A. It turns out that

The number $N(m,n)$ of cells associated with a regular homogeneous $m \times n$ system $Ax \le 0$ is independent of the system and is given by a simple recurrence:

$$\begin{array}{rcl} N(1,2) & = & 2 \\ m \ge 2, n \ge 2 & \Longrightarrow & N(m,n) = N(m-1,n) + N(m-1,n-1) \quad [N(m,1) = 2,\ m \ge 1]. \end{array}$$

Next, when A is drawn at random from probability distribution P on $\mathbf{R}^{m\times n}$ which possesses *symmetric* density p, that is, such that $p([a_1^\top; a_2^\top; ...; a_m^\top]) = p([\epsilon_1 a_1^\top; \epsilon_2 a_2^\top; ...; \epsilon_m a_m^\top])$ for all $A = [a_1^\top; a_2^\top; ...; a_m^\top]$ and all $\epsilon_i = \pm 1$, then *the probability for a vector $\sigma \in \Sigma$ to be an A-feasible signature is*

$$\pi(m,n) = N(m,n)/2^m.$$

In particular, the probability for the system $Ax \le 0$ to have a solution set with a nonempty interior (this is nothing but A-feasibility of the signature $[-1; ...; -1]$) is $\pi(m,n)$.
The inhomogeneous version of these results is as follows. An $m \times n$ system of linear inequalities $Ax \le b$ is called regular, if the matrix $[A, -b]$ is regular. Setting $g_i(x) = \sum_{j=1}^n A_{ij}x_j - b_i$, $i \le n$, the $[A,b]$-signature of x is, as above, the vector of signs of the reals $g_i(x)$. For $\sigma \in \Sigma$, the set $\mathcal{C}_\sigma = \{x : \sigma(x)) = \sigma\}$ is either empty, or is a nonempty open convex set; in the latter case, we call $\mathcal{C}_\sigma$ an $[A,b]$-cell, and call σ an $[A,b]$-feasible signature. Setting $\Pi_i = \{x : g_i(x) = 0\}$, we get m hyperplanes in $\mathbf{R}^n$, and the entire $\mathbf{R}^n$ is the union of those hyperplanes and all $[A,b]$-cells. It turns out that

The number $N(m,n)$ of cells associated with a regular $m \times n$ system $Ax \le b$ is independent of the system and is equal to $\frac{1}{2}N(m+1,n+1)$.

In addition, when $m \times (n+1)$ matrix $[A,b]$ is drawn at random from a probability distribution on $\mathbf{R}^{m\times(n+1)}$ possessing a symmetric density w.r.t. the Lebesgue distribution, *the probability for every $\sigma \in \Sigma$ to be $[A,b]$-feasible signature is*

$$\overline{\pi}(m,n) = N(m+1,n+1)/2^{m+1}.$$

In particular, the probability for the system $Ax \le b$ to be strictly feasible is $\overline{\pi}(m,n)$.

2. Accompanying exercise: Prove that if A is $m \times n$ regular matrix, then the system $Ax \le 0$ has a nonzero solution if and only if the system $Ax < 0$ is feasible. Derive from this fact that if $[A,b]$ is regular, then the system $Ax \le b$ is feasible if and only if it is strictly feasible, and that when the system $Ax \le 0$ has a nonzero solution, the system $Ax \le b$ is strictly feasible for every b.

Solution: Let us start with the first claim. The only nontrivial part of it is that for regular A, the existence of nonzero x such that $Ax \le 0$ implies feasibility of the system $Ax < 0$. Let us lead to contradiction the assumption that A is an $m \times n$ regular matrix such that the system $Ax \le 0$ has a nonzero solution $\overline{x}$, and at the same time the system $Ax < 0$ is infeasible. By the General Theorem of the Alternative, infeasibility of the system $Ax < 0$ implies that a nontrivial linear combination of rows of A with nonnegative coefficients is 0, or, which is the same, denoting by $a_i^\top$ the rows of A, the origin in $\mathbf{R}^n$ is a convex combination of a_i. W.l.o.g. we can assume that positive coefficients in this combination are associated with $a_1, ..., a_k$, for some $k \le m$. From the relations $\sum_{i=1}^k \lambda_i a_i = 0$, $\lambda_i > 0, i \le k$, and $a_i^\top \overline{x} \le 0$ it follows that $\overline{x}^\top a_i = 0$, $i = 1, ..., k$. Since $\overline{x} \ne 0$, it follows that a_i, $i \le k$, belong to an $(n-1)$-dimensional subspace of $\mathbf{R}^n$, so that the affine dimension of the affine span of $a_1, ..., a_k$ is at most $n-1$. Since 0 is a convex combination of $a_,.., a_k$, by Caratheodory Theorem 0 is a convex combination of $\overline{k} \le \min[k,n]$ of vectors from the collection $a_1, .., a_k$, implying that properly selected $\overline{k}$ rows in A are linearly dependent, contradicting regularity of A. As a corollary, if A is regular and $Ax \le 0$ for some nonzero x, the system $Ax < 0$ is solvable, which, of course, implies that the system $Ax \le b$ is solvable for every b. ■

To justify the second claim, it suffices to verify that if the system $Ax \le b$ is feasible and $[A,b]$ is regular, then the system is strictly feasible. To this end assume that the premise of this claim holds true, so that for some $\bar{x}$ it holds $\bar{b} := A\bar{x} \le b$. For small $\bar{\epsilon} > 0$ and all $e \in \mathbf{R}^n$, $\|e\|_\infty \le \bar{\epsilon}$, we have $\underbrace{[A, -\bar{b}; e^\top, -1]}_{B[e]}\underbrace{[\bar{x}; 1]}_{\bar{y}} = [A\bar{x} - \bar{b}; e^\top\bar{x} - 1] \le 0$. On the other hand, selecting e from the uniform distribution of the box $\|e\|_\infty \le \epsilon$, with $0 < \epsilon \le \bar{\epsilon}$, it is easily seen that when $\epsilon > 0$ is small enough,

the matrix $B[e]$ is regular with probability 1. Thus, we may assume that $B[e]$ is regular, and, as we have seen, the system $B[e]y \leq 0$ in variables y has a nonzero solution $\bar{y}$. By the already proved first claim, it follows that the system $B[e]y < 0$ has a solution $\widetilde{y}$. As a result, for every $\lambda \in (0,1)$, the vector $y_\lambda = (1-\lambda)\bar{y} + \lambda\widetilde{y}$ satisfies $B[e]y_\lambda < 0$. For small positive λ, y_λ is of the form $[x_\lambda; t_\lambda]$ with $t_\lambda > 0$; for such a λ, relation $B[e]y_\lambda < 0$ implies that $Ax_\lambda - t_\lambda \bar{b} < 0$, whence $A[x_\lambda/t_\lambda] < \bar{b} \leq b$, that is, the system $Ax \leq b$ is strictly feasible. ■

Note: by Accompanying Exercise, in the situations described in Intermezzo, probability $\overline{\pi}(m,n)$ for an $m \times n$ system $Ax \leq b$ to be strictly feasible is the same as the probability to be feasible, and the probability to have an unbounded feasible set (i.e., to be feasible and such that $Ah \leq 0$ for some nonzero h) is the same as the probability $\pi(m,n)$ for the signature $[-1,...,-1]$ to be A-feasible.

3. Use the results from Intermezzo to compute the expected values of F and B, see item 1.

Solution: Here are our results:

m	2	4	8	16
F	100	72	18	0
$\mathbf{E}\{F\}$	100	68.75	14.45	0.21
B	0	18	15	0
$\mathbf{E}\{B\}$	0	18.75	8.20	0.16

Exercise II.54. [computational study]

1. For $\nu = 1, 2, ..., 6$, generate 100 systems of linear inequalities $Ax \leq b$ with $n = 2^\nu$ variables and $m = 2n$ inequalities, the entries in A, b being drawn, independently of each other, from $\mathcal{N}(0.1)$. Fill the following table:

n	2	4	8	16	32	64
F						
$\mathbf{E}\{F\}$						
B						

F: # of feasible systems in sample;
B: # of feasible systems with bounded soultion sets

To compute the expected value of F, use the results from [Nem24, Exercise 2.23] cited in item 2 of Exercise II.53.

2. Carry out experiment similar to the one in item 1, but with $m = n + 1$ rather than $m = 2n$.

n	2	4	8	16	32	64
F						
$\mathbf{E}\{F\}$						
B						
$\mathbf{E}\{B\}$						

F: # of feasible systems in sample;
B: # of feasible systems with bounded soultion sets

Solution: Our results, rounded to 2 digits after the dot, are as follows:

1. $m = 2n$:

n	2	4	8	16	32	64
F	74	72	64	57	50	53
$\mathbf{E}\{F\}$	68.75	63, 67	59.82	57.00	54.97	53.52
B	17	16	7	5	3	3
$\mathbf{E}\{B\}$	18.75	13.67	9.82	7.00	4.97	3.52

F: # of feasible systems in sample;
B: # of feasible systems with bounded solution sets

2. $m = n + 1$:

n	2	4	8	16	32	64
F	92	96	100	100	100	100
$\mathbf{E}\{F\}$	87.50	96.88	100.00	100.00	100.00	100.00
B	11	4	0	0	0	0
$\mathbf{E}\{B\}$	12.50	3.13	0.20	0.00	0.00	0.00

F: # of feasible systems in sample;
B: # of feasible systems with bounded solution sets

Exercises from Part III

15.1 Around convex functions

Exercise III.1. Which of the functions below are convex on the indicated domains:

- $f(x) \equiv 1$ on $\mathbf{R}$

  Solution: convex
- $f(x) = x$ on $\mathbf{R}$

  Solution: convex
- $f(x) = |x|$ on $\mathbf{R}$

  Solution: convex
- $f(x) = -|x|$ on $\mathbf{R}$

  Solution: nonconvex
- $f(x) = -|x|$ on $\mathbf{R}_+ = \{x \in \mathbf{R} : \ x \geq 0\}$

  Solution: convex
- $f(x) = |2x - 3|$ on $\mathbf{R}$

  Solution: convex
- $f(x) = |2x^2 - 3|$ on $\mathbf{R}$

  Solution: nonconvex
- $\exp\{x\}$ on $\mathbf{R}$

  Solution: convex
- $\exp\{x^2\}$ on $\mathbf{R}$

  Solution: convex
- $\exp\{-x^2\}$ on $\mathbf{R}$

  Solution: nonconvex
- $\exp\{-x^2\}$ on $\{x \in \mathbf{R} : \ x \geq 100\}$

  Solution: convex
- $\ln(x)$ on $\{x \in \mathbf{R} : \ x > 0\}$

  Solution: nonconvex
- $-\ln(x)$ on $\{x \in \mathbf{R} : \ x > 0\}$

  Solution: convex

Exercise III.2.

1. Prove the following fact:
   For every $C_i \in \mathbf{S}^m_+$, $i \leq I$, satisfying $\sum_{i \in I} C_i = I_m$ and for every $\lambda_i \in \mathbf{R}$, we have

$$\mathrm{Tr}\left(\left(\sum\nolimits_{i \in I} \lambda_i C_i\right)^2\right) \leq \mathrm{Tr}\left(\sum\nolimits_{i \in I} \lambda_i^2 C_i\right).$$

Solution: Define $\phi_{ij} := \mathrm{Tr}(C_iC_j)$ so that $\phi_{ij} = \phi_{ji} \geq 0$ as $\mathbf{S}^m_+$ is a self-dual cone (see section D.2.2). Thus,

$$\begin{aligned}
\mathrm{Tr}\left(\left(\sum_{i\in I}\lambda_i C_i\right)^2\right) &= \sum_{i\in I}\sum_{j\in I}\lambda_i\lambda_j\phi_{ij}\\
&= \sum_{i\in I}\sum_{j\in I}[\lambda_i\sqrt{\phi_{ij}}][\lambda_j\sqrt{\phi_{ij}}]\\
&\leq \left(\sum_{i\in I}\sum_{j\in I}\phi_{ij}\lambda_i^2\right)^{1/2}\left(\sum_{i\in I}\sum_{j\in I}\phi_{ij}\lambda_j^2\right)^{1/2}\\
&= \sum_{i\in I}\sum_{j\in I}\phi_{ij}\lambda_i^2\\
&= \sum_{i\in I}\sum_{j\in I}\mathrm{Tr}(C_iC_j)\lambda_i^2\\
&= \sum_{i\in I}\mathrm{Tr}\left(C_i\sum_{j\in I}C_j\right)\lambda_i^2\\
&= \sum_{i\in I}\mathrm{Tr}(C_i)\lambda_i^2. \qquad [\text{since } \sum_{j\in I}C_j = I_m]
\end{aligned}$$

2. Recall from Example III.10.4 in section 10.2 that for $a_i \geq 0$, $\sum_i a_i > 0$ the function $\ln(\sum_i a_i \exp(\lambda_i))$ is a convex function of λ. Prove the following matrix analogy of this fact:

   For every $A_i \in \mathbf{S}^m_+$, $1 \leq i \leq I$ such that $\sum_i A_i \succ 0$, the function

   $$f(\lambda) = \ln \mathrm{Det}\left(\sum_i \exp(\lambda_i)A_i\right) : \mathbf{R}^I \to \mathbf{R}$$

   is convex.

   Solution: Invoking Examples C.7-8 from section C.1.6, we have

   $$\begin{array}{rcl}
   Df(\lambda)[d\lambda] & = & \mathrm{Tr}\left(\left[\sum_i \mathrm{e}^{\lambda_i}A_i\right]^{-1}\left[\sum_i d\lambda_i \mathrm{e}^{\lambda_i}A_i\right]\right)\\
   & & [B_i = \mathrm{e}^{\lambda_i}A_i \succeq 0, B = \sum_i B_i \succ 0, C_i = B^{-1/2}B_iB^{-1/2},\\
   & & \text{ so that } C_i \succeq 0, \sum_i C_i = I_m]\\
   & = & \mathrm{Tr}\left(B^{-1}\sum_i B_i d\lambda_i\right) = \mathrm{Tr}\left(\sum_i d\lambda_i C_i\right),\\
   D^2f(\lambda)[d\lambda, d\lambda)] & = & -\mathrm{Tr}\left(B^{-1}\left[\sum_i d\lambda_i B_i\right]B^{-1}\left[\sum_i d\lambda_i B_i\right]\right)\\
   & & +\mathrm{Tr}\left(B^{-1}\sum_i d\lambda_i^2 B_i\right)\\
   & = & -\mathrm{Tr}\left(B^{-1/2}\left[\sum_i d\lambda_i B_i\right]B^{-1}\left[\sum_i d\lambda_i B_i\right]B^{-1/2}\right)\\
   & & +\mathrm{Tr}\left(B^{-1/2}\left[\sum_i d\lambda_i^2 B_i\right]B^{-1/2}\right)\\
   & = & \mathrm{Tr}\left(\sum_i d\lambda_i^2 C_i\right) - \mathrm{Tr}\left(\left[\sum_i d\lambda_i C_i\right]^2\right),\\
   & \geq & 0 \text{ [by item 1]}
   \end{array}$$

   implying that f is convex (Corollary III.10.4).

3. Let A_i, $i \leq I$, be as in item 2. Is it true that the function

   $$g(x) = \ln \mathrm{Det}(\sum_i x_i^{-1}A_i) : \{x \in \mathbf{R}^I : x > 0\} \to \mathbf{R}$$

   is convex?

   Solution: The answer is "yes." Indeed, the function f from item 2 is convex and clearly is nondecreasing in λ_i; g is obtained from f by convex substitution of the argument $\lambda_i = -\ln(x_i)$, $i \leq I$.

4. Let B_i, $i \leq I$, be $m_i \times n$ matrices such that $\sum_i B_i^\top B_i \succ 0$, and let

   $$\Lambda = \{\lambda := (\lambda_1, ..., \lambda_I) : \lambda_i \in \mathbf{S}^{m_i}, \lambda_i \succ 0, i \leq I\}.$$

   Prove that the function

   $$h(\lambda) = \ln \mathrm{Det}\left(\sum_i B_i^\top \lambda_i^{-1} B_i\right) : \Lambda \to \mathbf{R}$$

   is convex.

Solution: We have

$$\begin{array}{ll} & \lambda\in\Lambda, t\geq h(\lambda)\\ \Longleftrightarrow & \exists V\succ 0: V^{-1}\succeq \sum_i B_i^\top\lambda_i^{-1}B_i, -\ln\mathrm{Det}(V)\leq t\\ \Longleftrightarrow & \exists V\succ 0: \left[\begin{array}{c|c|c|c} V^{-1} & B_1^\top & \cdots & B_I^\top\\ \hline B_1 & \lambda_1 & & \\ \hline \vdots & & \ddots & \\ \hline B_I & & & \lambda_I\end{array}\right]\succeq 0, -\ln\mathrm{Det}(V)\leq t\\ & \text{[by Schur Complement Lemma]}\\ \Longleftrightarrow & \exists V\succ 0: -\ln\mathrm{Det}(V)\leq t, \mathrm{Diag}\{\lambda_1,...,\lambda_I\}\succeq [B_1;...;B_I]V[B_1^\top,...,B_I^\top]\\ & \text{[by Schur Complement Lemma]}\end{array}$$

Taking into account that the function $-\ln\mathrm{Det}(V):\mathrm{int}\,\mathbf{S}^n_+\to\mathbf{R}$ is convex, we conclude that the epigraph of h is the projection of a convex set in (t,λ,V)-space onto the subspace of (t,λ)-variables and is therefore convex. ■

5. Let $B_i, i\leq I$, and Λ be as in the previous item. Prove that the matrix-valued function

$$F(\lambda)=\left[\sum_i B_i^\top\lambda_i^{-1}B_i\right]^{-1}:\Lambda\to\mathrm{int}\,\mathbf{S}^n_+$$

is $\succeq$-concave, that is, the $\succeq$-hypograph

$$\{(\lambda,Y):\lambda\in\Lambda, Y\preceq F(\lambda)\}$$

of the function is convex.

Solution: The values of F on λ are positive definite, implying that the set in question is convex if and only if the set

$$\mathcal{E}=\{(\lambda,Y):\lambda\in\Lambda,\exists V\succ 0: Y\preceq V\preceq F(\lambda)\}$$

is convex. When $V\succ 0$ and $\lambda\in\Lambda$, one has

$$V\preceq F(\lambda)\iff V^{-1}\succeq [F(\lambda)]^{-1}=\sum_i B_i^\top\lambda_i^{-1}B_i$$

(Exercise D.5), implying that

$$\begin{array}{l}\mathcal{E}=\{(\lambda\in\Lambda,Y):\exists V\succ 0: Y\preceq V\preceq F(\lambda)\}=\{(\lambda\in\Lambda,Y):\exists V\succ 0: Y\preceq V, V^{-1}\succeq\sum_i B_i^\top\lambda_i^{-1}B_i\}\\ =\{(\lambda\in\Lambda,Y):\exists V: Y\preceq V, \left[\begin{array}{c|c|c|c} V^{-1} & B_1^\top & \cdots & B_I^\top\\ \hline B_1 & \lambda_1 & & \\ \hline \vdots & & \ddots & \\ \hline B_I & & & \lambda_I\end{array}\right]\succeq 0\},\ \text{[Schur Complement Lemma]}\\ =\{(\lambda\in\Lambda,Y):\exists V\succ 0: V\succeq Y, \mathrm{Diag}\{\lambda_1,...,\lambda_I\}\succeq [B_1;...;B_I]V[B_1^\top,...,B_I^\top]\}\\ \text{[Schur Complement Lemma]}\end{array}$$

We see that $\mathcal{E}$ is the projection of the convex set in the space of (λ,Y,V)-variables onto the plane of (λ,Y)-variables, and thus $\mathcal{E}$ is convex. ■

Exercise III.3. A function f defined on a convex set Q is called log-convex on Q, if it takes real positive values on Q and the function $\ln f$ is convex on Q. Prove that

- a log-convex on Q function is convex on Q
- the sum (more generally, linear combination with positive coefficients) of two log-convex functions on Q also is log-convex on the set.

Solution: If $f(x)=e^{h(x)}$ and h is convex, then so is f, as superposition of a convex monotone function e^x and convex function h. If $f(x)=\lambda_1 f_1(x)+\lambda_2 f_2(x)$ with $f_i(x)=e^{h_i(x)}$, where h_i are convex and $\lambda_i>0$, $i=1,2$, then $f(x)=e^{h(x)}$ with $h(x)=\ln(e^{h_1(x)+\ln\lambda_1}+e^{h_2(x)+\ln\lambda_2})$. Since $\ln(e^u+e^v)$ is convex and monotone function of u,v, we conclude that h is convex along with $h_1(x)$, $h_2(x)$. ■

Exercise III.4. [Law of Diminishing Marginal Returns] Consider optimization problem

$$\mathrm{Opt}(r) = \max_x \{f(x) : G(x) \le r \ \&\ x \in X\} \qquad (P[r])$$

where $X \subset \mathbf{R}^n$ is nonempty convex set, $f(\cdot) : X \to \mathbf{R}$ is concave, and $G(x) = [g_1(x); ...; g_m(x)] : X \to \mathbf{R}^m$ is vector-function with convex components, and let $\mathcal{R}$ be the set of those r for which $(P[r])$ is feasible. Prove that

1. $\mathcal{R}$ is a convex set with nonempty interior and this set is monotone, meaning that when $r \in \mathcal{R}$ and $r' \ge r$, one has $r' \in \mathcal{R}$.

Solution: we clearly have $\mathcal{R} = \cup_{x \in X}\{r : r \ge G(x)\}$, and the right hand side set clearly has nonempty interior and is monotone. To prove that $\mathcal{R}$ is convex, let $r, r' \in \mathcal{R}$ and $\lambda \in [0,1]$. For properly selected $x, x' \in X$ we have $G(x) \le r$, $G(x') \le r'$, which combines with convexity of X and G to imply that $\lambda x + (1-\lambda)x' \in X$ and $G(\lambda x + (1-\lambda)x') \le \lambda r + (1-\lambda)r'$, so that $\lambda r + (1-\lambda)r' \in \mathcal{R}$. ■

2. The function $\mathrm{Opt}(r) : \mathcal{R} \to \mathbf{R} \cup \{+\infty\}$ satisfies the concavity inequality:

$$\forall (r, r' \in \mathcal{R}, \lambda \in [0,1]) : \mathrm{Opt}(\lambda r + (1-\lambda)r') \ge \lambda \mathrm{Opt}(r) + (1-\lambda)\mathrm{Opt}(r'). \qquad (!)$$

Solution: Let r, r', λ satisfy the premise in (!). There is nothing to prove when $\lambda = 0$ or when $\lambda = 1$, so let $\lambda \in (0,1)$. Let us select $s < \mathrm{Opt}(r)$ and $s' < \mathrm{Opt}(r')$, so that there exist $x, x' \in X$ such that $G(x) \le r$, $f(x) \ge s$. $G(x') \le r'$, $f(x') \ge s'$. Since X is convex, the components of G are convex, and f is concave, we have

$$\lambda x + (1-\lambda)x' \in X \ \&\ G(\lambda x + (1-\lambda)x') \le \lambda r + (1-\lambda)r' \ \&\ f(\lambda x + (1-\lambda)x') \ge \lambda s + (1-\lambda)s',$$

implying that $\mathrm{Opt}(\lambda r + (1-\lambda)r') \ge \lambda s + (1-\lambda)s'$. In the resulting inequality, properly selecting $s < \mathrm{Opt}(r)$ and $s' < \mathrm{Opt}(r')$, the right hand side can be made arbitrarily large when $\mathrm{Opt}(r)$ and/or $\mathrm{Opt}(r')$ are $+\infty$, and can be made arbitrarily close to $\lambda\mathrm{Opt}(r) + (1-\lambda)\mathrm{Opt}(r')$ when both $\mathrm{Opt}(r)$ and $\mathrm{Opt}(r')$ are finite, and the concavity inequality follows. ■

3. If $\mathrm{Opt}(r)$ is finite at some point $\overline{r} \in \mathrm{int}\,\mathcal{R}$, then $\mathrm{Opt}(r)$ is real-valued everywhere on $\mathcal{R}$. Moreover, when $X = \mathbf{R}^n$ and f and the components of G are affine, so that $(P[r])$ is an LP program, we can replace in the above claim the inclusion $r \in \mathrm{int}\,\mathcal{R}$ with the inclusion $r \in \mathcal{R}$: in the LP case, the function $\mathrm{Opt}(r)$ is either identically $+\infty$ everywhere on $\mathcal{R}$, or is real-valued at every point of $\mathcal{R}$.

Solution: Let $\mathrm{Opt}(r)$ be finite at a point $\overline{r} \in \mathrm{int}\,\mathcal{R}$ and let $r \in \mathcal{R}$; we need to prove that $\mathrm{Opt}(r) < \infty$. There is nothing to prove when $r = \overline{r}$, thus assume that $r \ne \overline{r}$. For properly selected $r_- \in \mathcal{R}$, the point $\overline{r}$ is a relative interior point of the segment $[r_-, r]$, which combines with the concavity inequality from the previous item to imply that both $\mathrm{Opt}(r_-)$ and $\mathrm{Opt}(r)$ are finite.
Now let $X = \mathbf{R}^n$, $f(x) = f^\top x$ and $G(x) = Ax - b$. Assuming that $\overline{r} \in \mathcal{R}$ is such that $\mathrm{Opt}(\overline{r}) < \infty$, we conclude that when $r = \overline{r}$, the LP program

$$\max_x \{f^\top x : Ax \le b + r\} \qquad (L[r])$$

is feasible and bounded. By LP duality, it means that the dual problem

$$\min_y \left\{[b + \overline{r}]^\top y : y \ge 0, A^\top y = f\right\}$$

is solvable and therefore feasible. But the feasible set of the LP dual to $(L[r])$ is independent of r, implying by Weak Duality that problems $(L[r])$ are above bounded. Thus, in the situation in question $\mathrm{Opt}(r)$ does not take value $+\infty$ at all and therefore is real-valued on $\mathcal{R}$. ■

Comment. Think about problem $(P[r])$ as about problem where r is the vector of resources you create, and $f(\cdot)$ is your profit, so that the problem is to maximize your profit given your resources and "technological constraints" $x \in X$. Now let $\overline{r} \in \mathcal{R}$ and e be a nonnegative vector, and let us look what happens when you select your vector of resources on the ray $R = \overline{r} + \mathbf{R}_+ e$, assuming that $\mathrm{Opt}(r)$ on this ray is real-valued. Restricted on this ray, your best profit becomes a function $\phi(t)$ of nonnegative variable t:

$$\phi(t) = \mathrm{Opt}(\overline{r} + te).$$

Since $e \geq 0$, this function is nondecreasing, as it should be: the larger t, the more resources you utilize, and the larger is your profit. A not so nice news is that $\phi(t)$ is concave in t, meaning that the slope of this function does not increase as t grows. In other words, if it costs you \$1 to pass from resources $\bar{x} + te$ to resources $\bar{x} + (t+1)e$, the return $\phi(t+1) - \phi(t)$ on one extra dollar of your investment goes down (or at least does not go up) as t grows. This is called *The Law of Diminishing Marginal Returns*.

Exercise III.5. [follow-up to Exercise III.4] There are n goods j with per-unit prices $c_j > 0$, per-unit utilities $v_j > 0$, and the maximum available amounts $\overline{x}_j$, $j \leq n$. Given budget $R \geq 0$, you want to decide on amounts x_j of goods to be purchased to maximize the total utility of the purchased goods, while respecting the budget and the availability constraints. Pose the problem as LO program and verify that the optimal value $\mathrm{Opt}(R)$ is piecewise linear function of R. What are the breakpoints of this function? What are the slopes between breakpoints?

Solution: Denoting by x_j the amount of good j we buy, maximizing the total utility becomes the LO program

$$\max_x \left\{ \sum_j v_j x_j : 0 \leq x_j \leq \overline{x}_j \, \forall j, \sum_j c_j x_j \leq R \right\}$$

As is immediately seen (check it!), the optimal solution to the problem is given by the following procedure: we sort the goods to make the ratios v_j/c_j ("utility per \$1 investment") nonincreasing. Assuming this is the case from the very beginning, we start with buying good # 1 until either the available amount of this good, or our budget becomes exhausted, whichever happens first. If this step does not exhaust the budget, we start to buy the second product until either its available amount, or the budget, is exhausted. Then, if we still have money, we start buying product # 3, and proceed in this fashion until either all available goods are bought, or the budget becomes zero. With this strategy, the breakpoints $R_1 < R_2 < ... < R_n$ of $\mathrm{Opt}(R) : [0, \infty) \to \mathbf{R}$ are given by the recurrence

$$R_k = R_{k-1} + \min[R - R_{k-1}, \overline{x}_k / c_k], \; 1 \leq k \leq n,$$

where $R_0 = 0$, and the slope of $\mathrm{Opt}(\cdot)$ on (R_{k-1}, R_k) is v_k/c_k; to the right of R_n, the slope is zero.

Exercise III.6. Let $\beta \in \mathbf{R}^n$ be such that $\beta_1 \geq \beta_2 \geq ... \geq \beta_n$. For $x \in \mathbf{R}^n$, let $x_{(k)}$ be the k-th largest entry in x. Consider the function

$$f(x) = \sum_k \beta_k x_{(k)} = [\beta_1 - \beta_2] s_1(x) + [\beta_2 - \beta_3] s_2(x) + ... + [\beta_{n-1} - \beta_n] s_{n-1}(x) + \beta_n s_n(x),$$

where, as always, $s_k(x) = \sum_{i=1}^k x_{(i)}$. As we know from Exercise I.29, the functions $s_k(x)$, $k < n$, are polyhedrally representable:

$$t \geq s_k(x) \iff \exists z \geq 0, s : x_i \leq z_i + s, i \leq n, \sum_i z_i + ks \leq t,$$

and $s_n(x)$ is just linear:

$$s_n(x) = \sum_i x_i$$

As a result, f admits the polyhedral representation

$$t \geq f(x) \iff \exists Z = [z_{ik}] \in \mathbf{R}^{n \times (n-1)}, s_k, t_k, k < n : \begin{cases} \forall (i \leq n, k < n) : z_{ik} \geq 0, x_i \leq z_{ik} + s_k, \\ \forall k < n : t_k \geq \sum_i z_{ik} + k s_k \\ t \geq \sum_{k=1}^{n-1} [\beta_k - \beta_{k+1}] t_k + \beta_n \sum_{i=1}^n x_i \end{cases}$$

This polyhedral representation has $2n^2 - n$ linear inequalities and $n^2 + n - 2$ extra variables. Now goes the exercise:

1. Find an alternative polyhedral representation of f with $n^2 + 1$ linear inequalities and $2n$ extra variables.

Solution: Let Π_n be the set of $n \times n$ doubly stochastic matrices. By Birkhoff Theorem, Π_n is the convex hull of $n \times n$ permutation matrices, implying that the set $X = \{Px : P \in \Pi_n\}$ is the convex hull of vectors obtained from x by permuting entries, which combines with $\beta_1 \geq ,,, \geq \beta_n$ to imply that

$$f(x) = \max_{P \in \Pi_n} \beta^\top Px,$$

that is, denoting by $\mathbf{e}$ the n-dimensional all-ones vector,

$$\begin{array}{rcl} f(x) & = & \max\limits_{P=[P_{ij}]} \left\{\beta^\top Px : P_{ij} \geq 0, P\mathbf{e} = \mathbf{e}, P^\top \mathbf{e} = \mathbf{e}\right\} \\ & = & \min\limits_{\lambda,\mu,[y_{ij}]} \left\{\mathbf{e}^\top[\lambda+\mu] : y_{ij} \geq 0, [\mathbf{e}\lambda^\top + \mu\mathbf{e}^\top]_{ij} - y_{ij} = [\beta x^\top]_{ij}, 1 \leq i,j \leq n\right\} \\ & & \text{[LP Duality]} \\ & = & \min\limits_{\lambda,\mu} \left[\mathbf{e}^\top[\lambda+\mu] : [\mathbf{e}\lambda^\top + \mu\mathbf{e}^\top]_{ij} \geq [\beta x^\top]_{ij},\ 1 \leq i,j \leq n\right\} \end{array}$$

Thus, $f(x)$ admits the polyhedral representation

$$t \geq f(x) \iff \exists \lambda, \mu \in \mathbf{R}^n : t \geq \mathbf{e}^\top[\lambda+\mu], [\mathbf{e}\lambda^\top + \mu\mathbf{e}^\top]_{ij} \geq [\beta x^\top]_{ij},\ 1 \leq i,j \leq n$$

and this representation has n^2+1 linear inequalities and $2n$ extra variables.

2. [computational study] Generate at random orthogonal $n \times n$ matrix U and vector β with nonincreasing entries and solve numerically the problem

$$\min_x \left\{ f(x) := \sum_k \beta_k x_{(k)} : \|Ux\|_\infty \leq 1 \right\}$$

utilizing the above polyhedral representations of f. For $n = 8, 16, 32, ..., 1024$, compare the running times corresponding to the 2 representations in question.

Solution: In our `CVX` experiments, U and β were generated according to

```
[U,D,V]=svd(randn(n,n));beta=-sort(randn(n,1))
```

and the ratio of the CPU time for the "long" polyhedral representation of f in use to the CPU time for the "short" one was as follows:

n	8	16	32	64	128	256	512	1024
CPU ratio	1.67	1.35	1.71	1.92	4.26	6.39	8.19	10.29

Exercise III.7. Let $a \in \mathbf{R}^n$ be a nonzero vector, and let $f(\rho) = \ln(\|a\|_{1/\rho})$, $\rho \in [0,1]$. Moment inequality, see section 13.3.3, states that f is convex. Prove that the function is also nonincreasing and Lipschitz continuous, with Lipschitz constant $\ln n$, or, which is the same, that

$$1 \leq p \leq p' \leq \infty \implies \|a\|_p \geq \|a\|_{p'} \geq n^{\frac{1}{p'} - \frac{1}{p}} \|a\|_p.$$

Solution: By homogeneity, it suffices to prove the inequality assuming $\|a\|_p = 1$, and by continuity it suffices to consider the case when $p \leq p' < \infty$. Setting $\alpha_i = |a_i|^p$, we get $1 = \|a\|_p = \left[\sum_i \alpha_i\right]^{1/p}$, $\|a\|_{p'} = \left[\sum_i \alpha_i^{\frac{p'}{p}}\right]^{1/p'}$. In terms of α_i the goal is to prove that when $\alpha_i \geq 0$ sum up to 1, then

$$1 \geq \sum_i \alpha_i^{\frac{p'}{p}} \geq \left[n^{\frac{1}{p'} - \frac{1}{p}}\right]^{p'},$$

which is immediate: due to $p' \geq p$ the function $g(\alpha) = \sum_i \alpha_i^{p'/p}$ is convex on the probabilistic simplex $\{\alpha \in \mathbf{R}^n_+ : \sum_i \alpha_i = 1\}$ and therefore attains its maximum on this simplex at a vertex (Theorem III.11.7), and attains its minimum on the simplex at the barycenter $n^{-1}[1; ...; 1]$ by Symmetry Principle (Proposition III.11.5 – permutations of coordinates are symmetries of the simplex and of $g(\cdot)$). ■

Exercise III.8. This Exercise demonstrates power of Symmetry Principle. Consider the situation as follows: you are given noisy observations

$$\omega = Ax + \xi,\ A = \mathrm{Diag}\{\alpha_i, i \le n\}$$

of unknown signal x known to belong to the unit ball $\mathbf{B} = \{x \in \mathbf{R}^n : \|x\|_2 \le 1\}$; here $\alpha_i > 0$ are given, and ξ is the standard (zero mean, unit covariance) Gaussian observation noise. Your goal is to recover from this observation the vector $y = Bx$, $B = \mathrm{Diag}\{\beta_i, i \le n\}$ being given. You intend to recover y by *linear estimate*

$$\widehat{y}_H(\omega) = H\omega,$$

where H is an $n \times n$ matrix you are allowed to choose. For example, selecting $H = BA^{-1} = \mathrm{Diag}\{\beta_i\alpha_i^{-1}\}$, you get an *unbiased* estimate:

$$\mathbf{E}\{\widehat{y}_H(Ax + \xi) - y\} = 0.$$

Let us quantify the quality of a candidate linear estimate $\widehat{y}_H$
— at a particular signal $x \in \mathbf{B}$ - by the quantity

$$\mathrm{Err}_x(H) = \sqrt{\mathbf{E}\{\|\widehat{y}_H(Ax + \xi) - Bx\|_2^2\}},$$

so that $\mathrm{Err}_x^2(H)$ is the expected squared $\|\cdot\|_2$-distance between the estimate and the estimated quantity,
— on the entire set $\mathbf{B}$ of possible signals – by *risk* $\mathrm{Risk}[H] = \max_{x\in\mathbf{B}} \mathrm{Err}_x(H)$.

1. Find closed form expressions for $\mathrm{Err}_x(H)$ and $\mathrm{Risk}(H)$.
2. Formulate the problem of finding the linear estimate with minimal risk as the problem of minimizing a convex function and prove that the problem is solvable, and admits an optimal solution H^* which is diagonal: $H^* = \mathrm{Diag}\{\eta_i, i \le n\}$.
3. Reduce the problem yielded by item 2 to the problem of minimizing easy-to-compute convex univariate function. Consider the case when $\beta_i = i^{-1}$ and $\alpha_i = [\sigma i^2]^{-1}$, $1 \le i \le n$, set $n = 10000$ and fill the following table:

σ	1.0	0.1	0.01	0.001	0.0001	0.00001	0.000001
$\mathrm{Risk}[H^*]$							
$\mathrm{Risk}[BA^{-1}]$							

where H^* is the minimum risk linear estimate as yielded by the solution to univariate problem you end up with, and $\mathrm{Risk}[BA^{-1}]$ is the risk of unbiased linear estimate.
You should see from your numerical results that minimal risk of linear estimation is much smaller than the risk of the unbiased linear estimate. Explain on qualitative level why allowing for bias reduces the risk.

Solution: 1: We have

$$\begin{array}{rcl}
\mathrm{Err}_x^2[H] & = & \mathbf{E}\{\|H[Ax+\xi] - Bx\|_2^2\} = \mathbf{E}\{\|[HA-B]x + H\xi\|_2^2\} \\
& = & \mathbf{E}\{\|[B-HA]x\|_2^2 + 2[H\xi]^\top[HA-B]x + [H\xi]^\top[H\xi]\} \\
& = & \|[B-HA]x\|_2^2 + \mathbf{E}\{\xi^\top H^\top H\xi\} \\
& = & \|[B-HA]x\|_2^2 + \mathbf{E}\{\mathrm{Tr}(\xi^\top H^\top H\xi)\} = \|[B-HA]x\|_2^2 + \mathbf{E}\{\mathrm{Tr}(H^\top H\xi\xi^\top)\} \\
& = & \|[B-HA]x\|_2^2 + \mathrm{Tr}(H^\top H\mathbf{E}\{\xi\xi^\top)\} = \|[B-HA]x\|_2^2 + \mathrm{Tr}(H^\top H)
\end{array}$$

and

$$\mathrm{Risk}^2[H] = \max_{x\in\mathbf{B}} \mathrm{Err}_x^2(H) = \mathrm{Tr}(H^\top H) + \max_{x,\|x\|_2\le 1} \|[B-HA]x\|_2^2 = \mathrm{Tr}(H^\top H) + \|[B-HA]\|^2,$$

where $\|\cdot\|$ is the spectral norm, see section D.1.4.
2: By the solution to item 1, the minimum risk linear estimate is yielded by an optimal solution to the problem

$$\mathrm{Opt} = \min_{H\in\mathbf{R}^{n\times n}} \left[R(H) = \|[B-AH]\|^2 + \mathrm{Tr}(H^\top H)\right]. \qquad (!)$$

the best achievable risk being $\sqrt{\text{Opt}}$. The objective tends to ∞ when $\|H\| \to \infty$, implying the existence of solution.

To prove that there exists an optimal solution H^* which is diagonal, let us apply Symmetry Principle (Proposition III.11.5). Consider the set $\mathcal{G}$ of all linear transformations $X \mapsto \overline{G}(X) := GXG : \mathbf{R}^{n\times n} \to \mathbf{R}^{n\times n}$ associated with 2^n diagonal $n \times n$ matrices G with diagonal entries ± 1. This clearly is a group, and its elements are symmetries of the feasible set $\mathbf{R}^{n\times n}$ of our optimization problem. Let us prove that every transformation $X \to \overline{G}(X)$, $\overline{G} \in \mathcal{G}$, is a symmetry of the objective as well. To this end note that multiplying a matrix from the left and/or from the right by orthonormal matrices, we clearly preserve the spectral norm of the matrix. Therefore for $\overline{G}(\cdot) \in \mathcal{G}$ we have

$$\begin{array}{rcl} R(\overline{G}(H)) & = & \|B - GHGA\|^2 + \text{Tr}([GHG]^\top [GHG]) \\ & = & \|GBG - GHG[GAG]\|^2 + \text{Tr}(GH^\top G^2 HG) \\ & & \quad [\text{due to } B = GBG \text{ and } A = GAG - A \text{ and } B \text{ are diagonal, and } G = G^\top] \\ & = & \|G[B - HA]G\|^2 + \text{Tr}(GH^\top HG) \ [\text{due to } G^2 = I_n] \\ & = & \|B - HA\|^2 + \text{Tr}(HH^\top) = R(H) \\ & & \quad [\text{we have used orthonormality of } G \text{ and the relation } \text{Tr}(GH^\top HG) \\ & & \quad = \text{Tr}(H^\top HG^2) = \text{Tr}(H^\top H) \text{ due to } G^2 = I_n.] \end{array}$$

By Symmetry Principle, the (solvable) convex problem in question has an optimal solution H^* which is $\mathcal{G}$-symmetric: $GH^*G = H^*$ for all diagonal G with diagonal entries ± 1. Observing that $[GH^*G]_{ij} = G_{ii}H^*_{ij}G_{jj}$, we get that $G_{ii}H^*_{ij}G_{jj} = H^*_{ij}$ for all i, j whenever $G_{kk} = \pm 1$ for all k, implying that $H^*_{ij} = -H^*_{ij}$ when $i \neq j$, that is, $H^*_{ij} = 0$ when $i \neq j$, so that H^* is diagonal. ■

3: By item 2, we do not spoil the optimal value in (!) when restricting ourselves with diagonal candidate solutions $H = \text{Diag}\{\eta_i\}$, thus arriving at the problem

$$\begin{array}{rcll} \text{Opt} & = & \min_{\eta_i, i\leq n} \left[\max_i [\beta_i - \eta_i\alpha_i]^2 + \sum_i \eta_i^2\right] & \\ & = & \min_{\rho, \{\eta_i\}} \left[\rho^2 + \sum_i \eta_i^2 : |\beta_i - \alpha_i\eta| \leq \rho\right] & \\ & = & \min_{\rho\geq 0:} \left\{\rho^2 + \sum_i \left[\frac{\max[\rho - |\beta_i|, 0]}{\alpha_i}\right]^2\right\} & (*) \end{array}$$

In the case when $\beta_i = i^{-1}$ and $\alpha_i = [\sigma i^2]^{-1}$, $1 \leq i \leq n$, and $n = 10000$ one has

σ	1.0	0.1	0.01	0.001	0.0001	0.00001	0.000001
Risk$[H^*]$	0.7071	0.28244	0.1124	0.04474	0.01781	0.00709	0.00282
Risk$[BA^{-1}]$	1.827e4	1.827e3	1.827e2	1.827e1	1.827e0	0.1827	0.01827

where H^* is the minimum risk linear estimate as yielded by the solution to $(*)$, and Risk$[BA^{-1}]$ is the risk of unbiased linear estimate.

Unbiased recovery in the case of diagonal B and A recovers an entry y_i in y as

$$\widehat{y}_i = \frac{\beta_i}{\alpha_i}\omega_i = y_i + \frac{\beta_i}{\alpha_i}\xi_i.$$

We see that while the recovery is unbiased, it significantly amplifies the noise, provided that β_i/α_i is large (with our data, this ratio is σi and indeed is large when $i \gg 1/\sigma$). On the other hand, we know in advance that x is bounded by 1 in $\|\cdot\|_2$, so that when β_i is small for some i, the bias in recovering y_i will be small even when we recover y_i by 0. The optimal linear estimate heavily utilizes our a priori information $\|x\|_2 \leq 1$ to find optimal tradeoff between the bias and the stochastic component of the recovery error $y_i - \widehat{y}_i$, this is why it not just beats the unbiased linear estimate (this always is the case – the latter estimate is linear!), but may beat it by huge margin, For example, the above table shows that unbiased estimate makes no sense when $\sigma \geq$ 1.e–4 – knowing in advance that $\|x\|_2 \leq 1$, we can estimate x by 0 with risk 1 (which does not require observations at all), which is better than the risk 1.82... of the unbiased linear estimate when $\sigma =$ 1.e–4 .

Exercise III.9. Given the sets of d-dimensional tentative nodes ($d = 2$ or $d = 3$) and of tentative bars of a TTD problem satisfying assumption $\mathfrak{R}$, let $\mathcal{V} = \mathbf{R}^M$ be the space of virtual displacements of the nodes, N be the number of tentative bars, and $W > 0$ be the allowed total bar volume, see Exercise I.16. Let, next, $\mathcal{C}(t, f) : \mathbf{R}^N_+ \times \mathcal{V} \to \mathbf{R} \cup \{+\infty\}$ be the compliance of truss $t \geq 0$ w.r.t. load f (we identify trusses with the corresponding vectors t of bar volumes). Prove that

1. $\mathcal{C}(t,f)$ is a convex lsc function, positively homogeneous of homogeneity degree 1, of $[t;f]$ with $\mathbf{R}^N_{++}\times\mathcal{V}\subset\mathrm{Dom}\,\mathcal{C}$, where $\mathbf{R}^N_{++}=\mathrm{int}\,\mathbf{R}^N_+=\{t\in\mathbf{R}^N: t>0\}$. This function is positively homogeneous, with degree -1, in t, when f is fixed, and positively homogeneous, of degree 2, in f when t is fixed. Besides this, $\mathcal{C}(t,f)$ is nonincreasing in $t\geq 0$: if $0\leq t'\leq t$, then $\mathcal{C}(t,f)\leq\mathcal{C}(t',f)$ for every f.

   Solution: As we know from Exercise I.16.2, the epigraph of $\mathcal{C}$ is

$$\mathrm{epi}\{\mathcal{C}\}:=\{[t;f;\tau]:\tau\geq\mathcal{C}(t,f)\}=\{[t;f;\tau]:t\geq 0,\mathcal{A}(t,f,\tau):=\left[\begin{array}{c|c}B\,\mathrm{Diag}(t)B^\top & f\\ \hline f^\top & 2\tau\end{array}\right]\succeq 0\}\quad(!)$$

   with given by the data $M\times N$ matrix B satisfying $BB^\top\succ 0$ (the latter is our default assumption $\mathfrak{R}$). We see that $\mathrm{epi}\{\mathcal{C}\}$ is closed convex set, implying that $\mathcal{C}$ is a convex lsc function (Proposition III.12.2). The inclusion $\mathbf{R}^N_{++}\times\mathcal{V}\subset\mathrm{Dom}\,\mathcal{C}$ is readily given by positive definiteness of the matrix $A(t)=B\,\mathrm{Diag}\{t\}B^\top$ for positive t's; whenever $A(t)\succ 0$, the matrix $\mathcal{A}(t,f,\tau)$ is, for every f, positive semidefinite whenever τ is large enough by the Schur Complement Lemma. Positive homogeneity, of degree 1, of $\mathcal{C}$ clearly follows from the fact that by the above description, $\mathrm{epi}\{\mathcal{C}\}$ is a closed cone. By (!) combined with the Schur Complement Lemma, whenever $[t;f;\tau]\in\mathrm{epi}\{\mathcal{C}\}$ and $\lambda>0,\mu$ are reals, we have $[\lambda t;\mu f;\lambda^{-1}\mu^2\tau]\in\mathrm{epi}\{\mathcal{C}\}$, implying the claims about homogeneity of $\mathcal{C}$ w.r.t. t and w.r.t. f. Finally, by the same (!) when $[t';f,\tau]\in\mathrm{epi}\{\mathcal{C}\}$ and $t\geq t'$, we have $[t;f;\tau]\in\mathrm{epi}\{\mathcal{C}\}$ as well, implying that $\mathcal{C}(t,\tau)$ is nonincreasing in $t\geq 0$. ■

2. The function $\mathrm{Opt}(W,f)=\inf_t\left\{\mathcal{C}(t,f):t\geq 0,\sum_i t_i=W\right\}$ – the optimal value in the TTD problem (5.2) – with W restricted to reside in $\mathbf{R}_{++}=\{W>0\}$ is convex continuous function with the domain $\mathbf{R}_{++}\times\mathcal{V}$. This function is positively homogeneous, of degree -1, in $W>0$ and homogeneous, of homogeneity degree 2, in f:

$$\forall(\lambda>0,\mu):\quad\mathrm{Opt}(\lambda W,\mu f)=\lambda^{-1}\mu^2\mathrm{Opt}(W,f),\ \forall(W,f)\in\mathbf{R}_{++}\times\mathcal{V}.$$

   Moreover, the infimum in $\inf_t\left\{\mathcal{C}(t,f):t\geq 0,\sum_i t_i=W\right\}$ is achieved whenever $W>0$.

   Solution: Consider the set

$$\mathcal{G}=\{[t;f;\tau;W]:W=\sum_i t_i,t\geq 0,\mathcal{A}(t,f,\tau)\succeq 0\}$$

   This set clearly is a closed convex cone, and the function

$$F(t,f,\tau,W)=\left\{\begin{array}{cc}\tau, & [t;f;\tau;W]\in\mathcal{G}\ \&\ W>0\\ +\infty, & \text{otherwize}\end{array}\right.$$

   is convex and nonnegative on this set. We clearly have

$$\mathrm{Opt}(W,f)=\inf_{t,\tau}F(t,f,\tau,W),\qquad(!!)$$

   which combines with convexity and nonnegativity of F and the rule on partial minimization (stability of convexity w.r.t. partial minimization, section 10.1) to imply that $\mathrm{Opt}(W,f)$ is convex (and of course nonnegative) function of (W,f). Moreover, by Exercise I.16.3, for every $W>0$

$$\inf_t\{\mathcal{C}(t,f):t\geq 0,\sum_i t_i=W\}=\inf_{t:t\geq 0,\sum_i t_i=W}\min_\tau\{\tau:\mathcal{A}(t,f,\tau)\succeq 0\}$$

   is achieved, implying, first, the "Moreover" claim of the statement we are justifying, and, second, the fact that $\mathrm{Opt}(W,f)$ is finite whenever $W>0$. Thus, $\mathrm{Opt}(W,f)$ is convex real-valued function in the domain $\mathbf{R}_{++}\times\mathcal{V}$, and since this domain is convex and open, Opt is continuous in this domain, as claimed. Homogeneity properties of this function we have announced are immediate consequences of the fact that $\mathcal{G}$ is a closed cone and that by Schur Complement Lemma for $\lambda>0,\mu\neq 0$ the matrices $\mathcal{A}(t,f,\tau)$ and $\mathcal{A}(\lambda t,\mu f,\lambda^{-1}\mu^2\tau)$ simultaneously are/are not positive semidefinite. ■

3. When on certain bridge there is just one car, of unit weight, the compliance of the bridge does not exceed 1, whatever be the position of the car. How large could the compliance of the bridge when there are 100 cars of total weight 70 on it?

Solution: The compliance is at most 4900. Indeed, a 100-force load with the total of forces' magnitudes 1 is a convex combination of loads with single force of magnitude 1 [8]. As we know from item 1, the compliance of a given truss is a convex function of the load, implying by Jensen's inequality that when our bridge is loaded by 100 (or any other number of) cars of total weight 1, the compliance does not exceed the maximum of compliances caused by a single car of unit weight, the maximum being taken over possible positions of this single car. For our bridge, this maximum is ≤ 1, implying that the compliance of the bridge loaded by a whatever number of cars of total weight 1 does not exceed 1. It remains to note that the compliance is positively homogeneous, of degree 2, function of load, so that with the total weight of cars not exceeding 70, the compliance does not exceed $70^2 = 4900$.

To formulate the next two tasks, let us associate with a free node p the set $\mathcal{F}^p$ of all single-force loads stemming from forces g of magnitude $\|g\|_2$ not exceeding 1 and acting at node p. For a set S of free nodes, $\mathcal{F}^S$ is the set of all loads with nonzero forces acting solely at the nodes from S and with the sum of $\|\cdot\|_2$-magnitudes of the forces not exceeding 1, so that

$$\mathcal{F}^S = \mathrm{Conv}(\cup_{p\in S}\mathcal{F}^p)$$

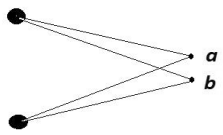
(why?)

4. Let $S = \{p_1, ..., p_K\}$ be a K-element collection of free nodes from the nodal set. Assume that for every node p from S and every load $f \in \mathcal{F}^p$ there exists a truss of a given total weight W such that its compliance w.r.t. f does not exceed 1. Which, if any, of the following statements are true?
   (i) For every load $f \in \mathcal{F}^S$, there exists a truss of total volume W with compliance w.r.t. f not exceeding 1
   (ii) There exists a truss of total volume W with compliance w.r.t. every load from $\mathcal{F}^S$ not exceeding 1
   (iii) For properly selected γ depending solely on d, there exists a truss of total volume γKW with compliance w.r.t. every load from $\mathcal{F}^S$ not exceeding 1

   Solution: The first and the third claims are correct, the second, in general, is wrong. Indeed, let $\mathcal{C}(t, f)$ be the compliance of truss t w.r.t. load f; as we know from item 1, this is a convex function of (t, f).

   To justify the first claim, given load $f \in \mathcal{F}^S$, we can find its representation $f = \sum_k \lambda_k f^k$ as a convex combination of loads $f^k \in \mathcal{F}^{p_k}$. By assumption on S, for every k there exists truss t^k of total volume W such that $\mathcal{C}(t^k, f^k) \leq 1$, implying by convexity that $\mathcal{C}(t := \sum_k \lambda_k t^k, f) \leq 1$. Since the total volume of t is W, t is the truss announced in claim 1.

   To demonstrate that the second claim is wrong in general, consider planar sets of tentative nodes and bars depicted on Figure S2III.1,



Fugure S2III.1. Tentative nodes and bars.

   where bold circles are fixed nodes, and $S = \{a, b\}$. Denoting by $\mathcal{T} = \{t \in \mathbf{R}^4_+ : \sum_i t_i = W\}$ the set of all trusses of total volume W, we can assume w.l.o.g. that

$$\max_{f\in\mathcal{F}^a} \min_{t\in\mathcal{T}} \mathcal{C}(t, f) = 1,$$

[8] "What is meant is not always put into writing" ("Boris Godunov" by Alexander Pushkin): we tacitly assume that possible locations of cars are among the nodes of the bridge, modeled as a truss, and that the stemming from cars forces acting at the bridge "look down."

note that by symmetry we also have

$$\max_{f\in\mathcal{F}^b}\min_{t\in\mathcal{T}}\mathcal{C}(t,f)=1,$$

so that we are in the situation postulated in item 4. Let $f^a\in\mathcal{F}^1$ and $f^b\in\mathcal{F}^2$ be the "critical loads" – those where the respective $\max_f$ are achieved. Clearly, there is no truss $\bar{t}$ of total volume W with $\mathcal{C}(\bar{t},f^a)\le 1$ and $\mathcal{C}(\bar{t},f^b)\le 1$ – were it existing, with our TTD data there clearly would exist truss t of total volume $W/2$ with compliance w.r.t one of the loads, f^a or f^b, not exceeding 1. It would imply the existence of truss of total volume W with compliance w.r.t. either f^a, or f^b not exceeding $1/2$ (recall that $\mathcal{C}(t,f)$ is homogeneous, of degree -1, with respect to t), which is not the case, since both loads are critical.

To justify the third claim, note that for every integer $\mu\ge d+1$ the unit $\|\cdot\|_2$-ball B in $\mathbf{R}^d$ is contained in the convex hull $\Delta=\mathrm{Conv}(\{r_{d,\mu}g^\iota,1\le\iota\le d+1\})$ of μ vectors of the $\|\cdot\|_2$-norm $r_{d,\mu}$ each, g^ι being unit normalizations of these vectors. For example, when $d=2$, specifying Δ as μ-side perfect polygon circumscribed around the unit circle and the vectors g^ι as the unit length normalization of the vertices of the polygon, we get $r_{d,\mu}=1/\cos(\pi/\mu)$.

Let us specify $f^{k\iota},1\le k\le K,1\le\iota\le\mu$ as single-force load with force g^ι acting at k-th node, p_k, of S, $1\le k\le K,1\le\iota\le\mu$, so that $f^{k\iota}\in\mathcal{F}^{p_k}$. Under the premise of item 4, there exist trusses $t^{k\iota}$ of total volume W each such that $\mathcal{C}(t^{k,\iota},f^{k,\iota})\le 1$. Let $\underline{t}=\sum_{k,\iota}t^{k\iota}$. Since $\mathcal{C}(t,f)$ clearly is nondecreasing in $t\ge 0$, we have $\mathcal{C}(\underline{t},f^{k\iota})\le 1$ for all k,ι, whence, by convexity of $\mathcal{C}(t,f)$ in f, $\mathcal{C}(\underline{t},f)\le 1$ for all $f\in U:=\mathrm{Conv}(\{f^{k\iota},k\le K,\iota\le\mu\})$. Due to the origin of g^ι, we have $U\supset r_{d,\mu}^{-1}\mathcal{F}^{p_k}\ \forall k\le K$, implying that $U\supset\mathrm{Conv}(\cup_k r_{d,\mu}^{-1}\mathcal{F}^{p_k})=r_{d,\mu}^{-1}\mathcal{F}^S$. Thus, $\mathcal{C}(\underline{t},f)\le 1$ for all $f\in r_{d,\mu}^{-1}\mathcal{F}^S$. By homogeneity of $\mathcal{C}(t,f)$ w.r.t. f and to t, it follows that $\mathcal{C}(r_{d,\mu}^2\underline{t},f)\le 1$ for all $f\in\mathcal{F}^S$, so that the compliance of the truss $\bar{t}=r_{d,\mu}^2\underline{t}$ w.r.t. every load from $\mathcal{F}^S$ does not exceed 1. It remains to note that the total volume of $\bar{t}$ is $\underbrace{r_{d,\mu}^2\mu}_{\gamma}KW$. We can now try different values of μ in order to minimize the factor γ over μ (and over geometry of Δ). For $d=2$, restricting ourselves with perfect μ-side polygons Δ circumscribed around the unit circle, the best μ is 5, resulting in $\gamma=5/\cos^2(\pi/5)\approx 7.6393$. When $d=3$, we restricted our search with Platonian solids Δ circumscribed around the unit 3D ball. The best solid was the octahedron, resulting in $\mu=6$ and $\gamma=18$. ■

★5. Prove the following statement:

> *In the situation of item 4 above, let $\gamma=4$ when $d=2$ and $\gamma=7$ when $d=3$. For every $k\le K$ there exists a truss $\widehat{t}^k$ of total volume γW such that the compliance of $\widehat{t}^k$ w.r.t. every load from $\mathcal{F}^{p_k}$ does not exceed 1. As a result, there exists truss $\widetilde{t}$ of total volume γKW with compliance w.r.t. every load from $\mathcal{F}^S$ not exceeding 1.*

Solution: Given $\epsilon\in(0,1)$, let $\mathcal{T}_\epsilon=\{t\in\mathbf{R}^N_+:\sum_i t_i=W,t_i\ge\epsilon W/N\ \forall i\}$.

1° Observe that for every $f\in\mathcal{F}^{p_k}$ there exists truss $t\in\mathcal{T}_\epsilon$ such that $\mathcal{C}(t,f)\le(1-\epsilon)^{-1}$. Indeed, given $f\in\mathcal{F}^{p_k}$, there exists truss t^f of total volume W such that $\mathcal{C}(t^f,f)\le 1$. Setting $\underline{t}=\epsilon[W/N;W/N;...W/N]$ and $\bar{t}=(1-\epsilon)t^f+\underline{t}$, we get a truss from $\mathcal{T}_\epsilon$ satisfying $\bar{t}\ge(1-\epsilon)t^f$. Since $\mathcal{C}(t,f)$ is nonincreasing and positively homogeneous, of degree -1, in $t>0$, we conclude that $\mathcal{C}(\bar{t},f)\le(1-\epsilon)^{-1}$, as claimed.

2° Let us fix a free node p of the nodal set, and let f_g, $g\in\mathbf{R}^d$, stand for single-force load where force g acts at node p. Recall that vectors from the space $\mathcal{V}=\mathbf{R}^M$ of nodal displacements are block vectors with d-dimensional blocks representing "physical displacements" of free nodes and indexed by these nodes. Let I_p be the set of indexes of those entries in a vector of nodal displacements which correspond to the block indexed by p.

When $t\in\mathcal{T}_\epsilon$, the stiffness matrix $A(t)=B\,\mathrm{Diag}\{t\}B^\top$ of truss t is positive definite (assumption $\mathfrak{R}$), so that the equilibrium displacement of truss t under a load f is $v=[A(t)]^{-1}f$, and the compliance is

$$\mathcal{C}(t,f)=\frac{1}{2}v^\top f=\frac{1}{2}f^\top A^{-1}(t)f.$$

As a result, for every $g \in \mathbf{R}^d$ and every truss $t \in \mathcal{T}_\epsilon$ one has

$$\mathcal{C}(t, f_g) = \frac{1}{2} g^\top [A^{-1}(t)]_{I_p} g,$$

where for $Q = [Q_{i,j}]_{i,j \le M} \in \mathbf{S}^M$, $[Q]_{I_p} = [Q_{ij}]_{i,j \in I_p} \in \mathbf{S}^d$ is the $d \times d$ principal submatrix of Q corresponding to rows and columns with indexes from I_p.
Next, let $\underline{A} = A(\underline{t})$, so that $0 \prec \underline{A} \preceq A(t)$ for every $t \in \mathcal{T}_\epsilon$ due to $t \ge \underline{t} > 0$. It follows that for all $t \in \mathcal{T}_\epsilon$ it holds $A^{-1}(t) \preceq \underline{A}^{-1}$, whence

$$\forall t \in \mathcal{T}_\epsilon : [A^{-1}(t)]_{I_p} \preceq \overline{Q} := [\underline{A}^{-1}]_{I_p}.$$

3^o Let us associate with node p the set $\mathcal{S}_p \subset \mathbf{S}^d_+$ given by

$$\mathcal{S}_p = \{Q \in \mathbf{S}^d : \exists t \in \mathcal{T}_\epsilon : [A^{-1}(t)]_{I_p} \preceq Q \preceq \overline{Q}\}.$$

We claim that $\mathcal{S}_p$ is a convex compact set in $\mathbf{S}^d_+$. The main component of verification is the following simple observation to be justified at the end of the proof:

(@) *Given symmetric positive definite $M \times M$ matrix $\underline{A}$ and a d-element subset I of its row indexes and denoting by $[C]_I$ the $d \times d$ principal submatrix of $C \in \mathbf{S}^M$ composed of rows and columns with indexes from I, the set*

$$\mathcal{S}^I = \{(Q, A) \in \mathbf{S}^d \times \mathbf{S}^M : A \succeq \underline{A} \ \&\ Q \succeq [A^{-1}]_I\}$$

is closed and convex.

By (@), the set

$$\mathcal{S}_p^+ = \left\{(Q, t) \in \mathbf{S}^d \times \mathcal{T}_\epsilon : Q \succeq [A^{-1}(t)]_{I_p}\right\}$$

is closed and convex (as the inverse image of closed and convex set $\mathcal{S}^{I_p}$ under the linear mapping $(Q, t) \mapsto (Q, A(t))$. Consequently, the projection $\mathcal{S}[p]$ of $\mathcal{S}_p^+$ onto the Q-space is convex, so that $\mathcal{S}_p$ is convex as well – this is the intersection of $\mathcal{S}[p]$ with the convex set $\{Q : Q \preceq \overline{Q}\}$. $\mathcal{S}_p$ clearly is bounded – it is contained in the set $\{0 \preceq Q \preceq \overline{Q}\}$. To see that $\mathcal{S}_p$ is closed, let $Q_i \in \mathcal{S}_p$ converge to Q as $i \to \infty$, and let us prove that $Q \in \mathcal{S}_p$, that is, that $Q \preceq \overline{Q}$ (which is evident) and that $Q \succeq [A^{-1}(t)]_{I_p}$ for some $t \in \mathcal{T}_\epsilon$. To see that the latter is the case, recall that for every i there exist $t^i \in \mathcal{T}_\epsilon$ such that $Q_i \succeq [A^{-1}(t^i)]_{I_p}$. Taking into account that $\mathcal{T}_\epsilon$ is compact and passing to a subsequence we can assume that $\lim_{i \to \infty} t^i$ exists; this limit clearly can be taken as the desired t. Thus, $\mathcal{S}_p$ is convex compact set.
4^o Now – the main step. Assume that p is a node from S and that we are under the premise of item 4, that is, for every $g \in \mathbf{R}^d$ with $\|g\|_2 \le 1$ there exists a truss t' of total volume W such that $\mathcal{C}(t', f_g) \le 1$. Invoking 1^o, we conclude that

$$\forall (g \in \mathbf{R}^d, \|g\|_2 \le 1) \exists t \in \mathcal{T}_\epsilon : \mathcal{C}(t, f_g) \le (1 - \epsilon)^{-1} \qquad (\#)$$

Denoting $B = \{g \in \mathbf{R}^d : \|g\|_2 \le 1\}$, consider the family of sets

$$\mathcal{Q}[g] = \{Q \in \mathcal{S}_p : \frac{1}{2} g^\top Q g \le \gamma (1 - \epsilon)^{-1}\}$$

parameterised by vectors $g \in B$. By their origin, the sets from the family are convex and compact. The crucial fact which we are about to prove is that

(!!) *Every γ of sets from the family have a point in common.*

To prove (!!), let $g_\ell \in B$, $1 \le \ell \le \gamma$, and let us prove that the sets $\mathcal{Q}[g_\ell]$, $\ell = 1, ..., \gamma$, have a point in common. For every ℓ, by (#), there exists $t^\ell \in \mathcal{T}_\epsilon$ such that $\mathcal{C}(t^\ell, f_{g_\ell}) \le (1 - \epsilon)^{-1}$, implying that setting

$$Q_\ell = [A^{-1}(t^\ell)]_{I_p},$$

we get

$$Q_\ell \preceq \overline{Q} \ \& \ \frac{1}{2} g_\ell^\top Q_\ell g_\ell = \frac{1}{2} f_{g_\ell}^\top A^{-1}(t^\ell) f_{g_\ell} \le (1-\epsilon)^{-1}$$

Now let $t = \frac{1}{\gamma}\sum_{\ell=1}^{\gamma} t^\ell$ and $Q = [A^{-1}(t)]_{I_p}$, so that $t \in \mathcal{T}_\epsilon$, $(Q,t) \in \mathcal{S}_p^+$, and $Q \preceq \overline{Q}$, implying that $Q \in \mathcal{S}_p$. For every ℓ we have

$$\begin{array}{c} t \ge \frac{1}{\gamma} t^\ell > 0 \implies A(t) \succeq \frac{1}{\gamma} A(t^\ell) \implies A^{-1}(t) \preceq \gamma A^{-1}(t^\ell) \\ \implies Q \preceq \gamma Q_\ell \implies \frac{1}{2} g_\ell^\top Q g_\ell \le \gamma g_\ell^\top Q_\ell g_\ell \le \gamma (1-\epsilon)^{-1} \end{array}$$

Thus, $Q \in \mathcal{S}_p$ and

$$\frac{1}{2} g_\ell^\top Q g_\ell \le \gamma (1-\epsilon)^{-1} \ \forall \ell \le \gamma$$

implying that $Q \in \mathcal{Q}[g_\ell]$ for all $\ell \le \gamma$, that is, $\cap_{\ell \le \gamma} \mathcal{Q}[g_\ell]$ is nonempty, as claimed.

5^o The rest of the proof is easy. Note that $\mathcal{Q}[g]$ are convex compact subsets of $\mathbf{S}^d$ and the latter linear space has dimension $\gamma - 1$. Applying Helly Theorem II, there exists a common point Q of all the sets $\mathcal{Q}[g]$, $g \in B$. Due to $\mathcal{Q}[g] \subset \mathcal{S}_p$ for every $g \in B$, we get $Q \in \mathcal{S}_p$, so that there exists $\bar{t} \in \mathcal{T}_\epsilon$ such that

$$Q \succeq [A^{-1}(\bar{t})]_{I_p}$$

Consequently,

$$\forall g \in B : \mathcal{C}(\bar{t}, f_g) = \frac{1}{2} f_g^\top A^{-1}(\bar{t}) f_g = \frac{1}{2} g^\top [A^{-1}(\bar{t})]_{I_p} g \le \frac{1}{2} g^\top Q g \le \gamma (1-\epsilon)^{-1},$$

where the concluding inequality is due to $Q \in \mathcal{Q}[g]$. Thus, there exists truss $\bar{t} \in \mathcal{T}_\epsilon$ such that the compliance of this truss w.r.t. every load f_g, $g \in B$, is $\le \gamma(1-\epsilon)^{-1}$.

6^o The remaining reasoning is quite straightforward. The truss $\bar{t}$ we have build depends on ϵ; setting $\epsilon_i = 1/(i+1)$, $i = 1, 2, ...$ let us denote by $\bar{t}^i$ the truss given by the above construction as applied with $\epsilon = \epsilon_i$. All these trusses are of total volume W, and passing to a subsequence, we can assume that $\bar{t}^i \to \bar{t}$ as $t \to \infty$. Due to $\mathcal{C}(\bar{t}^i, f_g) \le \gamma(1-\epsilon)^{-1}$, we have

$$\forall i \forall g \in B : \left[\begin{array}{c|c} B\,\mathrm{Diag}\{\bar{t}^i\}B^\top & f_g \\ \hline f_g^\top & 2\gamma(1-\epsilon)^{-1} \end{array}\right] \succeq 0,$$

implying that

$$\forall g \in B : \left[\begin{array}{c|c} B\,\mathrm{Diag}\{\bar{t}\}B^\top & f_g \\ \hline f_g^\top & 2\gamma \end{array}\right] \succeq 0$$

that is, $\mathcal{C}(\bar{t}, f_g) \le \gamma$ for all $g \in B$. Besides this, truss $\bar{t}$, same as all trusses $\bar{t}^i$, is of total volume W. Setting $\widehat{t} = \gamma t$, we get truss of total volume γW and compliance, w.r.t. every load f_g with $\|g\|_2 \le 1$, not exceeding 1.

Summing up the K trusses given by the above construction as applied to every one of the K free nodes composing the set S, we get a truss $\widetilde{t}$ of total volume $\gamma K W$ with compliance w.r.t. every load from $\mathcal{F}^S$ not exceeding 1.

Paying debts: proof of (@). Closedness of $\mathcal{S}^I$ is evident; all we need is to prove that the set is convex.

Let us make the following

> **Observation:** The mapping $X \mapsto X^{-1} : \mathrm{int}\,\mathbf{S}_+^M \to \mathrm{int}\,\mathbf{S}_+^M$ is $\succeq$-convex: whenever $X, Y \in \mathrm{int}\,\mathbf{S}_+^M$ and $\lambda \in [0,1]$, one has $[\lambda X + ((1-\lambda)Y]^{-1} \preceq \lambda X^{-1} + (1-\lambda)Y^{-1}$.

The "'bare hands" proof of this important fact (to be put into perspective in Part IV) is as follows. For $X, Y \succ 0$ we have, by Schur Complement Lemma, $\left[\begin{array}{c|c} X^{-1} & I \\ \hline I & X \end{array}\right] \succeq 0$, $\left[\begin{array}{c|c} Y^{-1} & I \\ \hline I & Y \end{array}\right] \succeq 0$, whence

$\left[\begin{array}{c|c} \lambda X^{-1}+(1-\lambda)Y^{-1} & I \\ \hline I & \lambda X+(1-\lambda)Y \end{array}\right] \succeq 0$, implying, by the same Schur Complement Lemma, that $\lambda X^{-1}+(1-\lambda)Y^{-1} \succeq [\lambda X+(1-\lambda)Y]^{-1}$.

Due to Observation, the mapping $X \mapsto [X^{-1}]_I : \operatorname{int}\mathbf{S}^M_+ \to \operatorname{int}\mathbf{S}^d_+$ is $\succeq$-convex:

$$\begin{array}{l} X \succ 0, Y \succ 0, \lambda \in [0.1] \implies [\lambda X+(1-\lambda)Y]^{-1} \preceq \lambda X^{-1}+(1-\lambda)Y^{-1} \\ \implies [(\lambda X+(1-\lambda)Y)^{-1}]_I \preceq [\lambda X^{-1}+(1-\lambda)Y^{-1}]_I = \lambda[X^{-1}]_I+(1-\lambda)[Y^{-1}]_I \\ \quad [\text{since principal submatrix of a positive semidefinite matrix is positive semidefinite as well}] \end{array}$$

which clearly implies the convexity of $\mathcal{S}^I = \{(Q,A) : A \succeq \underline{A}, Q \succeq [A^{-1}]_I\}$. ■

Some remarks are in order.

1. A careful reader hopefully recognizes the "driving force" behind the above proof – it is the same as the one used in essentially less technical, and thus much more transparent section 2.3.1.
2. In the above proof, it was completely unimportant that B was the unit ball of $\|\cdot\|_2$ – it could be a whatever nonempty subset of $\mathbf{R}^d$. In fact the proof justifies the following claim:

   Given (1) *the data of a TTD problem satisfying assumption* $\mathfrak{R}$ *and with nodes "living" in* $\mathbf{R}^d$ *(*$d=2$ *or* $d=3$*),* (2) *a collection of* K *free nodes* p_k, $k \le K$, *from the nodal set, and* (3) K *nonempty subsets* $B_k \subset \mathbf{R}^d$, *let* $\mathcal{F}^k$ *be the set of all single-force loads where a force from* B_k *acts at the node* p_k, *and let* $\mathcal{F} = \mathrm{Conv}\{\cup_{k\le K}\mathcal{F}^k\}$. *Assume that for every* k *and every load* $f \in \mathcal{F}^k$ *there exists a truss of total volume* W *with compliance w.r.t.* f *not exceeding 1. Then there exists truss of total volume* γKW *with compliance w.r.t. every load from* $\mathcal{F}$ *not exceeding 1, with* $\gamma = 4$ *when* $d=2$ *and* $\gamma = 7$ *when* $d=3$.
3. Finally we remark that the values of γ yielded by the above proof are essentially better than the values yielded by much more transparent (and fully adjusted to the unit Euclidean balls in the role of B_k's) solution to item 4. There is no reason to believe that these values are the smallest ones for which the above claim is true. This being said, it is easy to demonstrate that for $d=2$ the best possible in this respect value of γ is at least 2.

15.2 Support, characteristic, and Minkowski functions

Exercise III.10. [characteristic and support functions of convex sets] Let $X \subset \mathbf{R}^n$ be a nonempty convex set. *Characteristic* (a.k.a *indicator*) *function* of X is, by definition, the function

$$\chi_X(x) = \begin{cases} 0 & , x \in X \\ +\infty & , x \notin X \end{cases}$$

As is immediately seen, this function is convex and proper. The Legendre transform of this function is called the *support function* $\phi_X(x)$ of X:

$$\phi_X(x) = \sup_u [x^\top u - \chi_X(u)] = \sup_{u\in X} x^\top u.$$

1. Prove that χ_X is lower semicontinuous (lsc) if and only if X is closed, and that the support functions of X and $\operatorname{cl} X$ are the same.

Solution: Lower semicontinuity of convex function is, by Proposition III.12.2, exactly the same as closedness of the epigraph of the function. The epigraph of $\chi_X(\cdot)$ is exactly $X \times \mathbf{R}_+$, and this set is closed if and only if X is so. And of course

$$\phi_X(x) = \sup_{u\in X} x^\top u = \sup_{u\in \operatorname{cl} X} x^\top u,$$

so that the support functions of X and $\operatorname{cl} X$ are the same.

In the remaining part of Exercise, we are interested in properties of support functions, and in view of item 1, it makes sense to assume from now on that X, on the top of being nonempty and convex, is also closed.

Prove the following facts:

2. $\phi_X(\cdot)$ is proper lsc convex function which is positively homogeneous of degree 1:

$$\forall(x \in \mathrm{Dom}\,\phi_x, \lambda \geq 0) : \phi_X(\lambda x) = \lambda\phi_X(x).$$

In particular, the domain of ϕ_X is a cone. Demonstrate by example that this cone not necessarily is closed (look at the support function of the closed convex set $\{[v; w] \in \mathbf{R}^2 : v > 0, w \leq \ln v\}$).

Solution: $\phi_X(\cdot)$ is convex, proper and lsc, as Legendre transform of a whatever proper convex function. And of course whenever x is such that $\sup_{u\in X} x^\top u < \infty$, we have $\sup_{u\in X}[\lambda x]^\top u = \lambda \sup_{u\in X} x^\top u$ for all $\lambda \geq 0$.
Finally, for $X = \{[v; w] \in \mathbf{R}^2 : v > 0, w \leq \ln v\}$ we have

$$\begin{array}{rcl} \phi_X([x_1; x_2]) & = & \sup_{v,w}\{x_1 v + x_2 w : v > 0, w \leq \ln v\} \\ & = & \left\{\begin{array}{rll} +\infty & , x_1 > 0 & (a) \\ +\infty & , x_1 \leq 0, x_2 < 0 & (b) \\ +\infty & , x_1 = 0, x_2 \neq 0 & (c) \\ < +\infty & , x_1 < 0, x_2 \geq 0 & (d) \\ < +\infty & , x_1 = x_2 = 0 & (e) \end{array}\right. \end{array}$$

(to justify (a) and (c), set $[v; w] = [v; \ln v]$ and look what happens when $v \to \infty$ and when $v \to +0$, to justify (b), look what happens when $[v; w] = [1; w]$ and $w \to -\infty$). We see that $\mathrm{Dom}\,\phi_X$ is the second quadrant $\{x_1 \leq 0, x_2 \geq 0\}$ with eliminated open ray $\{[0; x_2] : x_2 > 0\}$, and this set is just a cone, not a closed one.

3. Vice versa, every proper convex lsc function ϕ which is positively homogeneous of degree 1,

$$(x \in \mathrm{Dom}\, f, \lambda \geq 0) \Longrightarrow \phi(\lambda x) = \lambda\phi(x)$$

is the support function of a nonempty closed convex set, specifically, its subdifferential $\partial\phi(0)$ taken at the origin. In particular, $\phi_X(\cdot)$ "remembers" X: if X, Y are nonempty closed convex sets, then $\phi_X(\cdot) \equiv \phi_Y(\cdot)$ if and only if $X = Y$.

Solution: Let ϕ be proper lsc convex and positively homogeneous, of degree 1, function, and let $\chi(x)$ be the Legendre transform of ϕ. As every Legendre transform of proper convex function, χ is proper, convex and lsc. In addition, from properness and positive homogeneity of ϕ it follows that $0 \in \mathrm{Dom}\,\phi$ and $\phi(0) = 0$, whence

$$\chi(u) = \sup_x \left\{u^\top x - \phi(x)\right\} \geq u^\top 0 - \phi(0) = 0.$$

It remains to prove that χ takes just two values, 0 and $+\infty$; given this, we immediately conclude that χ is the characteristic function of its (nonempty, convex, and closed due to properness, convexity and lower semicontinuity of χ, see item 1 of Exercise) domain. Indeed, we already know that $\chi(\cdot) \geq 0$; what remains to prove is that if $\chi(u) > 0$ for some u, then in fact $\chi(u) = \infty$. Relation $\chi(u) > 0$ amounts to existence of x such that $u^\top x - \phi(x) > 0$; but then, due to positive homogeneity of ϕ, for $\lambda > 0$ if holds $u^\top[\lambda x] - \phi(\lambda x) = \lambda[u^\top x - \phi(x)] \to \infty$, $\lambda \to \infty$, that is, $\chi(u) = +\infty$, as claimed.
Finally, from Proposition III.13.3 it follows that ϕ, being proper convex lsc function, is the Legendre transform χ^* of χ, that is, of the characteristic function of nonempty closed convex domain. By item **B** from chapter 13, see p. 213, the subdifferential of $\phi \equiv \chi^*$ taken at the origin is the set of all minimizers of χ, and this set for characteristic function is nothing but its domain. ■

4. Let X, Y be two nonempty closed convex sets. Then $\phi_X(\cdot) \geq \phi_Y(\cdot)$ if and only if $Y \subseteq X$.

Solution: For proper lsc convex functions f, g and their Legendre transforms x^*, g^* the relation $f(\cdot) \leq g(\cdot)$ clearly implies that $f^*(\cdot) \geq g^*(\cdot)$; since f and g are the Legendre transforms of their Legendre transforms (Proposition III.13.3), the latter relation, in turn, implies that $f \leq g$. Thus, for proper lsc convex functions f, g the relation $f \leq g$ is equivalent to $f^* \geq g^*$. In particular, $\phi_X(\cdot) \geq \phi_Y(\cdot)$ if and only if $\chi_X(\cdot) \leq \chi_Y(\cdot)$, and the latter relation clearly takes place if and only if $Y \subset X$. ■

5. $\mathrm{Dom}\,\phi_X = \mathbf{R}^n$ if and only if X is bounded.

Solution: When X is bounded, $\phi_X(\cdot)$ clearly is real-valued on the entire space. Vice versa, if the convex function $\phi_X(\cdot)$ is real valued on the entire space, $\partial\phi_X(0)$ is bounded by Proposition III.12.10; it remains to note that by item 3 of Exercise, $X = \partial\phi_X(0)$. ■

6. Let X be the unit ball of some norm $\|\cdot\|$. Then ϕ_X is nothing but the norm $\|\cdot\|_*$ conjugate to $\|\cdot\|$. In particular, when $p \in [1,\infty]$ and $X = \{x \in \mathbf{R}^n : \|x\|_p \le 1\}$, we have $\phi_X(x) \equiv \|x\|_q$, $\frac{1}{q} + \frac{1}{p} = 1$.

Solution: This is nothing but straightforward rewording of Fact III.13.4.

7. Let $x \mapsto Ax + b : \mathbf{R}^n \to \mathbf{R}^m$ be an affine mapping, and let $Y = AX + b = \{Ax + b : x \in X\}$. Then
$$\phi_Y(v) = \phi_X(A^\top v) + b^\top v.$$

Solution: Indeed, $\phi_Y(v) = \sup_{y\in Y} u^\top y = \sup_{x\in X} u^\top[Ax+b] = b^\top u + \sup_{x\in X}[A^\top v]^\top x = b^\top u + \phi_X(A^\top v)$. ■

Exercise III.11. [Minkowski functions of convex sets] The goal of this Exercise is to acquaint the reader with important special family of convex functions – Minkowski functions of convex sets.

Consider a proper *nonnegative* lower semicontinuous function $f : \mathbf{R}^n \to \mathbf{R} \cup \{+\infty\}$ which is *positively homogeneous of degree 1*, meaning that
$$x \in \mathrm{Dom}\, f,\ t \ge 0 \implies tx \in \mathrm{Dom}\, f \ \text{ and } \ f(tx) = tf(x).$$
Note that from the latter property of f and its properness it follows that $0 \in \mathrm{Dom}\, f$ and $f(0) = 0$.

We can associate with f its *basic sublevel set*
$$X = \{x \in \mathbf{R}^n : f(x) \le 1\}.$$
Note that X "remembers" f, specifically
$$\forall t > 0 : f(x) \le t \iff f(t^{-1}x) \le 1 \iff t^{-1}x \in X,$$
whence also
$$\begin{array}{c}\forall x \in \mathbf{R}^n : f(x) = \inf\{t : t > 0, t^{-1}x \in X\}\\ [\inf\{t : t > 0, t \in \varnothing\} = +\infty \text{ by definition}]\end{array} \tag{15.1}$$
Note that the basic sublevel set of our f cannot be arbitrary: it is convex and closed (since f is convex lsc) and contains the origin (since $f(0) = 0$).

Now, given a closed convex set $X \subset \mathbf{R}^n$ containing the origin, we can associate with it a function $f : \mathbf{R}^n \to \mathbf{R} \cup \{+\infty\}$ by construction from (15.1), specifically, as
$$f(x) = \inf\{t : t > 0, t^{-1}x \in X\} \tag{15.2}$$
This function is called *the Minkowski function* (M.f.) of X.

Here goes your first task:

1. Prove that when $X \subset \mathbf{R}^n$ is convex, closed, bounded, and contains the origin, function f given by (15.2) is proper, nonnegative, convex lsc function positively homogeneous of degree 1, and X is the basic sublevel set of f. Moreover, f is nothing but the support function ϕ_{X_*} of the polar X_* of X.

Solution: The polar X_* of X is closed convex set containing the origin, and therefore its support function $\overline{f}$, as the support function of any nonempty convex set, is convex lsc and positively homogeneous of degree 1 by Exercise III.10.2. Nonnegativity of $\overline{f}$ is readily given by the inclusion $0 \in X_*$. Thus, all which remains to verify is that in fact $f = \overline{f}$ and X is the basic sublevel set of $\overline{f}$. Verification of the equality $f = \overline{f}$ is immediate:
$$\begin{array}{l}\forall t > 0 : \overline{f}(x) \le t \iff \overline{f}(t^{-1}x) \le 1 \iff \sup_{y\in X_*}[t^{-1}x]^\top y \le 1\\ \iff t^{-1}x \in \mathrm{Polar}\,(X_*) = X,\end{array}$$
which combines with (15.2) to imply that whenever $t > 0$, relation $t \ge f(x)$ is the same as the relation $t \ge \overline{f}(x)$; since f and $\overline{f}$ are nonnegative, it follows that $f \equiv \overline{f}$. To see that X is the basic sublevel set of $f \equiv \overline{f}$, note that the basic sublevel set of the support function of X_* clearly is $\mathrm{Polar}\,(X_*) = X$.

Your next tasks are as follows:

2. What are the Minkowski functions of
   - the singleton $\{0\}$?
   - a linear subspace?
   - a closed cone $\mathbf{K}$?
   - the unit ball of a norm $\|\cdot\|$?
3. Prove that the Minkowski functions f_X, f_Y of closed convex and containing the origin sets X, Y are linked by the relation $f_X \geq f_Y$ if and only if $X \subseteq Y$.
4. When the Minkowski function of a set X (convex, closed, bounded, and containing the origin) does not take value $+\infty$?
5. What is the set of zeros of the Minkowski function of a set X (convex, closed, bounded, and containing the origin)?
6. What is the Minkowski function of the intersection $\cap_{k\leq K} X_k$ of closed convex sets containing the origin?

Solution: 2: The Minkowski function (M.f.) of a closed cone (in particular, of a linear subspace) is nothing but the characteristic function of this set. The M.f. of the unit $\|\cdot\|$-ball is the norm $\|\cdot\|$.

3: This is immediate consequence of the fact that f_X, f_Y are the support functions of the polars X_*, Y_* of X, Y combined with the the result of Exercise III.10.4 and the fact that passing to polars reverses inclusions.

4: The M.f. f_X of a closed convex set X containing the origin is real-valued if and only if X contains a neighbourhood of the origin. Indeed, if X contains a centered at the origin $\|\cdot\|_2$-ball of raduis $r > 0$, f_X, by item 3, does not exceed the real-valued M.f. $r^{-1}\|\cdot\|_2$ of this ball. And if f_X is real-valued, small enough positive multiples of $\pm e_i$ (e_i are the standard basic orths) belong to X, so that the origin is an interior point of X.

5: The set of zeros of the M.f. of X is exactly the recessive cone of X.

6: The M.f. in question is the maximum of the M.f.'s of X_k.

Exercise III.12.

1. Recall that the closed conic transform
$$\overline{\mathrm{ConeT}}(X) = \mathrm{cl}\,\{[x;t] \in \mathbf{R}^n \times \mathbf{R} :\ t > 0,\ x/t \in X\},$$
of a nonempty convex set $X \subset \mathbf{R}^n$ (see section 1.5) is a closed cone such that
$$\mathrm{cl}(X) = \{x : [x;1] \in \overline{\mathrm{ConeT}}(X)\}.$$
What is the cone dual to $\overline{\mathrm{ConeT}}(X)$?
2. Let $X \subset \mathbf{R}^n$ be a nonempty closed convex set and $X^+ = \overline{\mathrm{ConeT}}(X)$. Prove that
$$X_t^+ := \{x : [x;t] \in X^+\} = \begin{cases} tX, & t > 0 \quad (a) \\ \mathrm{Rec}(X), & t = 0 \quad (b) \\ \varnothing, & t < 0 \quad (c) \end{cases}$$
3. Let $X_1, ..., X_K$ be closed convex sets in $\mathbf{R}^n$ with nonempty intersection X. Prove that
$$\overline{\mathrm{ConeT}}(X) = \bigcap_k \overline{\mathrm{ConeT}}(X_k).$$
4. 4 Let $X = \bigcap_{k\leq K} X_k$, where $X_1, ..., X_K$ are closed convex sets in $\mathbf{R}^n$ such that $X_K \cap \mathrm{int}\, X_1 \cap \mathrm{int}\, X_2 ... \cap \mathrm{int}\, X_{K-1} \neq \varnothing$. Prove that $\phi_X(y) \leq a$ if and only if there exist y_k, $k \leq K$, such that
$$y = \sum_k y_k \quad \text{and} \quad \sum_k \phi_{X_k}(y_k) \leq a. \qquad (*)$$
In words: *the supremum of a linear form on $\bigcap_k X_k$ does not exceed some a if and only if the form can be decomposed into the sum of K forms with the sum of their suprema over the respective sets X_k not exceeding a.*

5. Prove the following polyhedral version of the claim in item 4:
Let $X_k = \{x \in \mathbf{R}^n : A_k x \le b_k\}$, $k \le K$, be polyhedral sets with nonempty intersection X. A linear form does not exceed some $a \in \mathbf{R}$ everywhere on X if and only if the form can be decomposed into the sum of K linear forms with the sum of their maxima on respective sets X_k not exceeding a.

Solution: 1: This is the cone

$$\{[y;s] \in \mathbf{R}^n_y \times \mathbf{R}_s : s \ge \phi_X(-y)\},$$

where

$$\phi_X(y) = \sup_{x \in X} y^\top x$$

is the support function of X.

Indeed, from the definition of the closed conic transform it immediately follows that $[y;s]^\top [x;t] \ge 0$ for all $[x;t] \in \overline{\mathrm{ConeT}}(X)$ if and only if $[y;s]^\top [x;1] \ge 0$ for all $x \in X$, that is, if and only if

$$0 \le s + \inf_{x \in X} y^\top x = s - \sup_{x \in X} [-y]^\top x = s - \phi_X(-y). \blacksquare$$

2: A point $[x;t]$ belongs to X^+ if and only if there exist a sequence $[x_i;t_i]$ converging to $[x;t]$ and such that $t_i > 0$ and $x_i = t_i y_i$ with $y_i \in X$. When $t > 0$, the points $y_i = x_i/t_i$ have the limit $y = x/t$ as $i \to \infty$, and $y \in X$ since X is closed; vice versa, if $y \in X$ and $t > 0$, the point $[ty;t]$ clearly belongs to $X^+ \cap \Pi_t$; we have proved (a). (c) is evident. (b) is stated in Fact II.6.19.

3: Let Π_s be the hyperplane $\{[x;s] : x \in \mathbf{R}^n\}$ in $\mathbf{R}^{n+1}$. By (a) and (c) in the previous item, we have $\overline{\mathrm{ConeT}}(X) \cap \Pi_t = \cap_k [\overline{\mathrm{ConeT}}(X_k) \cap \Pi_t]$ for $t \ne 0$. Since X_k are closed convex sets with nonempty intersection, we have $\mathrm{Rec}(\cap_k X_k) = \cap_k \mathrm{Rec}(X_k)$, whence $\overline{\mathrm{ConeT}}(X) \cap \Pi_0 = \cap_k [\overline{\mathrm{ConeT}}(X_k) \cap \Pi_0]$ as well.

4: There is nothing to prove when $a = \infty$. Now let $a \in \mathbf{R}$. When $(*)$ takes place and $x \in X$, for every k we have $x \in X_k$, so that $y_k^\top x \le \phi_{X_k}(y_k) \le a_k$. Summing up the resulting inequalities and taking into account that $\sum_k y_k = y$, we get $y^\top x \le \sum_k a_k \le a$. The resulting inequality holds true for every $x \in X$, implying $\phi_X(y) \le a$.

Vice versa, let $\phi_X(y) \le a \in \mathbf{R}$. Let $X_k^+ = \overline{\mathrm{ConeT}}(X_k)$, $k \le K$. By item 3, $\overline{\mathrm{ConeT}}(X) = \cap_k \overline{\mathrm{ConeT}}(X_k)$, and by item 1 we have $[-y;a] \in [\overline{\mathrm{ConeT}}(X)]_*$ The cones $M^k = \overline{\mathrm{ConeT}}(X_k)$ are closed, and their interiors clearly contain the sets $\{[x;1] : x \in \mathrm{int}\, X_k\}$. It follows $M^K \cap \mathrm{int}\, M^1 \cap ... \cap \mathrm{int}\, M^{K-1} \ne \varnothing$. We see that the linear from with the vector of coefficients $[-y;a]$ and cones $M^1, ..., M^K$ satisfy the premise of the Dubovitski-Milutin Lemma. By this lemma, there exists a decomposition

$$[-y;a] = \sum_k [-y_k;a_k]$$

with $[-y_k;a_k] \in M_*^k$, that is, invoking item 1, with $\phi_{X_k}(y_k) \le a_k$, $k \le K$. $\blacksquare$

5: We could prove this claim by slight modification of the reasoning for item 4, but it is easier to get it immediately from LP duality: for $\bar y \in \mathbf{R}^n$ and $a \in \mathbf{R}$ we have

$$\begin{array}{ll} & a \ge \sup_{x \in X} \bar y^\top x \implies a \ge \max_x \{\bar y^\top x : A_k x \le b_k, k \le K\} \\ \Longrightarrow & a \ge \min_{z_1,...,z_K} \{\sum_k b_k^\top z_k : z_k \ge 0\, \forall k, \sum_k A_k^\top z_k = \bar y\} \text{ [LP Duality Theorem]} \\ \Longrightarrow & \exists (z_1 \ge 0, ..., z_K \ge 0) : \sum_k \underbrace{A_k^\top z_k}_{y_k} = \bar y, \sum_k \underbrace{b_k^\top z_k}_{a_k} \le a \\ \Longrightarrow & \exists y_1, ..., y_K, a_1, ..., a_K : a_k \ge \max_x \{y_k^\top x_k : A_k x \le b_k\},\ k \le K, \sum_k y_k = \bar y \\ & \text{[LP Duality Theorem]} \end{array}$$

We see that *if* $\bar y^\top x \le a \in \mathbf{R}\, \forall x \in X$, *then* the linear form $\bar y^\top x$ can be decomposed into the sum of linear forms $y_k^\top x$ with the sum of maxima of the forms on the respective sets X_k not exceeding a. The inverse statement is evident.

Exercise III.13. Let $X \subset \mathbf{R}^n$ be a nonempty polyhedral set given by polyhedral representation

$$X = \{x : \exists u : Ax + Bu \le r\}.$$

Build polyhedral representation of the epigraph of the support function of X. For non-polyhedral extension, see Exercise IV.36.

Solution: We have

$$\begin{array}{rcl} t \ge \phi_X(y) & \iff & t \ge \mathrm{Opt}(P) := \max_{x,u}\{y^\top x : Ax + Bu \le r\} \\ & \iff & t \ge \mathrm{Opt}(D) := \min_\lambda \left\{r^\top\lambda : \lambda \ge 0, A^\top\lambda = y, B^\top\lambda = 0\right\} \\ & & [\text{LP Duality Theorem; note that } (P) \text{ is feasible due to } X \ne \varnothing] \\ & \iff & \exists\lambda : r^\top\lambda \le t, \lambda \ge 0, A^\top\lambda = y, B^\top\lambda = 0 \\ & & [\text{since by the above, } (D) \text{ is solvable whenever } t \ge \phi_X(y)] \end{array}$$

and we end up with polyhedral representation of $\mathrm{epi}\{\phi_X\}$.

Exercise III.14. Compute in closed analytic form the support functions of the following sets:

1. The ellipsoid $\{x \in \mathbf{R}^n : (x-c)^\top C(x-c) \le 1\}$ with $C \succ 0$

   Solution: $\phi(y) = \sqrt{y^\top C^{-1} y} + c^\top y$

2. The probabilistic simplex $\{x \in \mathbf{R}^n_+ : \sum_i x_i = 1\}$

Solution: $\phi(y) = \max_{i \le n} y_i$.

3. The nonnegative part of the unit $\|\cdot\|_p$-ball: $X = \{x \in \mathbf{R}^n_+ : \|x\|_p \le 1\}$, $p \in [1, \infty]$

Solution: $\phi(y) = \|[y]_+\|_q$, where $\frac{1}{p} + \frac{1}{q} = 1$ and $[[y_1; ...; y_n]]_+ = [\max[y_1; 0]; ...; \max[y_n, 0]]$.

4. The positive semidefinite part of the unit $\|\cdot\|_{p,\mathrm{Sh}}$ norm: $X = \{x \in \mathbf{S}^n_+ : \|x\|_{p,\mathrm{Sh}} \le 1\}$

Solution: $\phi(y) = \|[y]_+\|_{q,\mathrm{Sh}}$, where $\frac{1}{p} + \frac{1}{q} = 1$ and $[y]_+$ is the "positive semidefinite part of y" – the matrix obtained from symmetric matrix y by keeping intact all eigenvectors and nonnegative eigenvalues, and zeroing out the negative eigenvalues (what is called the function $[\cdot]_+$ as applied to a symmetric matrix, see section D.1.5).

5. The paraboloid $\{x \in \mathbf{R}^{n+1} : x_{n+1} \ge \frac{1}{2}\sum_{i=1}^n x_i^2\}$ $(n \ge 1)$.

Solution: $\phi(y) = \begin{cases} -\frac{\sum_{i=1}^n y_i^2}{2y_{n+1}} & , y_{n+1} < 0 \\ 0 & , y = 0 \\ +\infty & , \text{all other cases} \end{cases}$

15.3 Around subdifferentials

Exercise III.15. Let f be a convex function and $\bar{x} \in \mathrm{Dom}\, f \subset \mathbf{R}^n$. Prove that the property of $g \in \mathbf{R}^n$ to be a subgradient of f at $\bar{x}$ is local: the inequality

$$f(x) \ge f(\bar{x}) + g^\top(x - \bar{x}) \qquad (*)$$

hods true for all $x \in \mathbf{R}^n$ iff it holds true for all x in a neighborhood of $\bar{x}$.

Solution: In one direction the claim is evident. Now assume that $(*)$ holds true for all x in a neighborhood of $\bar{x}$, and let us prove that it holds true for all x. Indeed, let $\bar{f}(x) = f(x) - f(\bar{x}) - g^\top(x - \bar{x})$, so that $\bar{f}$ is convex along with f by calculus of convexity. Validity of $(*)$ is a neighborhood of $\bar{x}$ means that $\bar{x}$ is a local minimizer of $\bar{f}$: $\bar{f}(x) \ge \bar{f}(\bar{x}) = 0$ for all x from a neighborhood of $\bar{x}$. By unimodality (Theorem III.11.1 applied with $Q = \mathbf{R}^n$ and $x^* = \bar{x}$ to $\bar{f}$ in the role of f) $\bar{x}$ is a global minimizer of $\bar{f}$, so that $\bar{f}(x) \ge \bar{f}(\bar{x}) = 0$ for all x. Recalling what $\bar{f}$ is, we see that $(*)$ holds true for all x. ■

Exercise III.16. [subdifferentials of norms] Let $\|\cdot\|$ be a norm on $\mathbf{R}^n$, and $\|\cdot\|_*$ be its conjugate (see Fact III.13.4). Prove that

1. The subdifferential of $\|\cdot\|$ taken at the origin is the unit ball B_* of $\|\cdot\|_*$, or, which is the same, the polar
$$\left\{u:\ u^\top x \le 1\,\forall(x:\|x\|\le 1)\right\}$$
of the unit ball B of the norm $\|\cdot\|$.
2. When $x\neq 0$, the subdifferential of $\|\cdot\|$ taken at x is the set $\{u\in B_*: u^\top x=\|x\|\}$. In particular, the subdifferential of $\|\cdot\|$ remains intact when replacing x with tx, $t>0$, and is reflected w.r.t. the origin when x is replaced with tx, $t<0$.

Solution: 1: By Fact III.13.4, the Legendre transform of $\|\cdot\|$ is the characteristic function of B_*, so that $\|\cdot\|$, being real-valued convex continuous (and thus lsc) function, by Proposition III.13.3 is the Legendre transform of the characteristic function of the closed nonempty convex set B_*, or, which is the same, $\|\cdot\|$ is the support function of B_*. By item 3 of Exercise III.10, $\|\cdot\|$ is the support function of its subdifferential, taken at the origin, which also is a closed nonempty convex set. As we know from Exercise III.10.4, the support functions of nonempty closed convex sets coincide iff the sets coincide, so that the subdifferential of $\|\cdot\|$ taken at the origin is B_*. ■
2: Let $x\neq 0$ and X be the subdifferential of $\|\cdot\|$ taken at x; this is a nonempty convex compact set (Proposition III.12.10). When $g\in X$, we should have
$$g^\top(ty-x)\le\|ty\|-\|x\|,\ t>0,$$
which, after dividing both sides by t and passing to limit as $t\to\infty$, implies that $g^\top y\le\|y\|$ for all y, so that $g\in B_*=\{h: h^\top y\le 1\,\forall(y,\|y\|\le 1)\}$ (see Fact III.13.4). On the other hand, for $\epsilon\in(0,1)$ one has $g^\top[(1-\epsilon)x-x]\le\|(1-\epsilon)x\|-\|x\|=-\epsilon\|x\|$, implying thai $g^\top x\ge\|x\|$. Strict inequality is impossible due to already proved $g\in B_*$, and we conclude that $g^\top x=\|x\|$. Thus, $X\subset\{g\in B_*: g^\top x=\|x\|\}$. On the other hand, when $g\in B_*$ is such that $g^\top x=\|x\|$, we have for every $y\in\mathbf{R}^n$
$$\|y\|\ge g^\top y=g^\top(y-x)+g^\top x=g^\top(y-x)+\|x\|,$$
where the first inequality is due to $g\in B_*$. Thus, $\|x\|+g^\top(y-x)\le\|y\|$ for all y, implying that $g\in X$. ■

Exercise III.17. [Shatten norms] Let $p\in[1,\infty]$. The space $\mathbf{S}^n$ of symmetric $n\times n$ matrices can be equipped with *Shatten p-norms* – matrix analogies of the standard $\|\cdot\|_p$-norms on $\mathbf{R}^n$. Specifically, Shatten p-norm $\|\cdot\|_{p,\mathrm{Sh}}$ of symmetric matrix X is defined as
$$\|X\|_{p,\mathrm{Sh}}=\|\lambda(X)\|_p,$$
where $\lambda(X)$, as always, is the vector of eigenvalues of X.

1. Prove that Shatten norms indeed are norms, and the norm conjugate to $\|\cdot\|_{p,\mathrm{Sh}}$ is $\|\cdot\|_{q,\mathrm{Sh}}$, $\frac{1}{p}+\frac{1}{q}=1$:
$$\|X\|_{q,\mathrm{Sh}}=\max_Y\{\mathrm{Tr}(XY):\|Y\|_{p,\mathrm{Sh}}\le 1\}\tag{15.3}$$
2. Verify that $\|\cdot\|_{2,\mathrm{Sh}}$ is nothing but the Frobenius norm of X, and $\|X\|_{\infty,\mathrm{Sh}}$ is the same as the spectral norm of X.

Solution: 1: The facts that $\|\cdot\|_{p,\mathrm{Sh}}$ is positive outside of the origin and satisfies $\|\lambda X\|_{p,\mathrm{Sh}}=|\lambda|\|X\|_{p,\mathrm{Sh}}$ are evident. Therefore, all we need to justify all claims in item 1 is to justify (15.3), which, as a byproduct, implies convexity of $\|\cdot\|_{p,\mathrm{Sh}}$, which for positively homogeneous, of degree 1, functions implies the Triangle inequality. To justify (15.3), let $X=U\Lambda U^\top$ be the eigenvalue decomposition of $X\in\mathbf{S}^n$, so that $\Lambda=\mathrm{Diag}\{\lambda(X)\}$. Denoting by $\mathrm{Dg}\{Z\}$ the vector of diagonal entries in matrix Z, we have
$$\begin{array}{rcl}\forall(Y\in\mathbf{S}^n):\mathrm{Tr}(XY) & = & \mathrm{Tr}(U\Lambda U^\top Y)=\mathrm{Tr}(\Lambda[U^\top YU])=\lambda^\top(X)\mathrm{Dg}\{U^\top YU\}\\ & \le & \|\lambda(X)\|_q\|\mathrm{Dg}\{U^\top YU\}\|_p\le\|\lambda(X)\|_q\|\lambda(Y)\|_p,\end{array}$$
where the concluding inequality is due to Proposition III.14.3 as applied with $f(x)=\|x\|_p$. We see that the right hand side in (15.3) is $\le$ the left hand side. On the other hand, we can select $g\in\mathbf{R}^n$ in such a way that $\|g\|_p=1$ and $\lambda^\top(X)g=\|\lambda(X)\|_q$ (since $\|\cdot\|_q$ is the conjugate of $\|\cdot\|_p$). Setting

$Y = U\,\mathrm{Diag}\{g\}U^\top$, we get $\lambda(Y) = g$, $\|Y\|_{p,\mathrm{Sh}} = 1$, and $U^\top YU = \mathrm{Diag}\{g\}$, whence by the above computation, $\mathrm{Tr}(XY) = \lambda^\top(X)\lambda(Y) = \|\lambda(X)\|_q$, so that the right hand side in (15.3) is $\geq$ the left hand side. Thus, (15.3) does hold true. ■

2: Taken together, eigenvalue decomposition and the fact that multiplication of matrix from the left and from the right by orthogonal matrices preserves the Frobenius norm (Fact D.2) demonstrate that the Frobenius norm of a symmetric matrix is the same as $\|\cdot\|_2$-norm of its vector of eigenvalues. The fact that $\|\cdot\|_{\infty,\mathrm{Sh}}$ is the spectral norm is evident. ■

Exercise III.18. [chain rule for subdifferentials] Let $Y \subset \mathbf{R}^m$ and $X \subset \mathbf{R}^n$ be nonempty convex sets, $\overline{y} \in Y$, $\overline{x} \in X$, let $f(\cdot) : Y \to \mathbf{R}$ be a convex function, and let $A(\cdot) : X \to Y$ with $A(\overline{x}) = \overline{y}$. Further, let $\mathbf{K}$ be a closed cone in $\mathbf{R}^n$. A function f is called $\mathbf{K}$-*monotone* on Y, if for $y, y' \in Y$ such that $y' - y \in \mathbf{K}$ it holds that $f(y') \geq f(y)$, and A is called $\mathbf{K}$-*convex* on X if for all $x, x' \in X$ and $\lambda \in [0,1]$ it holds that $\lambda A(x) + (1-\lambda)A(x') - A(\lambda x + (1-\lambda)x') \in \mathbf{K}$.

Prove that

1. A is $\mathbf{K}$-convex on X if and only if for every $\phi \in \mathbf{K}_*$ the real-valued function $\phi^\top A(x)$ is convex on X.

   Solution: Indeed, since $\mathbf{K}$ is closed, we have $\mathbf{K} = (\mathbf{K}_*)_*$, so that $\lambda A(X) + (1-\lambda)A(x') - A(\lambda x + (1-\lambda)x') \in \mathbf{K}$ if and only if $\lambda\phi^\top A(x) + (1-\lambda)\phi^\top A(x') - \phi^\top A(\lambda x + (1-\lambda)x') \geq 0$ for all $\phi \in \mathbf{K}_*$.

2. Let A be $\mathbf{K}$-convex on X and differentiable at $\overline{x}$. Let $A'(\overline{x})$ denote the Jacobian of A at $\overline{x}$. Prove that

$$\forall x \in X : \ A(x) - [A(\overline{x}) + A'(\overline{x})[x - \overline{x}]] \in \mathbf{K}. \qquad (*)$$

   Solution: By item 1, for $\phi \in \mathbf{K}_*$ the function $\phi^\top A(x)$ is convex on X, and by the standard Calculus it is differentiable at $\overline{x}$ with the derivative $\phi^\top A'(x)$. Therefore by Gradient inequality one has

$$\forall x \in X : \phi^\top \left[A(x) - [A(\overline{x}) + A'(\overline{x})[x - \overline{x}]]\right] = \phi^\top A(x) - [\phi^\top A(\overline{x}) + \phi^\top A'(\overline{x})[x - \overline{x}]] \geq 0,$$

   and $(*)$ follows.

3. Let f be $\mathbf{K}$-monotone on Y and let A be $\mathbf{K}$-convex on X. Prove that the function $f\circ A\,(x) = f(A(x))$ is real valued and also convex on X.

   Solution: Indeed, for $x, x' \in X$ and $\lambda \in [0,1]$ the points $y = A(x)$, $y' = A(x')$, $w = \lambda y + (1-\lambda)y'$ and $z = A(\lambda x + (1-\lambda)x')$ belong to Y since Y is convex and A maps X into Y, and $w - z \in \mathbf{K}$ since A is $\mathbf{K}$-convex. Since f is $\mathbf{K}$-monotone, $w - z \in \mathbf{K}$ implies that $f(w) \geq f(z)$. Besides this, recalling what w is and that f is convex, $\lambda f(y) + (1-\lambda)f(y') \geq f(w)$. The bottom line is that $\lambda f(y) + (1-\lambda)f(y') \geq f(z)$, that is,

$$f\circ A\,(\lambda x + (1-\lambda)x') = f(z) \leq \lambda f(y) + (1-\lambda)f(y') = \lambda f\circ A\,(x) + (1-\lambda)f\circ A\,(x').$$

   The resulting inequality hods true for all $x, x' \in X$ and $\lambda \in [0,1]$, so that $f\circ A$ is convex on X.

4. Let f be $\mathbf{K}$-monotone on Y. Prove that $\partial f(\overline{y}) \subseteq \mathbf{K}_*$, provided $\overline{y} \in \mathrm{int}\, Y$.

   Solution: Indeed, let $g \in \partial f(\overline{y})$ and $h \in \mathbf{K}$. Since $\overline{y} \in \mathrm{int}\, Y$, we have $\overline{y} - th \in Y$ for small positive t, and $f(\overline{y} - th) \leq f(\overline{y})$ by $\mathbf{K}$-monotonicity of f. Besides this, $f(\overline{y} - th) \geq f(\overline{y}) - tg^\top h$ due to $g \in \partial f(\overline{y})$. Thus, for all small positive t it holds

$$f(\overline{y}) \geq f(\overline{y}) - tg^\top h,$$

   implying that $g^\top h \geq 0$. This relation holds true for every $g \in \partial f(\overline{y})$ and $h \in \mathbf{K}$, implying that $\partial f(\overline{y}) \subseteq \mathbf{K}_*$.

5. [chain rule] Let $\overline{y} \in \mathrm{int}\, Y$, $\overline{x} \in \mathrm{int}\, X$; let f be $\mathbf{K}$-monotone on Y, and let A be $\mathbf{K}$-convex on X and differentiable at $\overline{x}$. Prove that

$$\partial f\circ A\,(\overline{x}) = [A'(\overline{x})]^\top \partial f(\overline{y}) = \{[A'(\overline{x})]^\top g : \ g \in \partial f(\overline{y})\} \qquad (!)$$

Solution: Let us first verify that the right hand side set in (!) is contained in the left hand side one. Indeed, let $g \in \partial f(\overline{y})$, $x \in X$, and $y = A(x)$. We have

$$\begin{array}{l} f\circ A\,(x) = f(y) \geq f(\overline{y}) + g^\top[y-\overline{y}] = f(\overline{y}) + g^\top[A(x) - A(\overline{x})] \\ \geq f(\overline{y}) + g^\top A'(\overline{x})[x-\overline{x}] \text{ [since } g \in \mathbf{K}_* \text{ and due to } (*)] \\ = f\circ A\,(\overline{x}) + g^\top A'(\overline{x})[x-\overline{x}]. \end{array}$$

The resulting inequality holds true for all $x \in X$ and $g \in \partial f(\overline{y})$, implying that $[A'(\overline{x})]^\top \partial f(\overline{y})) \subset \partial f\circ A\,(\overline{x})$.

Now let us prove that the left hand side set in (!) is contained in the right hand side set, let it be called D. $\overline{y} \in \operatorname{int} Y$, so that $\partial f(\overline{y})$ is a nonempty convex compact set; therefore D also is nonempty convex compact set. Assume, on the contrary to what should be proved, that there exists $e \in \partial f\circ A\,(\overline{x}) \setminus D$. By Separation Theorem, there exists $h \in \mathbf{R}^n$ such that

$$h^\top e > \alpha = \max_{z\in D} h^\top z = \max_{g\in\partial f(\overline{y})} g^\top A'(\overline{x})h.$$

For small positive t from differentiability of A at $\overline{x}$ it follows that

$$y_t := A(\overline{x}+th) = \overline{y} + tA'(\overline{x})h + \epsilon_t,\ \|\epsilon_t\|_2/t \to 0, t \to +0.$$

Since f is convex and real-valued in a neighbourhood of $\overline{y}$, it is Lipschitz continuous, with some constant L, in such a neighbourhood, which combines with the above relation to imply that

$$f\circ A\,(\overline{x}+th) = f(y_t) = f(\overline{y} + tA'(\overline{x})h) + \delta_t,\ \delta_t/t \to 0,\ t \to +0,$$

whence

$$\lim_{t\to+0} \frac{f\circ A\,(\overline{x}+th) - f\circ A\,(\overline{x})}{t} = \lim_{t\to+0} \frac{f(\overline{y}+tA'(\overline{x})h) - f(\overline{y})}{t}.$$

By Theorem III.12.12, the left hand side in this equality is $\max_{d\in\partial f\circ A(\overline{x})} d^\top h$ (and is therefore $\geq e^\top h$ due to the origin of e), and the right hand side is $\max_{g\in\partial f(\overline{y})} g^\top A'(\overline{x})h = \alpha$. Thus, $e^\top h \leq \alpha$, which is the desired contradiction.

Exercise III.19. Recall that the sum $S_k(X)$ of $k \leq n$ largest eigenvalues of the $X \in \mathbf{S}^n$ is a convex function of X, see Remark III.14.4. Point out a subgradient of $S_k(\cdot)$ at a point $\overline{X} \in \mathbf{S}^n$. As a special case, find a subgradient of the maximal eigenvalue $\lambda_{\max}(X)$ of $X \in \mathbf{S}^n$ treated as a function of X.

Solution: Let $\overline{X} = \overline{U}\operatorname{Diag}\{\lambda(\overline{X}\}\overline{U}^\top$ be the eigenvalue decomposition of $\overline{X}$. Setting

$$P = \overline{U}\operatorname{Diag}\{\underbrace{1,...,1}_{k},0,...,0\}\overline{U}^\top,$$

we get $\operatorname{Tr}(\overline{X}P) = \sum_{i=1}^k \lambda_k(\overline{X}) = S_k(\overline{X})$. On the other hand, for $X \in \mathbf{S}^n$ we have

$$\begin{array}{rcl} \operatorname{Tr}(XP) & = & \operatorname{Tr}(X\overline{U}\operatorname{Diag}\{1,...,1,0,..,0\}\overline{U}^\top) = \operatorname{Tr}([\overline{U}^\top X\overline{U}]\operatorname{Diag}\{1,...,1,0,...,0\}) \\ & \leq & s_k(\operatorname{Dg}\{\overline{U}^\top X\overline{U}\}), \end{array}$$

where, as always, $s_k(x)$ is the sum of k largest entries in a vector x. By Proposition III.14.3, $s_k(\operatorname{Dg}\{\overline{U}^\top X\overline{U}\}) \leq s_k(\lambda(X)) = S_k(X)$. Thus,

$$\forall X \in \mathbf{S}^n : S_k(X) \geq \operatorname{Tr}(XP) = \operatorname{Tr}(\overline{X}P) + \operatorname{Tr}(P[X-\overline{X}]) = S_k(\overline{X}) + \operatorname{Tr}(P[X-\overline{X}]).$$

Recalling what is the inner product on $\mathbf{S}^n$, we conclude that $P \in \partial S_k(\overline{X})$.

To get a subgradient of $\lambda_{\max}(X)$, note that $\lambda_{\max}(X) \equiv S_1(X)$, so that the above computation says that if $e(X)$ is leading eigenvector of X (i.e., unit $\|\cdot\|_2$-norm eigenvector of X with eigenvalue $\lambda_{\max}(X)$), then $e(X)e^\top(X) \in \partial\lambda_{\max}(X)$.

15.4 Around Legendre transform

Exercise III.20. Compute Legendre transforms of the following univariate functions:

1. $f(x) = -\ln x$, $\mathrm{Dom}\, f = (0, \infty)$

Solution: $f^*(y) = \sup_{x>0}[xy + \ln x]$. When $y \geq 0$, the supremum is $+\infty$ (look what happens when $x \to +\infty$). When $y < 0$, the sufficient condition for some $x > 0$ to maximize $\phi_y(x) = xy + \ln x$ is to be a root of $\phi'_y(x)$ (Theorem III.11.2 as applied to convex differentiable function $-\phi_y(\cdot)$). The equation $\phi'_y(x) = 0$ reads $y + 1/x = 0$, resulting in $x = -1/y$ and $\max_{x>0} \phi_y(x) = \phi_y(-1/y) = -\ln(-y) - 1$. Thus,

$$f^*(y) = -\ln(-y) - 1, \ \mathrm{Dom}\, f^* = (-\infty, 0).$$

2. $f(x) = \mathrm{e}^x$, $\mathrm{Dom}\, f = \mathbf{R}$.

Solution: Setting $\phi_y(x) = xy - \mathrm{e}^x$, we have $\sup_x \phi_y(x) = +\infty$ when $y < 0$ (look what happens when $x \to -\infty$). When $y = 0$, we clearly have $\sup_x \phi_y(x) = 0$. Finally, when $y > 0$, the maximizer of $\phi_y(\cdot)$, same as in the previous item, can be found via Fermat rule – as a root of the equation $\phi'_y(x) = 0$. This equation reads $y - \mathrm{e}^x = 0$, resulting in $x = \ln y$ and $\sup_x \phi_y(x) = y \ln y - y$. Thus,

$$f^*(y) = y \ln y - y, \ \mathrm{Dom}\, f^* = [0, \infty); \ \text{here, as always, } 0 \ln 0 = 0 \text{ by definition.}$$

3. $f(x) = x \ln x$, $\mathrm{Dom}\, f = [0, \infty)$ ($0 \ln 0 = 0$ by definition).

Solution: To maximize $xy - x \ln x$ over $x \geq 0$ we can use the Fermat rule resulting in the equation $y - 1 - \ln x = 0$. Thus, the maximizer is $x = \mathrm{e}^{y-1}$, resulting in

$$f^*(y) = \mathrm{e}^{y-1}, \ \mathrm{Dom}\, f^* = \mathbf{R}.$$

We could get the same result without computation: from item 2 we know that the Legendre transform of e^x is $y \ln y - y$, implying that the Legendre transform of $x \ln x - x$ is e^y; and linear perturbation of a function (in our case, adding x to $x \ln x - x$) results in shift of the Legendre transform.

4. $f(x) = x^p/p$, $\mathrm{Dom}\, f = [0, \infty)$; here $p > 1$.

Solution: $f^*(y) = \sup_{x \geq 0}[\phi_y(x) := xy - x^p/p]$. When $y \leq 0$, we clearly have $\sup_{x \geq 0} \phi_y(x) = \phi_y(0) = 0$. When $y > 0$, the maximizer of $\phi_y(x)$ over $x \geq 0$ is given by Fermat rule resulting in the equation $y = x^{p-1}$. Thus, for $y > 0$ we have $\sup_{x \geq 0} \phi_y(x) = y^{1+\frac{1}{p-1}} - y^{\frac{p}{p-1}}/p = y^q/q$, where $q = \frac{p}{p-1}$, or, which is the same, $\frac{1}{p} + \frac{1}{q} = 1$. We end up with

$$f^*(y) = [y_+]^q/q, \ \mathrm{Dom}\, f^* = \mathbf{R}; \ \text{here } y_+ = \max[y, 0], \ q = \tfrac{p}{p-1}.$$

Exercise III.21. Compute Legendre transforms of the following functions:

- [log-barrier for nonnegative orthant $\mathbf{R}^n_+$] $f(x) = -\sum_{i=1}^n \ln x_i : \mathrm{int}\, \mathbf{R}^n_+ \to \mathbf{R}$

Solution:

$$f^*(z) = \sup_{x>0} \sum_i [z_i x_i + \ln x_i] = \begin{cases} -n - \sum_i \ln(-z_i), & z < 0 \\ +\infty & \text{, otherwise} \end{cases},$$

thus, $f^*(z) = f(-z) - n$.

- [log-det barrier for semidefinite cone $\mathbf{S}^n_+$] $f(x) = -\ln \mathrm{Det}(x) : \mathrm{int}\, \mathbf{S}^n_+ \to \mathbf{R}$ (start with proving convexity of f).

Solution: Convexity of f was already established twice – first time via computing second order directional derivative in section C.2.2, second time in chapter 14. We have

$$f^*(z) = \sup_{x \succ 0} \left[\mathrm{Tr}(zx) + \ln \mathrm{Det}(x)\right].$$

It is immediately seen that $f^*(z) = +\infty$ unless $z \in -\mathrm{int}\, \mathbf{S}^n_+$. Indeed, restricting maximization over $x \succ 0$

by maximization over $x \succ 0$ commuting with z and looking what happens when x and z are represented in the orthonormal eigenbasis of z, we get

$$f^*(z) \geq \sup_{\xi>0} \left[\sum\nolimits_i \xi_i\zeta_i + \sum\nolimits_i \ln \xi_i\right],$$

where ζ_i are the eigenvalues of z, and from item 1 we know that the right hand side sup is $+\infty$ unless all ξ_i are negative. Now let $z \prec 0$. In this case we can maximize the concave function $\mathrm{Tr}(zx) - f(x) = \mathrm{Tr}(zx) + \ln \mathrm{Det}(x)$ over $x \succ 0$ by solving the Fermat equation; as we know from Example C.9 in section C.1.6, $\nabla f(x) = -x^{-1}$, so that the Fermat rule results in $x = -z$ and $f^*(z) = -\ln \mathrm{Det}(-z) - n = f(-z) - n$.

Exercise III.22. [computing the Legendre transform of the log-barrier $-\ln(x_n^2 - x_1^2 - ... - x_{n-1}^2)$ for Lorentz cone] Consider the optimization problem

$$\max_{x,t} \left\{\xi^\top x + \tau t + \ln(t^2 - x^\top x) : \ (t,x) \in X = \{(t,x) : t > \sqrt{x^\top x}\}\right\},$$

where $\xi \in \mathbf{R}^n$, $\tau \in \mathbf{R}$ are parameters. Is the problem convex[9]? For what values of ξ, τ this problem is solvable? What is the optimal value? Is it convex in the parameters?

Solution: Problem is convex, since the function $f(t,x) = -\ln(t^2 - x^\top x)$ is convex (direct computation of the second order directional derivative[10]); the domain of the problem is open. Therefore the problem is solvable if and only if the Fermat system

$$\begin{array}{ll} \xi - f'_x(t,x) = 0 & \Longleftrightarrow \ \frac{2x}{t^2 - x^\top x} = \xi \\ \tau - f'_\tau(t,x) = 0 & \Longleftrightarrow \ -\frac{2t}{t^2 - x^\top x} = \tau \end{array} \qquad (*)$$

in variables t, x has a solution with $t > \sqrt{x^\top x}$; it follows that τ should be negative. Assuming that it is the case, the second equation says that $\frac{1}{t^2 - x^\top x} = -\frac{\tau}{2t}$, whence the first equation says that $x = -\frac{t}{\tau}\xi$. It follows that

$$-\frac{2t}{\tau} = t^2 - x^\top x = t^2 - \frac{t^2}{\tau^2}\xi^\top\xi = \frac{t^2}{\tau^2}(\tau^2 - \xi^\top\xi). \qquad (1)$$

In order for this equation be solvable one should have $\tau^2 > \xi^\top\xi$, which combines with $\tau < 0$ to yield that $-\tau > \sqrt{\xi^\top\xi}$. Under the latter assumption, (1) implies that

$$t = -\frac{2\tau}{\tau^2 - \xi^\top\xi}, \qquad (2)$$

whence also

$$x = \frac{2\xi}{\tau^2 - \xi^\top\xi} \qquad (3)$$

Thus, the space of parameters for which the problem is solvable is given by

$$-\tau > \sqrt{\xi^\top\xi},$$

the solution is given by (2) - (3), and the optimal value is (direct computation)

$$-\ln(\tau^2 - \xi^\top\xi) + 2\ln 2 - 2.$$

[9] A *maximization* problem with objective $f(\cdot)$ and certain constraints and domain is called convex if the equivalent minimization problem with the objective $(-f)$ and the original constraints and domain is convex.

[10] intelligent reasoning: in the domain $t > \sqrt{x^\top x}$ we have $f(t,x) = -\ln t - \ln(t - t^{-1}x^\top x) = -\ln t + g(t^{-1}x^\top x - t)$, where the function $g(s) = \begin{cases} -\ln(-s), & s < 0 \\ +\infty, & s \geq 0 \end{cases}$ is convex and nondecreasing. The function $t^{-1}x^\top x - t$ is convex in the domain $t > 0$ as the perspective transform of $x^\top x - 1$. Now convexity of f is readily given by calculus of convexity-preserving operations.

The optimal value is convex in the parameters τ, ξ (by its origin, it is supremum of linear forms, parameterized by x, t, of the parameters τ, ξ).

Exercise III.23. Consider the optimization problem

$$\max_{x,y} \{f(x,y) = ax + by + \ln(\ln y - x) + \ln(y) : \ (x,y) \in X = \{(x,y) : y > \exp\{x\}\}\},$$

where $a, b \in \mathbf{R}$ are parameters. Is the problem convex? For what values of a, b this problem is solvable? What is the optimal value? Is it convex in the parameters?

Solution: The objective is concave (direct computation), the domain is convex, so that the problem is convex; the domain of the problem is open. Therefore a, b correspond to a solvable problem if and only if the Fermat system

$$\begin{array}{ll} f'_x(x,y) = 0 & \Longleftrightarrow \ a = \frac{1}{\ln y - x} \\ f'_y(x,y) = 0 & \Longleftrightarrow \ b = -\frac{1}{y}\left[1 + \frac{1}{\ln y - x}\right] \end{array} \tag{4}$$

in variables x, y has a solution with $y > 0$, $\ln y > x$. From the first equation, a should be positive, and if this is the case, the second equation says that b should be negative and $y = -\frac{1+a}{b}$. Thus, a should be positive, b should be negative, and in this case the solution to (4) is

$$x = \ln\left(-\frac{1+a}{b}\right) - \frac{1}{a}, \ y = -\frac{1+a}{b},$$

whence the optimal value is

$$(a+1)\ln\left(-\frac{1+a}{b}\right) - \ln a - a - 2.$$

This quantity, due to its origin, is supremum of linear forms of a, b and therefore is convex in the domain $a > 0, b < 0$.

Exercise III.24. Compute Legendre transforms of the following functions:

- ["geometric mean"] $f(x) = -\prod_{i \le n} x_i^{\pi_i} : \mathbf{R}^n_+ \to \mathbf{R}$, where $\pi_i > 0$ sum up to 1 and $n > 1$.

  Solution: Convexity of f was established in Example III.10.5. The Legendre transform is

  $$f^*(y) = \sup_{x \ge 0}\{\sum_i y_i x_i + \prod_i x_i^{\pi_i}\} \tag{$*$}$$

  The right hand side is $+\infty$ unless $y < 0$ (assuming that, say, $y_1 \ge 0$, look what happens when x runs through the ray $\{[t; 1; ...; 1] : t \ge 0\}$). Assuming $y < 0$ and setting $z = -y$, we have $f^*(-z) = \sup_{x \ge 0}\{\prod_i x_i^{\pi_i} - \sum_i z_i x_i\}$. What we are maximizing over x, is a homogeneous, of homogeneity degree 1, function of $x \ge 0$ (recall that $\sum_i \pi_i = 1$); therefore the supremum is either 0 or $+\infty$, depending on whether what we are maximizing is or is not nonpositive on $\mathbf{R}^n_+$, or, which is the same, is or is not nonpositive on the set $X_z = \{x \ge 0 : \sum_i z_i x_i = 1\}$. Making educated guess that the maximizer x_z of $\prod_i x_i^{\pi_i}$ over X_z is positive, Karush-Kuhn-Tucker optimality conditions (see discussion after Proposition III.11.3) as applied to our maximization problem (rewritten as $\min_{x \in X_z}[\sum_i z_i x_i + f(x)]$) result in the system

  $$\pi_i \underbrace{[\prod_j x_j^{\pi_j}]}_{\alpha} x_i^{-1} = \lambda z_i, \ i \le n, \sum_i z_i x_i = 1$$

  in variables x, λ; x-component of a solution to this system, if positive, is the desired x_z by Proposition III.11.3. From the system, $\sum_i z_i x_i = \lambda^{-1}\alpha$, that is, $\lambda = \alpha$ and therefore $x_i = \pi_i / z_i$. The vector $[\pi_1/z_1; ...; \pi_n/z_n]$ indeed is positive and is therefore the desired maximizer x_z of ϕ over X_z the maximum being $\prod_i [\pi_i/z_i]^{\pi_i} - 1$. As we remember, $f^*(-z)$ is $+\infty$ when this maximum is positive and is zero otherwise. The bottom line is that the domain of f^* is $\{y \in \mathbf{R}^n : y < 0, \prod_i [-\pi_i/y_i]^{\pi_i} \le 1\}$ and in this domain f^* is identically equal to 0.

  ‘

- ["inverse geometric mean"] $f(x) = \prod_{i\le n} x_i^{-\pi_i} : \operatorname{int}\mathbf{R}^n_+ \to \mathbf{R}$, where $\pi_i > 0$.

  Solution: Convexity of f is stated in Example III.10.6. We have $f^*(y) = \sup_{x>0}[\sum_i y_i x_i - \prod_i x_i^{-\pi_i}]$. This supremum is $+\infty$ when some of y_i are positive (look what happens when, say, $y_1 > 0$, $x_2 = x_3 = ... = x_n = 1$ and $x_1 \to \infty$). Assuming $y \le 0$, a necessary and sufficient condition for $x > 0$ to maximize the concave function $\phi(x) = \sum_i y_i x_i - \underbrace{\prod_i x_i^{-\pi_i}}_{\psi(x)}$ on its domain $\operatorname{int}\mathbf{R}^n_+$ is to solve the Fermat equation $\nabla\phi(x) = 0$, that is, to satisfy

  $$\pi_i \psi(x) x_i^{-1} = -y_i,\ i \le n,$$

  resulting in

  $$f^*(y) = \begin{cases} -(1+\sum_i \pi_i)\left[\prod_i [-y_i/\pi_i]^{\pi_i}\right]^{\frac{1}{1+\sum_i \pi_i}}, & y \le 0 \\ +\infty, & \text{otherwise} \end{cases}.$$

Exercise III.25. Prove the following version of the results of section 13.2:

> Suppose that $f : \mathbf{R}^n \to \mathbf{R} \cup \{+\infty\}$ is a proper convex lsc function with open domain G, and f is twice continuously differentiable, with positive definite Hessian, on G. Assume also that
> (!) Whenever y is such that the function $y^\top x - f(x)$ is, as a function of x, bounded from above, the function achieves its maximum over x.
> Prove that the domain G_* of the Legendre transform f^* of f is an open convex set, f^* is twice continuously differentiable, with positive definite Hessian, on G_*, and the mappings $x \mapsto \nabla f(x)$, $y \mapsto \nabla f^*(y)$ are inverse to each other one-to-one correspondences between G and G_*.

Hint: Use Implicit Function Theorem (Theorem IV.21.5).

Solution: When $\bar{x} \in G$, the gradients, taken at $\bar{x}$, of the functions $\phi_p(x) = \frac{\partial f(x)}{\partial x_p}$, $1 \le p \le n$, are linearly independent (as these gradients are the columns of the Hessian, taken at $\bar{x}$, of f, and this Hessian is positive definite and thus is nonsingular). Applying the Implicit Function Theorem, we conclude that the image G_* of G under the mapping $x \mapsto \nabla f(x)$ is open, and the mapping establishes one-to-one correspondence, with a continuously differentiable inverse, between an open neighbourhood of $\bar{x}$ and an open neighbourhood of $\nabla f(\bar{x})$. Besides this, the mapping $x \mapsto \nabla f(x)$ maps different points of G into different points of G_*; indeed, assuming $\nabla f(x) = \nabla f(x') =: d$ with $x, x' \in G$, both x and x' are, by the Fermat rule, minimizers of the strictly convex function $f(x) - d^\top x$, whence $x = x'$ as the minimizer of a strictly convex function is unique. Next, by (!) combined with the Fermat rule, G_* is exactly the set $\operatorname{Dom} f^*$ of those d for which the function $d^\top x - f(x)$ is bounded from above. Thus, the mapping $x \mapsto \nabla f(x)$ establishes one-to-one continuously differentiable correspondence, with continuously differential inverse $d \to g(d)$, between G and the open set $\operatorname{Dom} f^* = G_*$. By item C in section 13.2, $\nabla f(g(y)) = y$ implies that $g(y) \in \partial f^*(y)$. Thus, f^* is a convex lsc function with open domain allowing for a continuously differentiable selection $y \mapsto g(y)$ of subgradients, whence f^* is differentiable with $\nabla f(y) = g(y)$. Indeed, for $y \in G_*$, by convexity of f, for small $t > 0$ we have $h^\top g(y+th) \ge Df^*(y)[h] \ge h^\top g(y-th)$; passing to limit as $t \to +0$, we get $Df^*(y)[h] \equiv h^\top g(y)$. Thus, first order partial derivatives of f^* – they are just the entries of $g(\cdot)$ and are continuous on G_*, whence f^* is continuously differentiable on the open convex domain G_* with continuously differentiable gradient $\nabla f^*(\cdot) \equiv g(\cdot)$, which is the inverse of the mapping $x \mapsto \nabla f(x) : G \to G_*$.

Exercise III.26. The goal of this exercise is to investigate the relation between smoothess of a convex function and strong convexity of its Legendre transform.
Just for starters:

1. Let $\|\cdot\|$ be a norm on $\mathbf{R}^n$. Recall that unit ball of the conjugate norm $\|y\|_* = \max_x\{y^\top x : \|x\| \le 1\}$ is the polar of the unit ball of $\|\cdot\|$, and the norm conjugate to $\|\cdot\|_*$ is the norm $\|\cdot\|$ itself. Your task is to prove that the functions $\frac{1}{2}\|x\|^2$ and $\frac{1}{2}\|d\|_*^2$ are Legendre transforms of each other.

Solution: We have

$$\sup_x\{d^\top x - \frac{1}{2}\|x\|^2\} = \sup_t \sup_x\{td^\top f - \frac{1}{2}\|x\|^2 t^2\} = \sup_{x\neq 0} \frac{1}{2}\frac{[d^\top x]^2}{\|x\|^2} = \frac{1}{2}\max_x^2\{d^\top x : \|x\| \leq 1\} = \frac{1}{2}\|d\|_*^2.$$

The subsequent tasks need certain preamble.
Smooth convex functions. Let f be a convex function, $\|\cdot\|$ be a norm on $\mathbf{R}^n$, and L be a nonnegative real. We say that f *is* $(L, \|\cdot\|)$*-smooth*, if $\mathrm{Dom}\, f = \mathbf{R}^n$ and

$$\forall(x, z \in \mathbf{R}^n, e \in \partial f(x)) : f(z) \leq f(x) + e^\top[z - x] + \frac{L}{2}\|z - x\|^2.$$

It is easily seen that *a convex function $f : \mathbf{R}^n \to \mathbf{R}$ is $(L, \|\cdot\|)$-smooth if and only if it is continuously differentiable, and the mapping $x \mapsto \nabla f(x)$ is Lipschitz continuous, with constant L, from the norm $\|\cdot\|$ to the norm $\|\cdot\|_*$:*

$$\|\nabla f(x) - \nabla f(y)\|_* \leq L\|x - y\| \ \forall x, y$$

same as *if and only if f is continuously differentiable and*

$$[x - y]^\top[\nabla f(x) - \nabla f(y)] \leq L\|x - y\|^2 \ \forall x, y.$$

A twice continuously differentiable function $f : \mathbf{R}^n \to \mathbf{R}$ is $(L, \|\cdot\|)$-smooth if and only if $0 \leq \frac{d^2}{dt^2}\big|_{t=0} f(x + th) \leq L\|h\|^2$ for all x, h.
Example: Convex quadratic function $f = \frac{1}{2}x^\top Qx - q^\top x + c$ ($Q \succeq 0$) is $(L, \|\cdot\|_2)$-smooth whenever the eigenvalues of Q are upper-bounded by L.
Strongly convex functions. Let $g : \mathbf{R}^n \to \mathbf{R} \cup \{+\infty\}$ be a proper convex lsc function, $\|\cdot\|_*$ be the conjugate of a norm $\|\cdot\|$, and L be a positive real. We say that g is $(L, \|\cdot\|_*)$*-strongly convex*, if for every $\bar{y} \in \mathrm{Dom}\, g$ it holds

$$\forall(y \in \mathbf{R}^n, e \in \partial g(\bar{y})) : g(y) \geq g(\bar{y}) + [y - \bar{y}]^\top e + \frac{1}{2L}\|y - \bar{y}\|_*^2.$$

It can be proved that *a proper convex lsc function g is $(L, \|\cdot\|_*)$-strongly convex if and only if*

$$[e' - e]^\top[y' - y] \geq \frac{1}{L}\|y' - y\|_*^2 \ \forall(y, y' \in \mathrm{Dom}\, g, e \in \partial g(y), e' \in \partial g(y'))$$

A twice continuously differentiable on $\mathrm{rint}\,\mathrm{Dom}\, g$ convex lsc function g is $(L, \|\cdot\|_*)$-strongly convex if and only if $\frac{d^2}{dt^2}\big|_{t=0} g(y + th) \geq L^{-1}\|h\|_*^2$ for all $y \in \mathrm{rint}\,\mathrm{Dom}\, g$ and all h from the linear subspace parallel to $\mathrm{Aff}(\mathrm{Dom}\, g)$.
Example: Convex quadratic form $f(x) = \frac{1}{2}x^\top Qx - q^\top x + c : \mathbf{R}^n \to \mathbf{R}$ is $(L, \|\cdot\|_*)$-strongly convex if and only if $Q \succ 0$ and all eigenvalues of Q are lower-bounded by L^{-1}.
Note: When f is convex quadratic form with the matrix of the quadratic part equal to $Q \succ 0$, the Legendre transform f^* of f is convex quadratic form with the matrix of the quadratic part equal to Q^{-1}.
From the examples above, *a quadratic form with positive definite matrix of the quadratic part, the form is $(L, \|\cdot\|_2)$-smooth if and only if the Legendre transform f^* of f is $(L, \|\cdot\|_2)$-strongly convex.*

The point of the exercise is to justify the following far-reaching extension of the latter observation:

> The Legendre transform f^* of an $(L, \|\cdot\|)$-smooth convex function $f : \mathbf{R}^n \to \mathbf{R}$ is $(L, \|\cdot\|_*)$-strongly convex. Vice versa, if the Legendre transform f^* of a proper convex lsc function $f : \mathbf{R}^n \to \mathbf{R} \cup \{+\infty\}$ is $(L, \|\cdot\|_*)$-strongly convex, then f is $(L, \|\cdot\|)$-smooth.

2. Justify the above claim.

Solution: Justifying the first claim: Let $f : \mathbf{R}^n \to \mathbf{R}$ be $(L, \|\cdot\|)$-smooth convex function, and let us prove that f^* is $(L, \|\cdot\|_*)$-strongly convex. Let $\bar{d} \in \mathrm{Dom}\, f^*$ be such that $\partial f^*(\bar{d}) \neq \varnothing$, and let $\bar{x} \in \partial f^*(\bar{d})$. By item C in section 13.2, $\bar{x}$ is a maximizer over $x \in \mathbf{R}^n$ of the function $\bar{d}^\top x - f(x)$, whence $f^*(\bar{d}) = \bar{d}^\top\bar{x} - f(\bar{x})$. Besides this, as $\bar{x} \in \partial f^*(\bar{d})$, $\bar{d}$ is a maximizer of $d^\top\bar{x} - f^*(d)$ over d, whence

$\bar{d} \in \partial f(\bar{x})$ by the same item C in section 13.2. Consequently, $f(x) \le f(\bar{x}) + \bar{d}^\top[x - \bar{x}] + \frac{L}{2}\|x - \bar{x}\|^2$ (as f is $(L, \|\cdot\|)$-smooth), whence

$$\begin{array}{l} f_*(d) = \sup_x[d^\top x - f(x)] \ge \sup_x\{d^\top x - f(\bar{x}) - \bar{d}^\top[x - \bar{x}] - \frac{L}{2}\|x - \bar{x}\|^2\} \\ = \sup_x\{[d - \bar{d}]^\top[x - \bar{x}] - \frac{L}{2}\|x - \bar{x}\|^2\} + d^\top\bar{x} - f(\bar{x}) = \frac{1}{2L}\|d - \bar{d}\|_*^2 + d^\top\bar{x} - f(\bar{x}) \\ = \frac{1}{2L}\|d - \bar{d}\|_*^2 + [d - \bar{d}]^\top\bar{x} + [\bar{d}^\top\bar{x} - f(\bar{x})] = f^*(\bar{d}) + [d - \bar{d}]^\top\bar{x} + \frac{1}{2L}\|d - \bar{d}\|_*^2 \end{array}$$

(we have used the fact that the Legendre transform of $\frac{L}{2}\|\cdot\|^2$ is $\frac{1}{2L}\|\cdot\|_*^2$). Thus,

$$f^*(d) \ge \ f^*(\bar{d}) + \bar{x}^\top[d - \bar{d}] + \frac{1}{2L}\|d - \bar{d}\|^2$$

whenever $\bar{x} \in \mathrm{Argmax}_x[\bar{y}^\top x - f(x)] = \partial f_*(\bar{y})$ (the latter relation is again given by item C in section 13.2). Thus, f^* is $(L, \|\cdot\|_*)$-strongly convex.

Justifying the second claim: Let f^* be $(L, \|\cdot\|_*)$-strongly convex, and let us prove that f is $(L, \|\cdot\|)$-smooth. Let us prove, first, that $\mathrm{Dom}\, f = \mathbf{R}^n$. Indeed, as f is proper convex lsc, it is the Legendre transform of f^*, Selecting $\bar{d} \in \mathrm{rint}\,\mathrm{Dom}\, f^*$ and $\bar{e} \in \partial f^*(\bar{d})$, we have

$$f^*(d) \ge f^*(\bar{d}) + \bar{e}^\top[d - \bar{d}] + \frac{1}{2L}\|d - \bar{d}\|_*^2,$$

implying that for every x the function $x^\top d - f^*(d)$ is bounded from above, and thus $\mathrm{Dom}(f_*)_* = \mathrm{Dom}\, f = \mathbf{R}^n$. Now let $\bar{x} \in \mathbf{R}^n$ and $\bar{d} \in \partial f(\bar{x})$. Then $\bar{x}$ is a maximizer of $\bar{d}^\top x - f(x)$ over x, whence $\bar{d} \in \mathrm{Dom}\, f^*$, $f(\bar{x}) = \bar{d}^\top\bar{x} - f^*(\bar{d})$, and $\bar{x} \in \partial f^*(\bar{d})$ by item C in section 13.2. As f^* is strongly convex and $\bar{x} \in \partial f^*(\bar{d})$, we have $f^*(d) \ge f^*(\bar{d}) + \bar{x}^\top[d - \bar{d}]^\top + \frac{1}{2L}\|d - \bar{d}\|_*^2$ for every d. Therefore, as f is the Legendre transform of f^*, we have

$$\begin{array}{l} f(x) = \sup_d[d^\top x - f^*(d)] \le \sup_d\left\{d^\top x - f^*(\bar{d}) - \bar{x}^\top[d - \bar{d}] - \frac{1}{2L}\|d - \bar{d}\|_*^2\right\} \\ = \sup_d\left\{[d - \bar{d}]^\top[x - \bar{x}] - \frac{1}{2L}\|d - \bar{d}\|_*^2\right\} + \bar{d}^\top[x - \bar{x}] + [\bar{d}^\top\bar{x} - f^*(\bar{d})] \\ = \frac{L}{2}\|x - \bar{x}\|^2 + [x - \bar{x}]^\top\bar{d} + f(\bar{x}). \end{array}$$

Thus,

$$f(x) \le f(\bar{x}) + [x - \bar{x}]^\top\bar{d} + \frac{L}{2}\|x - \bar{x}\|^2.$$

The resulting inequality holds true for every $x \in \mathbf{R}^n$ and every $\bar{d} \in \partial f(\bar{x})$, implying that f is $(L, \|\cdot\|)$-smooth.

Your concluding task is as follows:

3. Verify that the function $f(x) = \ln(\sum_{i=1}^n \mathrm{e}^{x_i})$ is $(1, \|\cdot\|_\infty)$-smooth, compute its Legendre transform f^*, and make conclusions about strong convexity of f^* (the latter plays important role in the design of *proximal First Order algorithms* for minimization of convex functions over the probabilistic simplex).

Solution: Direct computation shows that setting $p_i = \mathrm{e}^{x_i}/(\sum_j \mathrm{e}^{x_j})$, so that $p_i > 0$ and $\sum_i p_i = 1$, we have $\frac{d^2}{dt^2}\big|_{t=0} f(x + th) = \sum_i p_i h_i^2 - (\sum_i p_i h_i)^2$; convexity of f was established in Example III.10.4. We see that $\frac{d^2}{dt^2}\big|_{t=0} f(x + th) \le \sum_i p_i h_i^2 \le \max_i h_i^2 = \|h\|_\infty^2$, implying that f is $(1, \|\cdot\|_\infty)$-smooth. To compute the Legendre transform f^* of f, note that when $d_i > 0$ and $\sum_i d_i = 1$, the solution to the problem $\max_x[d^T x - f(x)]$ can be found by Fermat rule and is $x_i = \ln(d_i)$; thus, $\mathrm{Dom}\, f^*$ contains the relative interior $\{d > 0, \sum_i d_i = 1\}$ of the probabilistic simplex $\Delta = \{d \in \mathbf{R}^n : d \ge 0, \sum_i d_i = 1\}$, and $f^*(d) = \sum_i d_i \ln(d_i)$ on $\mathrm{rint}\,\Delta$. A natural guess is that $\mathrm{Dom}\, f^* = \Delta$ and $f^* = \sum_i d_i \ln(d_i)$ everywhere on Δ. To verify this guess, note that when $d \in \mathrm{rbd}\,\Delta$, so that the entries in d are nonnegative, sum up to 1, and some of d_i are zeros, the evident way to maximize $\sum_i d_i x_i - \ln(\sum_i \mathrm{e}^{x_i})$ is to push x_i such that $d_i = 0$ to $-\infty$, which reduces the maximization to maximizing $\sum_{i \in I} d_i x_i - \ln(\sum_{i \in I} \mathrm{e}^{x_i})$, with $I = \{i : d_i > 0\}$. This problem we have already solved, and we get $f^*(d) = \sum_i d_i \ln(d_i)$ everywhere on Δ (here, as always, $0\ln(0)$ is set to $0 = \lim_{s \to +0} s \ln s$). It remains to verify that $f^* = +\infty$ outside of Δ. Indeed, when d is not nonnegative, say, $d_1 < 0$, the function $f_d(x) = \sum_i d_i x_i - \ln(\sum_i \mathrm{e}^{x_i})$ is not bounded from above (look what happens when $x_1 \to -\infty$ and $x_i = 0$, $i \ge 2$). When d is nonnegative and $\sum_i d_i \ne 1$, we clearly have $f_d([s; ...; s]) \to \infty$ as $s \to \infty$ when $\sum_i d_i > 1$, and as $s \to -\infty$ when

$\sum_i d_i < 1$. The bottom line is that $f^*(d) = \sum_i d_i \ln(d_i) : \Delta \to \mathbf{R}$, and this function is $(1, \|\cdot\|_1)$-strongly convex.

15.5 Miscellaneous exercises

Exercise III.27. [multi-factor Hölder inequality]

Given positive reals $q_1, ..., q_n$ and $p \in [1, \infty)$, we define the weighted p-norm of a vector $x \in \mathbf{R}^n$ as

$$|x|_p = \left(\sum_{j=1}^n q_j |x_j|^p \right)^{1/p}$$

This clearly is a norm which becomes the standard norm $\|\cdot\|_p$ when $q_j = 1$, $j \le n$. Same as $\|x\|_p$, the quantity $|x|_p$ has limit, namely, $\|x\|_\infty$, as $p \to \infty$, and we define $|\cdot|_\infty$ as this limit.

Now let p_i, $i \le k$, be positive reals such that

$$\sum_{i=1}^k \frac{1}{p_i} = 1.$$

1. Prove that for nonnegative reals $a_1, ..., a_k$ one has

$$a_1 a_2 ... a_k \le \frac{a_1^{p_1}}{p_1} + \ldots + \frac{a_k^{p_k}}{p_k}$$

or, equivalently (set $b_i = a_i^{p_i}$)

$$\forall b \ge 0 : b_1^{1/p_1} b_2^{1/p_2} ... b_k^{1/p_k} \le \frac{b_1}{p_1} + \frac{b_2}{p_2} + ... + \frac{b_k}{p_k}.$$

Note: the special case $p_i = k$, $i \le k$, of this inequality is the inequality between the geometric and the arithmetic means.

Solution: Set $\lambda_i = 1/p_i$, so that $\lambda_i \ge 0$ and $\sum_i \lambda_i = 1$. The claim is evident if some of $a_i \ge 0$ are equal to 0. Assuming $a_i > 0$ for all i and taking into account that $\ln(s)$ is concave on the positive ray, we have

$$\sum_i \lambda_i \ln(a_i^{p_i}) \le \ln \left(\sum_i \lambda_i a_i^{p_i} \right)$$

whence, taking the exponents of both sides and recalling what λ_i are,

$$a_1 ... a_k \le \sum_i \frac{a_i^{p_i}}{p_i}.$$ ■

2. Let $x^1, ..., x^k \in \mathbf{R}^n$, and let $x^1 x^2 \ldots x^k$ be the entrywise product of $x^1, ..., x^k$:

$$[x^1 x^2 \ldots x^k]_j = x_j^1 x_j^2 \cdots x_j^k,\ 1 \le j \le n.$$

Prove that

$$|x^1 x^2 \ldots x^k|_1 \le \sum_{i=1}^k \frac{|x_i^i|^{p_i}}{p_i}. \qquad (*)$$

Solution: Set $x = x^1 x^2 \ldots x^k$ and $y_j^i = q_j^{1/p_i} |x_j^i|$, so that $\prod_{i=1}^k y_j^i = q_j \prod_{i=1}^k |x_j^i|$ due to $\sum_i 1/p_i = 1$. We have

$$\begin{array}{rcl} q_j |x_j| & = & q_j \prod_i |x_j^i| = \prod_i y_j^i \\ & \le & \sum_i [y_j^i]^{p_i}/p_i \text{ [by item 1]} \\ & = & \sum_i q_j |x_j^i|^{p_i}/p_i, \text{ [by definition of } y_j^i] \end{array}$$

that is, $q_j |x_j| \le \sum_{i=1}^k q_j |x_j^i|^{p_i}/p_i$. Summing up these inequalities over $j = 1, ..., n$, we get $(*)$. ■

3. Prove *multi-factor Hölder inequality*: for vectors $x^i \in \mathbf{R}^n$, $i \le k$, one has

$$|x^1 x^2 \dots x^k|_1 \le |x^1|_{p_1} |x^2|_{p_2} \cdots |x^k|_{p_k} \qquad (\#)$$

Solution: $(\#)$ clearly holds true when some of x^i are zero vectors. Assuming $|x^i|_{p_i} > 0$ for all i, observe that both sides in $(\#)$ are positively homogeneous, of degree 1, w.r.t. every one of x^i: when multiplying x^i by t, both sides are multiplied by $|t|$. As a result, to verify $(\#)$ for nonzero x^i is the same as to verify this inequality when $|x^i|_{p_i} = 1$ for all i. But in this case, invoking item 2,

$$|x^1 x^2 \dots x^k|_1 \le \sum_{i=1}^k |x^i|_{p_i}^{p_i}/p_i = \sum_{i=1}^k 1/p_i = 1,$$

exactly as stated by $(\#)$ in the case of $|x^i|_{p_i} = 1$, $i \le k$. ■

Note: we have proved $(\#)$ for positive reals $p_1, ..., p_k$ with $\sum_i 1/p_i = 1$. From the reasoning it is immediately seen that the $(\#)$ remains true when $p_i = \infty$ for some i (and, of course, $1/p_i$ is set to 0 for these i).

Exercise III.28. [Muirhead's inequality]
For any $u \in \mathbf{R}^n$ and $z \in \mathbf{R}^n_{++} := \{z \in \mathbf{R}^n : z > 0\}$ define

$$f_z(u) = \frac{1}{n!} \sum_\sigma z_{\sigma(1)}^{u_1} \cdots z_{\sigma(n)}^{u_n},$$

where the sum is over all permutations σ of $\{1, \dots, n\}$. Show that if P is a doubly stochastic $n \times n$ matrix, then

$$f_z(Pu) \le f_z(u), \quad \forall (u \in \mathbf{R}^n, z \in \mathbf{R}^n_{++}).$$

Solution: For $z \in \mathbf{R}^n_{++}$, $f_z(u)$ clearly is convex and permutation symmetric function of u; it remains to apply Lemma III.14.1.

Exercise III.29. Prove that a convex lsc function f with polyhedral domain is continuous on its domain. Does the conclusion remain true when lifting either one of the assumptions that (a) convex f is lsc, and (b) $\mathrm{Dom}\, f$ is polyhedral?

Solution: We should prove that if $\mathrm{Dom}\, f$ is polyhedral, $x_i \in \mathrm{dom}\, f$ converge to $\bar{x}$ as $i \to \infty$, then $f(x) = \lim_{i\to\infty} f(x_i)$. Passing to a subsequence, it suffices to prove this relation when the sequence $f(x_i)$ has a limit (finite or infinite) as $i \to \infty$. Finally, restricting f from $\mathrm{Dom}\, f$ onto the intersection of $\mathrm{Dom}\, f$ with appropriate box, the situation reduces to the one where $\mathrm{Dom}\, f$ is polyhedral and bounded. Let $V = \mathrm{Ext}(\mathrm{Dom}\, f)$; then V is nonempty finite set: $V = \{v_1, ..., v_N\}$, and $\mathrm{Dom}\, f = \mathrm{Conv}(V)$ (by Krein-Milman Theorem). Since f is lsc, we have $s := \lim_{i\to\infty} f(x_i) \ge f(\bar{x})$. We want to prove that in fact $s = f(\bar{x})$; given that $s \ge f(\bar{x})$, all we need is to lead to a contradiction the assumption that $s > f(\bar{x})$. Assume that $s > f(\bar{x})$; then for some $\delta > 0$ we have $f(x_i) \ge f(\bar{x}) + \delta$ for all but finitely many values of i. Representing x_i as a convex combination $\sum_{j=1}^N \lambda_j^i v_j$ of v_j and passing to a subsequence, we can assume that the N sequences $\{\lambda_j^i\}_{i\ge1}$ have limits λ_j as $i \to \infty$, so that $\lambda_j \ge 0$, $\sum_j \lambda_j = 1$, and $\bar{x} = \lim_{i\to\infty} x_i = \sum_j \lambda_j v_j$, and, in addition, that $f(x_i) \ge f(\bar{x}) + \delta$ for all i. Now let $J = \{j : \lambda_j > 0\}$. For every $\theta > 1$, we have

$$x_i^\theta := \bar{x} + \theta(x_i - \bar{x}) = \sum_{j=1}^N \lambda_{j,\theta}^i v_j, \ \lambda_{j,\theta}^i = \lambda_j + \theta[\lambda_j^i - \lambda_j].$$

Note that $\sum_j \lambda_{j,\theta}^i = 1$ for all i, same as $\lambda_{j,\theta}^i \ge 0$ for all i provided that $j \notin J$. When $j \in J$, we have $\lambda_j > 0$, and therefore $\lambda_{j,\theta}^i \ge 0$ for all large enough values of i, due to $\lambda_j - \lambda_j^i \to 0$, $i \to \infty$. The bottom line is that for fixed θ, all coefficients $\lambda_{j,\theta}^i$, $1 \le j \le N$, are nonnegative for all large enough values of i. Consequently, x_i^θ for large i is a convex combination of v_j and therefore belongs to $\mathrm{Dom}\, f$. For i such that $x_i^\theta \in \mathrm{Dom}\, f$ by convexity of f we have

$$\delta \le f(x_i) - f(\bar{x}) \le \theta^{-1}[f(x_i^\theta) - f(\bar{x})]$$

due to $x_i^\theta - \bar{x} = \theta[x_i - \bar{x}]$. We conclude that for every $\theta > 1$ and all large enough values of i we have $x_i^\theta \in \mathrm{Dom}\, f$ and $f(x_i^\theta) \geq f(\bar{x}) + \theta\delta$. As a result, f is not bounded from above on $\mathrm{Dom}\, f$, which is the desired contradiction, since $\max_{x\in \mathrm{Dom}\, f} f(x) = \max_{j\leq N} f(v_j) < \infty$ by convexity of f combined with $\mathrm{Dom}\, f = \mathrm{Conv}\{v_1, ..., v_N\}$. ■

A convex non-lsc function with polyhedral domain can be discontinuous, e.g., $f(x) = \begin{cases} 1 & , x = 0 \\ 0 & , 0 < x \leq 1 \end{cases}$, $\mathrm{Dom}\, f = [0, 1]$. Similarly, a convex lsc function with non-polyhedral domain, even a closed one, can be discontinuous. To give an example, consider the following construction: we take the convex hull E of the set $\{[x; y; 0] : (x-1)^2 + y^2 \leq 1\} \cup \{[0; 0; -1]\}$ and set $E^+ = \{[x; y; t] : \exists \tau : [x; y; \tau] \in E \;\&\; t \geq \tau\}$. Clearly, E^+ is closed and convex and is the epigraph of some function f with the domain $D = \{[x; y] : (x-1)^2 + y^2 \leq 1\}$. Since $E^+ = \mathrm{epi}\{f\}$ is closed and convex, f is convex lsc. At the same time, the intersection of E^+ and the line $\{[0; 0; t] : t \in \mathbf{R}\}$ is the ray $\{[0; 0; t] : t \geq -1\}$, so that $f(0, 0) = -1$, and the intersection of E^+ and a line $\{[a; b; t] : t \in \mathbf{R}\}$ with $a > 0$, b satisfying $(a-1)^2 + b^2 = 1$ is the ray $\{[a; b; t] : t \geq 0\}$, that is, $f(a, b) = 0$ whenever $[a; b]$ is a boundary point of D distinct from $[0; 0]$. Since the boundary point $[0; 0]$ of D is the limit of a sequence of distinct from it boundary points of D, f is not continuous on D.

Exercise III.30. Let $a_1, ..., a_n > 0$, $\alpha, \beta > 0$. Solve the optimization problem

$$\min_x \left\{ \sum_{i=1}^n \frac{a_i}{x_i^\alpha} : x > 0, \sum_i x_i^\beta \leq 1 \right\}$$

Solution: Passing to variables $y_i = x_i^\beta$, we convert the problem to a *convex* program

$$\min_y \left\{ \sum_i a_i y_i^{-\alpha/\beta} : y > 0, \sum_i y_i \leq 1 \right\}$$

KKT conditions (where we guess that the constraint is active) read

$$\begin{array}{l} -\frac{\alpha}{\beta} a_i y_i^{-\frac{\alpha}{\beta}-1} + \lambda = 0,\ i = 1, ..., n \\ \sum_i y_i = 1 \end{array}$$

whence

$$y_i = \frac{a_i^{\frac{\beta}{\alpha+\beta}}}{\sum_j a_j^{\frac{\beta}{\alpha+\beta}}} \implies x_i = \frac{a_i^{\frac{1}{\alpha+\beta}}}{\left(\sum_j a_j^{\frac{\beta}{\alpha+\beta}}\right)^{1/\beta}}$$

Since the problem in y-variables is convex, the KKT point we have found is a globally optimal solution. The optimal value is

$$\left(\sum_j a_j^{\frac{\beta}{\alpha+\beta}}\right)^{\frac{\alpha+\beta}{\beta}}.$$

Exercise III.31. [computational study] Consider the following situation: there are K "radars" with k-th of them capable to locate targets within ellipsoid $E_k = \{x \in \mathbf{R}^n : (x - c_k)^\top C_k (x - c_k) \leq 1\}$ ($C_k \succ 0$); the measured position of target is

$$y_k = x + \sigma_k \zeta_k,$$

where x is the actual position of the target, and ζ_k is the standard (zero mean, unit covariance) Gaussian observation noise; ζ_k's are independent across k. Given measurements $y_1, ..., y_K$ of target's location x known to belong to the "common field of view" $E = \cap_k E_k$ of the radars, which we

assume to possess a nonempty interior, we want to estimate a given linear form $e^\top x$ of x by using linear estimate

$$\widehat{x} = \sum_k h_k^\top y_k + h.$$

We are interested in finding the estimate (e.g., the parameters $h_1, ..., h_K$, h) minimizing the risk

$$\mathrm{Risk2} = \max_{x\in E} \sqrt{\mathbf{E}\left\{\left[e^\top x - \textstyle\sum_k h_k^\top [x + \sigma_k \zeta_k] - h\right]^2\right\}}$$

1. Pose the problem as convex optimization program

   Solution: We have

   $$\begin{array}{l}\mathrm{Risk2}^2 = \max_{x\in E} \mathbf{E}\left\{\left[\left([e - \sum_k h_k]^\top x - h\right) - \left(\sum_k \sigma_k h_k^\top \zeta_k\right)\right]^2\right\} \\ = \max_{x\in E}\left[\left[e - \sum_k h_k\right]^\top x - h\right]^2 + \sum_k \sigma_k^2 h_k^\top h_k\right] \\ = \sum_k \sigma_k^2 h_k^\top h_k + \max_{x\in E}\left[[e - \sum_k h_k]^\top x - h\right]^2.\end{array}$$

   As a result, denoting by ϕ_E the support function of E, the problem of minimizing $\mathrm{Risk2}^2$ can be posed as convex optimization problem

   $$\mathrm{Opt} = \min_{h_1,...,h_K,h,t}\left\{t^2 + \sum_k \sigma_k^2 h_k^\top h_k : \phi_E(e - \sum_k h_k) \le t + h, \phi_E(\sum_k h_k - e) \le t - h\right\}$$

   By Exercise III.12.4, we have

   $$\phi_E(g) = \min_{g_1,...,g_K}\left\{\sum_k \phi_{E_k}(g_k) : \sum_k g_k = g\right\},$$

   and by Exercise III.14.1,

   $$\phi_{E_k}(g) = \sqrt{g C_k^{-1} g} + g^\top c_k,$$

   so that the problem of interest becomes

   $$\mathrm{Risk2}^2 = \min_{h_k,g_k,f_k,k\le K,h,t}\left\{t^2 + \sum_k \sigma_k^2 h_k^\top h_k : \begin{array}{l}\sum_k [\|C_k^{-1/2} g_k\|_2 + c_k^\top g_k] \le t + h, \\ \sum_k g_k + \sum_k h_k = e \\ \sum_k [\|C_k^{-1/2} f_k\|_2 + c_k^\top f_k] \le t - h, \\ -\sum_k f_k + \sum_k h_k = e\end{array}\right\}$$

2. Process the problem numerically and look at the results.
   Recommended setup:

   - $K = 3$, $n = 2$, $[c_1, c_2, c_3] = \begin{bmatrix} 1.000 & -0.500 & -0.500 \\ 0 & 0.866 & -0.866 \end{bmatrix}$,

     $C_1 = \begin{bmatrix} 0.2500 & 0 \\ 0 & 1.5000 \end{bmatrix}$, $C_2 = \begin{bmatrix} 1.1875 & 0.5413 \\ 0.5413 & 0.5625 \end{bmatrix}$, $C_3 = \begin{bmatrix} 1.1875 & -0.5413 \\ -0.5413 & 0.5625 \end{bmatrix}$

   - $\sigma_1 = 0.1, \sigma_2 = 0.2, \sigma_3 = 0.3$
   - $e = [1; 1]/\sqrt{2}$.

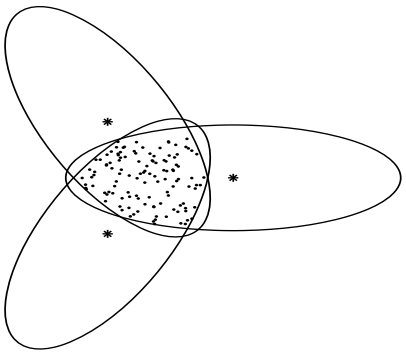


Figure 15.1. 3 radars and their common field of view (dotted)

Solution: Our results are as follows:

- Risk2 $= 0.0852$, $[h_1, h_2, h_3] = \left[\begin{array}{ccc} 0.5130 & 0.1283 & 0.0570 \\ 0.5136 & 0.1284 & 0.0571 \end{array}\right]$, $h = 0.0010$

Exercise III.32. For any $k \leq m$ and $X \in \mathbf{S}^m$, recall that $S_k(X)$ denotes the sum of k largest eigenvalues of the matrix X. Given $X \in \mathbf{S}^m$, define $R[X] := \left\{V^\top X V : V \in \mathcal{O}_m\right\}$ where $\mathcal{O}_m = \{V \in \mathbf{R}^{m\times m} : VV^\top = I_m\}$ is the set of all $m \times m$ orthogonal matrices. Prove that for any two symmetric matrices $X, Y \in \mathbf{S}^m$, we have

$$Y \in \mathrm{Conv}(R[X]) \text{ if and only if } S_k(Y) \leq S_k(X) \text{ for all } k < m \text{ and } \mathrm{Tr}(Y) = \mathrm{Tr}(X).$$

Solution: In one direction: suppose $Y \in \mathrm{Conv}(R[X])$, and let us prove that $S_k(Y) \leq S_k(X)$, $k \leq M$, with $S_m(Y) = S_m(X)$. Observe that a rotation $X \mapsto V^\top X V$, $V \in \mathcal{O}_m$, as every similarity transformation $X \mapsto Z^{-1}XZ$, preserves the vector of eigenvalues. It follows that the linear function $\mathrm{Tr}(Z) = S_m(Z)$ is equal to $S_m(X)$ on the entire $R[X]$ and is therefore equal to the same $S_m(X)$ on $\mathrm{Conv}(R[X])$. The bottom line is that $\mathrm{Tr}(Y) = S_m(Y) = S_m(X) = \mathrm{Tr}(X)$. Next, by the above argument $S_k(Z)$ is identically equal to $S_k(X)$ on the entire $R[X]$, and since $S_k(\cdot)$ is convex (see chapter 14), we conclude that $S_k(\cdot)$ is $\leq S_k(X)$ everywhere on $\mathrm{Conv}(R[X])$, whence $S_k(Y) \leq S_k(X)$, $k \leq m$.
In the opposite direction: let $S_k(Y) \leq S_k(X)$, $k \leq m$, and $S_m(Y) = S_m(X)$. By Majorization Principle, $\lambda(Y) = \pi\lambda(X)$ with doubly stochastic π. By Birkhoff Theorem (Theorem II.7.7), π is a convex combination of permutation matrices P_i: $\pi = \sum_i \alpha_i P_i$ with $\alpha_i \geq 0$ summing up to 1. Consequently, $\lambda(Y) = \sum_i \alpha_i P_i \lambda(X)$, or, which is clearly the same, $\mathrm{Diag}\{\lambda(Y)\} = \sum_i \alpha_i P_i \mathrm{Diag}\{\lambda(X)\} P_i^\top$. Next, $X = U\,\mathrm{Diag}\{\lambda(X)\}U^\top$ and $Y = V\,\mathrm{Diag}\{\lambda(Y)\}V^\top$ with $U, V \in \mathcal{O}_m$. The bottom line is

$$\begin{array}{l} Y = V\,\mathrm{Diag}\{\lambda(Y)\}V^\top = V\left[\sum_i \alpha_i P_i\,\mathrm{Diag}\{\lambda(X)\}P_i^\top\right]V^\top = V\left[\sum_i \alpha_i P_i U^\top X U P_i^\top\right]V^\top \\ = \sum_i \alpha_i \underbrace{VP_iU^\top}_{=:W_i^\top} XW_i \end{array}$$

The matrices W_i are products of matrices form $\mathcal{O}_m$ and thus $W_i \in \mathcal{O}_m$, and we conclude that $Y \in \mathrm{Conv}(R[X])$. ■

Exercises from Part IV

24.1 Around Conic Duality

Exercise IV.1. Given Linear Dynamical System

$$\begin{array}{rcl} x_0 & = & 0 \\ x_{t+1} & = & Ax_t + Bu_t,\ t = 0, 1, \ldots, N-1 \end{array} \tag{LDS}$$

$(A : n \times n, B : n \times m)$ with controls u_t subject to the "energy constraints"

$$\|u_t\|_2 \leq 1,\ 0 \leq t < N, \tag{EN}$$

pose the problem of minimizing $f^\top x_N$ (f is a given vector) as a conic problem on the product of Lorentz cones, write down the conic dual of this problem and answer the following questions:

1. Is the problem essentially strictly feasible?
2. Is the problem bounded?
3. Is the problem solvable?
4. Is the dual problem feasible?
5. Is the dual problem solvable?
6. Are the optimal values equal to each other?
7. What do the optimality conditions say?

Solution: Relation $\|u\|_2 \leq 1$ with $u \in \mathbf{R}^m$ is equivalent to $[u; 1] \in \mathbf{L}^{m+1}$, so that the problem of interest in the conic form reads

$$\min_{x,u,r} \{f^\top x_N : x_0 = 0,\ x_{t+1} = Ax_t + Bu_t,\ 0 \leq t \leq N-1, [u_t; 1] \in \mathbf{L}^{m+1}, 0 \leq t \leq N-1\} \tag{P}$$

To get the dual problem, we denote by $s_t \in \mathbf{R}^n$ the vectors of Lagrange multipliers for the state constraints $x_{t+1} - Ax_t - Bu_t = 0$, by s_{-1} the vector of Lagrange multipliers for the equality constraints $x_0 = 0$ and by $[y_t; z_t] \in \mathbf{L}_*^{m+1} = \mathbf{L}^{m+1}$ – the Lagrange multipliers for the conic constraints. Aggregating constraints of (P) with multipliers as the weights, we get the aggregated constraint

$$s_{-1}^\top x_0 + \sum_{t=0}^{N-1} s_t^\top [x_{t+1} - Ax_t - Bu_t] + \sum_{t=0}^{N-1} [y_t^\top u_t + z_t] \geq 0,$$

or, which is the same, the constraint

$$\begin{array}{l} s_{N-1}^\top x_N + [s_{N-2} - A^\top s_{N-1}]^\top x_{N-1} + [s_{N-3} - A^\top s_{N-2}]^\top x_{N-2} + \ldots + [s_{-1} - A^\top s_0]^\top x_0 \\ +\sum_{t=0}^{N-1} [y_t - B^\top s_t]^\top u_t \geq -\sum_{t=0}^{N-1} z_t \end{array} \tag{$*$}$$

To get the dual problem, we add to the restrictions $\|y_t\|_2 \leq z_t$ (that is, restrictions $[y_t; z_t] \in \mathbf{L}^{m+1}$) the restriction that the left hand side in $(*)$ identically in x's and u's is $f^\top x_N$ and maximize under this restriction the right hand side in $(*)$. Thus, the dual problem is

$$\max_{y_t, z_t, s_t} \left\{ -\sum_{t=0}^{N-1} z_t : \begin{array}{l} s_{N-1} = f, A^\top s_{t+1} = s_t,\ -1 \leq t \leq N-2, \\ y_t = B^\top s_t, 0 \leq t \leq N-1, \|y_t\|_2 \leq z_t, 0 \leq t \leq N-1 \end{array} \right\} \tag{D}$$

An optimal solution to the dual problem is evident:

$$s_t = [A^{N-1-t}]^\top f,\ -1 \leq t \leq N-1, y_t = B^\top [A^{N-1-t}]^\top f,\ z_t = \|B^\top [A^{N-1-t}]^\top f\|_2,$$

the optimal value is

$$-\sum_{t=0}^{N-1}\|B^\top[A^{N-1-t}]^\top f\|_2.$$

The answers to the remaining questions are as follows:

1. Is the problem essentially strictly feasible? – Yes, (P) is essentially strictly feasible, an essentially strictly feasible solution being, e.g. $u_t = 0$, $0 \le t \le N-1$, $x_t = 0$, $0 \le t \le N$
2. Is the problem bounded? – Yes, since the feasible set clearly is bounded
3. Is the problem solvable? – Yes, as every feasible problem with bounded feasible set (this set is automatically closed, and therefore the linear – and thus continuous – objective attains its minimum on this set)
4. Is the dual problem feasible? – Yes, by Conic Duality Theorem (not speaking about the fact that we see feasible solution by naked eyes)
5. Is the dual problem solvable? – Yes, by Conic Duality Theorem (not speaking about the fact that we see the optimal solution by naked eyes)
6. Are the optimal values equal to each other? – Yes, by Conic Duality Theorem
7. What do the optimality conditions say? – They say that at the primal-dual optimum the primal slacks $[u_t; 1]$ are orthogonal to the vectors $[y_t; z_t] = [B^\top[A^{N-1-t}]^\top f; \|B^\top[A^{N-1-t}]^\top f\|_2]$, that is,

   $$\|B^\top[A^{N-1-t}]^\top f\|_2 + u_t^\top B^\top[A^{N-1-t}]^\top f = 0,$$

   which combines with $\|u_t\|_2 \le 1$ and the Cauchy inequality to imply that whenever the vector $e_t = B^\top[A^{N-1-t}]^\top f$ is nonzero, we have $u_t = -e_t/\|e_t\|_2$, and when $e_t = 0$, u_t can be a whatever vector of norm not exceeding 1. Note that we got "closed form" solutions to both (P) and (D).

Exercise IV.2. Consider the conic constraint $Ax - b \in K$ where $K \subset \mathbf{R}^m$ is a regular cone and matrix A is of full column rank (i.e., has linearly independent columns, or, which is the same, has trivial kernel). Suppose that the constraint is feasible. Show that the following properties are all equivalent to each other:

1. the feasible region $\{x \in \mathbf{R}^n : Ax - b \in K\}$ is bounded;
2. $\mathrm{Im}(A) \cap K = \{0\}$, where $\mathrm{Im}(A) := \{Ax : x \in \mathbf{R}^n\}$;
3. the following system of vector inequalities is solvable

   $$A^\top\lambda = 0, \quad \lambda \in \mathrm{int}\, K_*.$$

Using these conclude that the property of whether a conic problem $\min_x\{c^\top x : Ax - b \in K\}$ has a bounded feasible region or not is independent of the choice of b, provided that the problem is feasible.

Solution: 1. $\Longleftrightarrow$ 2.: We are in the case when the feasible set $X = \{x : Ax - b \in K\}$ is nonempty (and clearly is closed). By Fact II.6.18 X is bounded if and only if X has no nonzero recessive directions, that is, if and only if the recessive cone of X (which is $\{h : Ah \in -K\}$ (why?)) is trivial. Since $h \mapsto Ah$ is an embedding, the latter happens if and only if $\mathrm{Im}(A) \cap [-K] = \{0\}$, or, which is the same, if and only if $\mathrm{Im}(A) \cap K = \{0\}$. ■

3. $\Longrightarrow$ 2.: With 3. in force, there exists $\lambda \in \mathrm{int}\, K_*$ such that $A^\top\lambda = 0$. If now $x \in \mathbf{R}^n$ is such that $y = Ax \in K$, we have $\lambda^\top y = [A^\top\lambda]x = 0$, and since $y \in K$ and $\lambda \in \mathrm{int}K_*$, we conclude that $y = 0$. The bottom line is that $\mathrm{Im}(A) \cap K = \{0\}$, that is, 2. takes place. ■

2. $\Longrightarrow$ 3.: Assume, on the contrary to what should be proved, that 2. does take place, and 3. does not. Then the convex nonempty set $\{\lambda : A^\top\lambda = 0\}$ does not intersect $\mathrm{int}\, K_*$, which also is a nonempty convex set, implying, by Separation Theorem, that there exists $y \in \mathbf{R}^m$,$y \ne 0$, such that

$$\sup_{\lambda: A^\top\lambda = 0} y^\top\lambda \le \inf_{u \in \mathrm{int}\, K_*} y^\top u.$$

since the right hand side infimum is finite and K_* is a cone, this infimum is 0, implying that $y \in (\mathrm{int}\, K_*)_* = K$. On the other hand, the supremum in the left hand side is taken over a linear subspace $\mathrm{Ker}A^\top$; it can be finite if and only if $y \in [\mathrm{Ker}A^\top]^\perp = \mathrm{Im}(A)$. Thus, y is a nonzero vector from $\mathrm{Im}(A) \cap K$, which is impossible by 2. ■

Finally, consider a feasible conic constraint $Ax-b\in K$ with tegular cone K. If $\mathrm{Ker}A\neq\{0\}$, the feasible set of this constraint is unbounded independently of what b is, since $\mathrm{Ker}A$ is the recessive subspace of the feasible set, provided the latter is nonempty. And if $\mathrm{Ker}A=\{0\}$, we, as we just have seen, are in the case where 1. is equivalent to 2., and the validity status of 2. is independent of what b is.

Exercise IV.3.

Given a cone K in a Euclidean space E with inner product $\langle\cdot,\cdot\rangle$, we call a pair of elements $x\in K$ and $y\in K_*$ *complementary* if $\langle x,y\rangle=0$.

In this exercise, we will examine complementarity relations for the second-order cones $\mathbf{L}^n$ and the positive semidefinite cone $\mathbf{S}^n_+$.

1. Consider $\mathbf{L}^n:=\left\{x=[\tilde{x};x_n]\in\mathbf{R}^{n-1}\times\mathbf{R}:\ x_n\geq\|\tilde{x}\|_2\right\}$; as we know, this cone is self-dual (Example II.6.9). Prove that $x,s\in\mathbf{L}^n$ satisfy $\langle x,s\rangle=0$ iff $x_n\tilde{s}+s_n\tilde{x}=0$ holds.
2. Consider the space of $n\times n$ symmetric matrices, i.e., $E=\mathbf{S}^n$ equipped with the Frobenius inner product $\langle X,Y\rangle=\mathrm{Tr}(XY)=\sum_{i=1}^n\sum_{j=1}^n X_{ij}Y_{ij}$. Let $K=\mathbf{S}^n_+:=\{X\in\mathbf{S}^n:x^\top Xx\geq 0,\forall x\in\mathbf{R}^n\}$ be the positive semidefinite cone; recall that this cone is self-dual (Example II.6.10). Prove that $X,Y\in\mathbf{S}^n_+$ are complementary, i.e., $\langle X,Y\rangle=0$, if and only if their matrix product is zero, i.e., $XY=YX=0$. In particular, matrices from a complementary pair commute and therefore share a common orthonormal eigenbasis.

Solution:

1. The statement clearly is true when $s_n=0$ or $x_n=0$. Assuming $x_n>0$, $s_n>0$, the relation $[\tilde{x};x_n]^\top[\tilde{s};s_n]=0$ for $[\tilde{x};x_n]$ and $[\tilde{s};s_n]$ from $\mathbf{L}^n$ means that $s_nx_n=-\tilde{s}^\top\tilde{x}$ together with $\|\tilde{x}\|_2\leq x_n$ and $\|\tilde{s}\|_2\leq s_n$, which, by Cauchy inequality (Theorem B.1) may happen if and only if $\tilde{x}=x_ne$ and $\tilde{s}=-s_ne$ for unit vector e, or, which is the same, when $x_n\tilde{s}+s_n\tilde{x}=0$. ■
2. For $X,Y\in\mathbf{S}^n$ we have $[XY]^\top=YX$, whence $XY=0$ if and only if $YX=0$. Clearly when $XY=YX=0$, we have $\mathrm{Tr}(XY)=0$. So we will prove the reverse direction. Assume that $X\succeq 0,Y\succeq 0$ and $\mathrm{Tr}(XY)=0$, and let us prove that $XY=0$. Indeed, we have

$$0=\mathrm{Tr}(XY)=\mathrm{Tr}(X^{1/2}[X^{1/2}Y^{1/2}]Y^{1/2})=\mathrm{Tr}([X^{1/2}Y^{1/2}][X^{1/2}Y^{1/2}]^\top)=\sum_{i,j}[X^{1/2}Y^{1/2}]^2_{ij},$$

whence $X^{1/2}Y^{1/2}=0$, so that $XY=X^{1/2}[X^{1/2}Y^{1/2}]Y^{1/2}=0$. ■

Exercise IV.4. By General Theorem of the Alternative, a system of m scalar linear constraints $Ax\geq b$ in variables $x\in\mathbf{R}^n$ (or, which is the same, the conic inequality $Ax\geq_{\mathbf{R}^m_+}b$) has no solutions if and only if it can be led to contradiction by aggregation: there exist nonnegative weights $\lambda_1,...,\lambda_m$ such that the associated weighted sum $\lambda^\top Ax\geq\lambda^\top b$ of inequalities from the system is a contradictory inequality, that is, $A^\top\lambda=0$ and $b^\top\lambda>0$. For a general conic constraint of the form

$$Ax\geq_{\mathbf{K}}b \tag{I}$$

where $\mathbf{K}\subset\mathbf{R}^m$ is a regular cone, a similar recipe for certifying infeasibility would read

$$\exists\lambda\in\mathbf{K}_*:A^\top\lambda=0\ \text{ and }\ b^\top\lambda>0. \tag{II}$$

The goal of this exercise is to investigate relation between feasibility statuses of (I) and of (II).

Your first task is easy:

1. Prove that if (II) is feasible, then (I) is infeasible.

Solution: Let λ satisfy (II). To prove that (I) has no solutions, assume, on the contrary, that x solves (I). Then $\lambda^\top Ax\geq\lambda^\top b$ due to $\lambda\in\mathbf{K}_*$, which with our λ results in $0\geq b^\top\lambda$, which is not the case due to $\lambda^\top b>0$; this is the desired contradiction.

The rest of your effort is aimed at investigating to which extent item 1 can be inverted: if and when it is true that when (II) has no solutions, then (I) is feasible? General Theorem of the Alternative says that this indeed is the case when $\mathbf{K}$ is the nonnegative orthant $\mathbf{R}^m_+$. In the general case, the situation is different.

2. Let (I) be the univariate conic inequality

$$Ax := [1;0;1]x \geq_{\mathbf{L}^3} b := [0;1;0] \tag{i}$$

where $\mathbf{L}^3$ is the 3D Lorentz cone. Write down the associated system (II) and check that both this system and (i) are infeasible. Conclude from this example that in general, solvability of (II) is only sufficient, but not necessary, condition for infeasibility of (I).

Solution: Recalling what $\mathbf{L}^3$ is, (i) is the scalar inequality $x \geq \sqrt{x^2+1}$; of course, it is infeasible. Now, the associated system (II) reads

$$\lambda_3 \geq \sqrt{\lambda_1^2+\lambda_2^2}, \lambda_1+\lambda_3 = 0, \lambda_2 > 0$$

in real variables $\lambda_1, \lambda_2, \lambda_3$; λ_1 can be immediately eliminated, resulting in clearly infeasible system $\lambda_3 \geq \sqrt{\lambda_3^2+\lambda_2^2}$, $\lambda_2 > 0$. ■

3. Prove that (II) is infeasible if and only if (I) is *nearly feasible*, meaning that for every $\epsilon > 0$ there exists b' such that $\|b'-b\|_2 \leq \epsilon$ and the conic constraint $Ax \geq_{\mathbf{K}} b'$ is feasible. Equivalently: (II) is infeasible if and only if b belongs to the closure $\overline{B}$ of the set $B = A\mathbf{R}^n - \mathbf{K}$ of those right hand side vectors in (I) for which (I) is feasible.

Solution: Taking into account item 1, the claim we want to prove is that (II) has no solutions if and only if $b \in \overline{B}$, or, which is the same,

(!) (II) has a solution iff $b \notin \overline{B}$.

Justification of (!) is immediate. In one direction: if (II) has a solution λ, then $A^\top\lambda = 0$ and $A^\top b' > 0$ for all b' close enough to b, implying, by item 1, that all these close enough to b vectors b' , when treated as the right hand sides in (I), result in infeasible conic constraint, that is, do not belong to B. Thus, there is a neighbourhood of b which does not intersect B, implying that $b \notin \overline{B}$.
In the opposite direction: assume that $b \notin \overline{B}$, and let us prove that (II) is feasible. Since $b \notin \overline{B}$, b is at a positive distance from the nonempty convex set $B = \{z : z = Ax - y, x \in \mathbf{R}^n, y \in \mathbf{K}\}$, implying by the Separation Theorem that $\{b\}$ can be strongly separated from B: for properly selected λ it holds

$$\lambda^\top b > \sup_{z\in B} \lambda^\top z = \sup_{x\in\mathbf{R}^n, y\in\mathbf{K}} \lambda^\top[Ax-y]. \tag{$*$}$$

the concluding supremum here is finite, implying that $A^\top\lambda = 0$ (otherwise we could make $\lambda^\top[Ax-y]$ arbitrarily large by properly selecting x and setting $y = 0$) and $\lambda \in \mathbf{K}_*$ (otherwise we could make $\lambda^\top[Ax-y]$ arbitrarily large by properly selecting $y \in \mathbf{K}$ and setting $x = 0$). Thus, $A^\top\lambda = 0$ and $\lambda \in \mathbf{K}_*$, implying that the supremum in $(*)$ is 0, that is, $\lambda^\top b > 0$; we conclude that λ solves (II), so that the latter system is feasible. ■

Conclusion: *Feasibility of* (II) *is necessary and sufficient for infeasibility of* (I) *if and only if the set $B = A\mathbf{R}^n - \mathbf{K}$ of the right hand sides in the conic constraint* (I) *resulting in constraint's feasibility is closed; in fact, feasibility of* (II) *is necessary and sufficient condition for b not to belong to the closure $\overline{B}$ of B.* Now, when $\mathbf{K} = \{y : Py \geq 0\}$ is a polyhedral cone, e.g., $\mathbf{R}^m_+$, B is polyhedral (since its definition in the case under consideration is its polyhedral representation as well) and therefore is closed, which explains why when the cone $\mathbf{K}$ is polyhedral infeasibility of (II) is equivalent to feasibility of (I). At the same time, when $\mathbf{K}$ is not polyhedral, B can be non-closed, as is the case in example from item 2. Let us look at the geometry of this example. (i) wants of us to find a point in the intersection of the cone $\mathbf{L}^3 = \{x \in \mathbf{R}^3 : x_3 \geq \sqrt{x_1^2+x_2^2}\}$ with the line $\ell = \{[t;-1;t] \in \mathbf{R}^3 : t \in \mathbf{R}\}$. ℓ belongs to the 2D plane $L = \{x \in \mathbf{R}^3 : x_2 = -1\}$, and the intersection of $\mathbf{L}^3$ with this plane is the set $\{[x_1;-1;x_3] : x_3^2 - x_1^2 \geq 1, x_3 \geq 0\}$, or, which is the same, the set $\{[x_1;-1;x_3] : (x_3-x_1)(x_3+x_1) \geq 1, x_3 - x_1 \geq 0\}$; introducing the coordinates $u = x_3 + x_1$, $v = x_3 - x_1$ on the 2D plane L, the intersection of L and $\mathbf{L}^3$ in these coordinates becomes the inner part $H = \{[u;v] : u \geq 1/v, v > 0\}$ of the branch $\Gamma = \{[u;v] : uv = 1, v > 0\}$ of hyperbola. In u,v-coordinates the line ℓ is just the line $v = 0$. Thus, geometrically the situation is as follows: to intersect ℓ and $\mathbf{L}^3$ is the same as to intersect H with the v-axis of the $[u;v]$-plane; the intersection clearly is empty, so that (i) is infeasible. At the same time, our line is an asymptote of Γ, so that the shift $v = \epsilon$ of the line $v = 0$ makes the intersection of the shifted line with H nonempty,

whatever small $\epsilon > 0$ be. The outlined shift of ℓ in our original x-coordinates reduces to passing from $b = [0;1;0]$ to $b_\epsilon = [0;1;-\epsilon]$. The bottom line is that $b \notin B$ and $b \in \overline{B}$, since $b = \lim_{\epsilon \to +0} b_\epsilon$ and $b_\epsilon \in B$.

The result of item 3 attracts our attention to the following question: *What are natural sufficient conditions which guarantee the closedness of the set $A\mathbf{R}^n - \mathbf{K}$?* Here is a simple answer:

4. Prove that when the only common point of the image space $L := \{y \in \mathbf{R}^m : \exists x : y = Ax\}$ of A and of $\mathbf{K}$ is the origin, the set $B := A\mathbf{R}^n - \mathbf{K} = L - \mathbf{K}$ is closed. Prove that the same holds true when the condition $L \cap \mathbf{K} = \{0\}$ is "heavily violated," meaning that $L \cap \operatorname{int} \mathbf{K} \neq \varnothing$.

Solution: Assume that $L \cap \mathbf{K} = \{0\}$, and let us prove that B is closed. Thus, let $b_i = y_i - z_i$ with $y_i \in L$ and $z_i \in \mathbf{K}$, and let $b_i \to b$ as $i \to \infty$; we need to prove that then $b \in B$. Consider two cases: (a) the sequence $\{z_i\}$ is bounded, and (b) the sequence $\{z_i\}$ is unbounded. In the case of (a), passing to a subsequence, we can assume that $z_i \to \overline{z}$ as $i \to \infty$; since $z_i \to \overline{z}$ and $b_i \to b$ as $i \to \infty$, we conclude that $y_i = b_i + z_i \to b + \overline{z} =: \overline{y}$ as $i \to \infty$. As $\mathbf{K}$ is closed and $z_i \in \mathbf{K}$, we have $\overline{z} \in \mathbf{K}$. By its origin, $\overline{y}$ is the limit of a converging sequence of points from L and thus $\overline{y} \in L$. We see that $b = \overline{y} - \overline{z} \in B$, as claimed. In the case of (b), passing to a subsequence, we can assume that $r_i := \|z_i\|_2 \to \infty$ as $i \to \infty$; since $z_i = y_i - b_i$ and the sequence $\{b_i\}$ converges and is therefore bounded, we conclude that $r_i^{-1}\|y_i\|_2 \to 1$ as $i \to \infty$. Passing to a subsequence, we can further assume that the unit vectors $\overline{z}_i := r_i^{-1} z_i$ converge as $i \to \infty$ to some unit vector $\overline{z}$, and the sequence of vectors $\overline{y}_i = r_i^{-1} y_i$ converges as $i \to \infty$ to some vector $\overline{y}$, which also is unit due to $r_i^{-1}\|y_i\|_2 \to 1$, $i \to \infty$. We have

$$r_i^{-1} y_i - r_i^{-1} z_i = r_i^{-1} b_i; \qquad (*)$$

since $\{b_i\}$ is a bounded sequence and $r_i \to \infty$, $i \to \infty$, passing to limit in $(*)$ we get $\overline{y} = \overline{z}$. By its origin, $\overline{y}$ is the limit of sequence of points from L and thus $\overline{y} \in L$, and $\overline{z}$ is the limit of a sequence of points from the closed cone $\mathbf{K}$ and therefore $\overline{z} \in \mathbf{K}$. The bottom line is that in the case of (b) the set $L \cap \mathbf{K}$ contains the unit vector $\overline{z} = \overline{y}$, which is impossible due to $L \cap \mathbf{K} = \{0\}$. Thus, (b) is impossible, and we are done.

Finally, in the case of $L \cap \operatorname{int} \mathbf{K} \neq \varnothing$ B is closed by a very simple reason – in this case $B = \mathbf{R}^m$. Indeed, if $a \in \operatorname{int} \mathbf{K} \cap L$, then $\lambda a - b \in \mathbf{K}$ for all large enough positive λ, that is, $b = \lambda a - z$ for certain $\lambda > 0$ and $z \in \mathbf{K}$. And since $a \in L$, we have $\lambda a \in L$ as well, that is, $b \in L - \mathbf{K}$. ■

Exercise IV.5. [follow-up to Exercise IV.4] Let $\mathbf{K} \subset \mathbf{R}^m$ be a regular cone, $P \in \mathbf{R}^{m \times n}$, $Q \in \mathbf{R}^{m \times k}$, and $p \in \mathbf{R}^m$. Consider the set

$$\overline{K} \quad = \quad \{x \in \mathbf{R}^n : \exists u \in \mathbf{R}^k : Px + Qu + p \in \mathbf{K}\}$$

This set clearly is convex. When the cone $\mathbf{K}$ is polyhedral, the above description of $\overline{K}$ is its polyhedral representation, so that the set $\overline{K}$ is polyhedral and as such is closed.

The goal of this exercise is to understand what happens with closedness of $\overline{K}$ when $\mathbf{K}$ is a general-type regular cone.

1. Is it true that $\overline{K}$ is closed whenever $\mathbf{K}$ is a regular cone?
   Hint: Look what happens when $\mathbf{K} = \mathbf{L}^3$, $P = I_3$, $Q = [0;1;1] \in \mathbf{R}^{3\times 1}$, and $p = [0;0;0]$
2. Prove when $\mathbf{K}$ is a regular cone and $\operatorname{Im} Q \cap \mathbf{K} = \{0\}$, $\overline{K}$ is closed.

Solution: 1: In the situation of Hint. denoting by L the linear subspace $\{x \in \mathbf{R}^3 : x_2 = 0\}$ we have

$$\overline{K} \cap L = \{x = [x_1;0;x_3] : \exists u \in \mathbf{R} : [x_1;u;x_3+u] \in \mathbf{L}^3\} = \{x = [x_1;0;x_3] : \exists u : x_3 + u \geq \sqrt{x_1^2 + u^2}\}.$$

From the concluding description of $\overline{K} \cap L$ we see that this set contains all triples $[1;0;\epsilon]$ with $\epsilon > 0$ and does not contain the triple $[1;0;0]$ and therefore is not closed; consequently, $\overline{K}$ is not closed as well (L is closed!).

2: Assuming $\mathbf{K}$ regular and $L \cap \mathbf{K} = \{0\}$, where $L = \operatorname{Im} Q$, the set $Z = \{b \in \mathbf{R}^m : \exists u \in L : u + b \in \mathbf{K}\}$ is closed (by Exercise IV.4.4); it remains to note that $\overline{K}$ is the inverse image of the closed set Z under the continuous mapping $x \mapsto Px + p$ and as such is closed along with Z. ■

Exercise IV.6. Let $\mathfrak{n}(x)$ be a norm on $\mathbf{R}^n$ such that $\mathfrak{n}$ is continuously differentiable outside of the origin, and let

$$\mathfrak{n}_*(y) = \max_x \{y^\top x : \mathfrak{n}(x) \le 1\}.$$

be the norm conjugate to $\mathfrak{n}$ (see Fact III.13.4), so that $\mathfrak{n}_*(\cdot)$ is a norm such that

$$x^\top y \le \mathfrak{n}(x)\mathfrak{n}_*(y)\, \forall x, y \in \mathbf{R}^n$$

and $(\mathfrak{n}_*)_* = \mathfrak{n}$, implying that for every $x \ne 0$ there exists $y \ne 0$ such that

$$x^\top y = \mathfrak{n}(x)\mathfrak{n}_*(y).$$

Here are your tasks:

1. Let M be a $d \times d$ matrix, $d \ge 2$, with diagonal entries equal to 1. Assume that $M\lambda \le 0$ for some nonzero vector $\lambda \ge 0$. How large could be $\min_{i,j} M_{ij}$?

   Solution: $\mu := \min_{i,j} M_{i,j} \le -\frac{1}{d-1}$. Indeed, assuming that $M\lambda \le 0$ with nonzero $\lambda \ge 0$, let k be the index of largest entry in λ. We have

   $$0 \ge \sum_j M_{kj}\lambda_j = \lambda_k + \sum_{j \ne k} M_{kj}\lambda_j \ge \lambda_k + \lambda_k(d-1)\mu \implies \mu \le -\frac{1}{d-1}.$$

   For the matrix M with diagonal entries equal to 1 and off-diagonal entries equal to $-1/(d-1)$ we have $M[1;...;1] = 0$, that is, the bound $\min_{i,j} M_{ij} \le -\frac{1}{d-1}$ is unimprovable.

2. For $d \ge 2$, let $p_1, ..., p_d$ be $\mathfrak{n}_*(\cdot)$-unit vectors, $w_1, ..., w_d$ be $\mathfrak{n}(\cdot)$-unit vectors, and let $p_i^\top w_i = 1$, $1 \le i \le d$. Assume that $0 \in \mathrm{Conv}\{p_1, ..., p_d\}$. How small could be $\max_{i \ne j} \mathfrak{n}(w_i - w_j)$?

   Solution: $\max_{i,j \le d} \mathfrak{n}(w_i - w_j) \ge \frac{d}{d-1}$. Indeed, consider the $d \times d$ matrix $M = [M_{ij} = w_i^\top p_j]_{i,j \le d}$, We have $0 = \sum_j \lambda_j p_j$ with properly selected $\lambda \ge 0$ such that $\sum_j \lambda_j = 1$, so that $M\lambda = 0$. Besides this, the diagonal entries in M are equal to 1. By the previous item, we have $M_{ij} \le -\frac{1}{d-1}$ for some i, j, that is,

   $$\mathfrak{n}(w_j - w_i) = \mathfrak{n}_*(p_j)\mathfrak{n}(w_j - w_i) \ge p_j^\top [w_j - w_i] = 1 - M_{ij} \ge 1 + \frac{1}{d-1} = \frac{d}{d-1}.$$

3. Let $x \in \mathbf{R}^n$ be nonzero.
   1. Let $g = \nabla \mathfrak{n}(x)$.
      1. What is $\mathfrak{n}_*(g)$?

         Solution: For every $h \in \mathbf{R}^n$ we have $\mathfrak{n}(x+h) \ge \mathfrak{n}(x) + g^\top h$, whence also $\mathfrak{n}(x) + \mathfrak{n}(h) \ge \mathfrak{n}(x) + g^\top h$, that is, $\mathfrak{n}(h) \ge g^\top h$, implying that $\mathfrak{n}_*(g) = \max_h \{g^\top h : \mathfrak{n}(h) \le 1\} \le 1$. On the other hand, $0 = \mathfrak{n}(0) \ge \mathfrak{n}(x) - g^\top x$, that is, $g^\top x \ge \mathfrak{n}(x)$. Besides this, $g^\top x \le \mathfrak{n}_*(g)\mathfrak{n}(x)$, and we get $\mathfrak{n}(x) \le g^\top x \le \mathfrak{n}_*(g)\mathfrak{n}(x)$, implying that $\mathfrak{n}_*(g) \ge 1$. The bottom line is that $\mathfrak{n}_*(g) = 1$.
      2. What is $g^\top x$?

         Solution: $g^\top x = \mathfrak{n}(x)$ – differentiate the identity $\mathfrak{n}(tx) = t\mathfrak{n}(x)$, $t > 0$, in t at $t = 1$.
      3. Let e be such that $\mathfrak{n}_*(e) \le \mathfrak{n}_*(g)$ and $e^\top x = g^\top x$. Is it true that $e = g$?

         Solution: Yes. e in question should satisfy $\mathfrak{n}_*(e) \le \mathfrak{n}_*(g)$, that is, $\mathfrak{n}_*(e) \le 1$ by item 3.1.1. Next, $e^\top x = g^\top x$, that is, $e^\top x = \mathfrak{n}(x)$ by item 3.1.2. Therefore for every h it holds $e^\top h = e^\top(x+h) - e^\top x \le \mathfrak{n}_*(e)\mathfrak{n}(x+h) - \mathfrak{n}(x) \le \mathfrak{n}(x+h) - \mathfrak{n}(x)$, that is, $\mathfrak{n}(x) + e^\top h \le \mathfrak{n}(x+h)$ for all h, so that e is a subgradient of f at x. Since $x \ne 0$ and $\mathfrak{n}$ is differentiable outside of the origin, we have $e = \nabla \mathfrak{n}(x) = g$.
   2. Given N points $y_i \in \mathbf{R}^n$, consider the problem of finding the smallest $\mathfrak{n}(\cdot)$-ball containing $y_1, ..., y_N$.

1. Write down the problem as a conic one, and write down the conic dual of this problem. Are both the problems solvable with equal optimal values?

   Solution: The problem in question is

   $$\begin{array}{rcl} \mathrm{Opt}(P) & = & \min_{x,t}\{t : \mathfrak{n}(x-y_i) \le t, 1 \le i \le N\} \\ & = & \min_{[x;t]}\{t : [x-y_i;t] \in \mathbf{K}, i \le N\} \\ & & \mathbf{K} = \{[u;t] \in \mathbf{R}^n \times \mathbf{R} : \mathfrak{n}(u) \le t\} \end{array} \qquad (P)$$

   its dual is

   $$\begin{array}{rcl} \mathrm{Opt}(D) & = & \max_{z_1,\ldots,z_N,s_1,\ldots,s_N}\left\{\sum_i z_i^\top y_i : \begin{array}{l}\sum_i z_i = 0, \sum_i s_i = 1 \\ [z_i;s_i] \in \mathbf{K}_*, i \le N\end{array}\right\} \\ & & \mathbf{K}_* = \{[z;s] \in \mathbf{R}^n \times \mathbf{R} : \mathfrak{n}_*(z) \le s\} \end{array} \qquad (D)$$

   and both problems are solvable with equal optimal values.
   Indeed, (P) is self-explanatory; the fact that $\mathbf{K}$ is a regular cone is evident. The fact that the dual cone is as indicated in (D) is immediate: denoting a vector from $\mathbf{R}^n \times \mathbf{R}$ by $[z;s]$, this vector is in the cone dual to $\mathbf{K}$ if and only if for every $t \ge 0$ one has

   $$0 \le \min_u\{[u;t]^\top[z;s] : [u;t] \in \mathbf{K}\} = \min_u\{st + u^\top z : \mathfrak{n}(u) \le t\} = st - t\mathfrak{n}_*(z),$$

   implying that the dual cone is as in (D). Now, to get the dual problem, we should equip the constraints $[x-y_i;t] \in \mathbf{K}$ of (P) with Lagrange multipliers $[z_i;s_i] \in \mathbf{K}_*$ in such a way, that the left hand side in the aggregated constraint $\sum_i [z_i;s_i]^\top[x-y_i;t] \ge 0$, that is, in the inequality

   $$\sum_i ts_i + \sum_i z_i^\top x \ge \sum_i z_i^\top y_i$$

   is identically in t, x equal to the primal objective t, and to maximize under this restriction (taken along with the restrictions $[z_i;s_i] \in \mathbf{K}_*$) the right hand side of the aggregated constraint, which results in (D).
   Problem (P) clearly is strictly feasible (to get a strictly feasible solution, set $x = 0$ and take $t > \max_i \mathfrak{n}(y_i)$) and solvable (since the sets of feasible solutions where the objective is upper-bounded by a given real are compact); by Conic Duality Theorem, the dual problem is solvable with the same optimal value as (P).

2. Assume that the data are such that the optimal value in (P) is equal to 1. How small can be $\max_{i,j} \mathfrak{n}(y_i - y_j)$?
   Hint: write down and analyze optimality conditions.

   Solution: $\max_{i,j} \mathfrak{n}(y_i - y_j) \ge \frac{n+1}{n}$.
   Indeed, let $[x;1]$ be primal optimal, and $[z_i;s_i]$, $i \le N$, be dual optimal. By optimality conditions (complementary slackness) the primal slacks $[x-y_i;1]$ should be orthogonal to the respective dual solutions $[z_i;s_i]$, that is,

   $$s_i = [y_i - x]^\top z_i, i \le N, \qquad (\#)$$

   and since $\mathfrak{n}(y_i - x) \le 1$ and $\mathfrak{n}_*(z_i) \le s_i$ for all i by primal and dual feasibility, we have $\mathfrak{n}(y_i - x) \le \mathrm{Opt}(P) = 1$, so that the right hand side in $(\#)$ is $\le \mathfrak{n}(y_i - x)\mathfrak{n}_*(z_i) \le \mathfrak{n}_*(z_i)$, and since $\mathfrak{n}_*(z_i) \le s_i$ by dual feasibility, $(\#)$ implies that $\mathfrak{n}_*(z_i) = s_i$ for all i. Next, setting $w_i = y_i - x$, we have $\mathfrak{n}(w_i) \le 1$, so that the right hand side in $(\#)$ is $\le \mathfrak{n}(w_i)\mathfrak{n}_*(z_i) = \mathfrak{n}(w_i)s_i$; therefore $(\#)$ implies that $\mathfrak{n}(w_i) = 1$ for all $i \in \mathcal{I} = \{i : s_i > 0\}$. Note that the set $\mathcal{I}$ is nonempty, since otherwise $\mathrm{Opt}(D)$ would be 0 and not 1. Now, from the constraints of (D) we have

   $$\sum_{i \in \mathcal{I}} z_i = 0,$$

and $\mathfrak{n}_*(z_i) = s_i > 0$ for $i \in \mathcal{I}$, so that setting $p_i = s_i^{-1} z_i$, $i \in \mathcal{I}$, we have

$$\mathfrak{n}(w_i) = 1,\ i \in \mathcal{I}\ \&\ \mathfrak{n}_*(p_i) = 1, i \in \mathcal{I}\ \&\ p_i^\top w_i = 1, i \in \mathcal{I}\ \&\ \sum_{i\in\mathcal{I}} s_i p_i = 0, \qquad (!)$$

where the relations $p_i^\top w_i = 1$, $i \in \mathcal{I}$, stem from (#) due to $s_i > 0$, $i \in \mathcal{I}$.
Since $0 < s_i$ for $i \in \mathcal{I} \neq \varnothing$, the last relation in (!) means that $0 \in \mathrm{Conv}\{p_i : i \in \mathcal{I}\}$. By Caratheodory Theorem, we can find a subset $I \subset \mathcal{I}$ of cardinality d, $2 \le d \le n+1$, such that $0 \in \mathrm{Conv}\{p_i, i \in I\}$. Assuming w.l.o.g. that $I = \{1, ..., d\}$, for $M = [w_i^\top p_j]_{i,j\le d}$ and some $\lambda \in \mathbf{R}^d_+$ with $\sum_i \lambda_i = 1$ we have

$$\mathfrak{n}_*(p_i) = 1, \mathfrak{n}(w_i) = 1, p_i^\top w_i = 1, i \le d\ \&\lambda \ge 0, \lambda \neq 0, M\lambda = 0 \qquad (!!)$$

By item 2, we have $\max_{i,j\le d} \mathfrak{n}(w_i - w_j) \ge \frac{d}{d-1} \ge \frac{n+1}{n}$, and it remains to note that $w_i - w_j = y_i - y_j$.

3. In the situation of item 3.2.2, assume that $\mathfrak{n}(x) = \|x\|_2$ is the standard Euclidean norm. How small can be $\max_{i,j} \mathfrak{n}(y_i - y_j)$ now?

   Solution: The concluding relation in the solution to the previous item now reads $\|p_i\|_2 = \|w_i\|_2 = 1$ and $p_i^\top w_i = 1$, $1 \le i \le d \le n+1$, whence $p_i = w_i$, $i \le d$. By item 1, (!!) implies that $\min_{i,j\le d} w_i p_j^\top \le -\frac{1}{d-1}$, that is, there exist $i, j \le d$ such that $w_i^\top p_j \le -\frac{1}{d-1}$, that is, $w_i^\top w_j \le -\frac{1}{d-1}$. Consequently,

$$\|w_i - w_j\|^2 = w_i^\top w_i + w_j^\top w_j - 2w_i^\top w_j \ge 2(1 + \frac{1}{d-1}) = \frac{2d}{d-1},$$

that is, $\max_{i,j} \|y_i - y_j\|_2 \ge \max_{i,j\le d} \|w_i - w_j\|_2 \ge \sqrt{\frac{2d}{d-1}} \ge \sqrt{\frac{2(n+1)}{n}}$.
Note: instead of asking how large is the maximum of pairwise distances between $y_i \in \mathbf{R}^n$ given that the smallest Euclidean ball containing $y_1, ..., y_N$ is of radius 1, we could ask how large could be radius of the smallest Euclidean ball containing the points $y_1, ..., y_N \in \mathbf{R}^n$ with pairwise $\|\cdot\|_2$-distances not exceeding 1, and in terms of the latter question, the above result states that this radius is at most $\sqrt{\frac{n}{2(n+1)}}$. This is called "Jung's Theorem;" the result is sharp, since the smallest radius Euclidean ball containing the $n+1$ vertices of the perfect simplex (simplex in $\mathbf{R}^n$ with distances 1 between every two vertices) is exactly $\sqrt{\frac{n}{2(n+1)}}$; to see this, realize the perfect simplex as $\{x \in \mathbf{R}^{n+1}_+ : \sum_i x_i = 1/\sqrt{2}\}$, and $\mathbf{R}^n$ - as the hyperplane $\sum_i x_i = 1/\sqrt{2}$ in $\mathbf{R}^{n+1}$.

24.2 Geometry of primal-dual pair of conic problems

Exercise IV.7. [geometry of primal-dual pair of conic problem] The goal of the Exercise is to reveal notable geometry of primal-dual pair of conic problem.

It is convenient to work with the primal problem in the form

$$\mathrm{Opt}(P) = \min_x \left\{ c^\top x : Ax - b \ge_{\mathbf{K}} 0,\ Px = p \right\} \qquad (P)$$

where $\mathbf{K}$ is a regular cone in certain $\mathbf{R}^N$. As is immediately seen, the conic dual of (P) reduces to the problem

$$\mathrm{Opt}(D) = \max_{y,z} \left\{ b^\top y + p^\top z : y \in \mathbf{K}_*,\ A^\top y + P^\top z = c \right\} \quad [11] \qquad (D)$$

From now on we make the following, in fact, rather weak,

[11] building conic dual to a conic problem is a purely mechanical process; however, this process as presented in section 18.4 operates with conic problem in a form slightly different from the one of (P), namely, with linear inequality constraints instead of linear equalities. To apply this process to

Assumption: *The systems of linear equality constraints in (P) and (D) are solvable.*

Let us fix $\overline{x}$ and $(\overline{y}, \overline{z})$ such that

$$P\overline{x} = p \;\&\; A^\top \overline{y} + P^\top \overline{z} = c. \tag{\#}$$

Your first task is as follows:

1. Pass in (P) from variables x to *primal slack* $\xi = Ax - b$. Specifically, prove that in terms of primal slack (P) becomes the problem

$$\begin{array}{c} \mathrm{Opt}(\mathcal{P}) = \min_\xi \left\{\overline{y}^\top \xi : \xi \in \mathbf{K} \cap [\mathcal{L} - \overline{\xi}]\right\} \\ \left[\mathcal{L} = \{\xi : \exists x : \xi = Ax, Px = 0\},\ \overline{\xi} = b - A\overline{x}\right] \end{array} \tag{$\mathcal{P}$}$$

   namely, prove that
   (i) Every feasible solution x to (P) induces feasible solution $\xi = Ax - b$ to $(\mathcal{P})$, and the value of the objective of (P) at x differs from the value of the objective of $(\mathcal{P})$ at $Ax - b$ by the independent of x constant:

$$\overline{y}^\top \xi = c^\top x - \left[\overline{y}^\top b + \overline{z}^\top p\right]. \tag{A}$$

   (ii) Vice versa, every feasible solution ξ to $(\mathcal{P})$ is of the form $Ax - b$ for some feasible solution x to (P).
   The bottom line is that (P) can be reformulated equivalently as $(\mathcal{P})$, and the optimal values of these two problems are linked by the relation

$$\mathrm{Opt}(\mathcal{P}) = \mathrm{Opt}(P) - \left[\overline{y}^\top b + \overline{z}^\top p\right].$$

Solution: Let x be feasible for (P) and $\xi = Ax - b$. Then ξ satisfies the inclusion $\xi \in \mathbf{K}$ and

$$\xi = A[x - \overline{x}] + [A\overline{x} - b] = A[x - \overline{x}] - \overline{\xi}$$

and $P[x - \overline{x}] = 0$, that is, $\xi \in \mathbf{K} \cap [\mathcal{L} - \overline{\xi}]$. This reasoning can be easily reversed to demonstrate that if $\xi \in \mathbf{K} \cap [\mathcal{L} - \overline{\xi}]$, then $\xi = Ax - b$ for some x feasible for (P). Besides this,

$$c^\top x = [A^\top \overline{y} + P^\top \bar{z}]^\top x = \overline{y}^\top [Ax - b] + \overline{y}^\top b + \overline{z}^\top Px = \overline{y}^\top \xi + \left[\overline{y}^\top b + \overline{z}^\top p\right],$$

as claimed in (A).
On the other hand, when ξ is feasible for $(\mathcal{P})$, we have $\xi \in \mathbf{K}$ and $\xi = Ax' - \overline{\xi}$ for some x' with $Px' = 0$, whence

$$\mathbf{K} \ni \xi = Ax' - \overline{\xi} = A[x' + \overline{x}] - b = Ax - b,$$

where $x = x' + \overline{x}$ satisfies $Px = p$. We conclude that $\xi = Ax - b$ with x feasible for (P). ■

Next task is as follows:

2. Pass from problem (D) in variables y, z to problem

$$\begin{array}{c} \max\limits_y \left\{\overline{\xi}^\top y : y \in \mathbf{K}_* \cap [\mathcal{L}^\perp + \overline{y}]\right\} \\ \left[\mathcal{L}^\perp := \{y : y^\top \xi = 0\, \forall \xi \in \mathcal{L}\} = \{y : \exists z : A^\top y + P^\top z = 0\}\right] \end{array} \tag{$\mathcal{D}$}$$

   in variable y only, specifically, prove that
   (i) The orthogonal complement $\mathcal{L}^\perp$ of $\mathcal{L}$ indeed is the linear subspace $\{y : \exists z : A^\top y + P^\top z = 0\}$.
   (ii) y-component of feasible solution (y, z) to (D) is a feasible solution to $(\mathcal{D})$, and vice versa –

   (P), it suffices to represent the linear equalities $Px = p$ by a pair of opposite linear inequalities $Px - p \geq 0, -Px + p \geq 0$. Applying the recipe from section 18.4 to the resulting problem, the dual reads

$$\max_{y, z', z''} \left\{b^\top y + [z' - z'']^\top p : A^\top y + P^\top [z' - z''] = c,\ y \in \mathbf{K}_*, z' \geq 0, z'' \geq 0\right\}.$$

   Passing from z', z'' to $z = z' - z''$, we reduce the latter problem to (D).

every feasible solution y to $(\mathcal{D})$ can be augmented by z to yield a feasible solution (y,z) to (D). Besides this, whenever (y,z) is feasible for (D), we have

$$b^T y + p^T z = \overline{\xi}^T y + c^T \overline{x}. \qquad (B)$$

The bottom line is that (D) can be reformulated equivalently as $(\mathcal{D})$, and the optimal values of these two problems are linked by the relation

$$\mathrm{Opt}(\mathcal{D}) = \mathrm{Opt}(D) - c^T \overline{x}.$$

Solution: (i): To prove that $\mathcal{L}^\perp = \{y : \exists z : A^T y + P^T z = 0\}$ is the same as to prove that the necessary and sufficient condition for equality $y^T \xi = 0$ treated as equality in variables ξ, x to be consequence of the system of linear equalities $\xi - Ax = 0, Px = 0$ in variables ξ, x is for y to admit selection of z such that $A^T y + P^T z = 0$, but this is what Linear Algebra (not speaking about Homogeneous Farkas Lemma) says: a homogeneous linear equation is a consequence of a system of homogeneous linear equations if and only is the vector of coefficients of this equation (in our case, the vector $[y^T, 0_{1\times n}]$) is linear combination of the vectors of coefficients of the equations from the system, which in the case in question boils down to $y^T A + z^T P = 0$ for certain z. ■

(ii) If (y,z) is feasible for (D), then $[A^T, P^T][y - \overline{y}; z - \overline{z}] = 0$, that is, $y \in \mathcal{L}^\perp + \overline{y}$ by already proved (i), and $y \in \mathbf{K}_*$, that is, y is feasible for $(\mathcal{D})$. Besides this, $A^T y + P^T z = c$, $\overline{\xi} = b - A\overline{x}$, and $P\overline{x} = p$, whence

$$\overline{\xi}^T y - [b^T y + p^T z] = [b - A\overline{x}]^T y - [b^T y + p^T z] = -\overline{x}^T A^T y - p^T z = \overline{x}^T [P^T z - c] - p^T z = -\overline{x}^T c,$$

as required in (B).

Vice versa, if y is feasible for $(\mathcal{D})$, then $y \in \mathbf{K}_*$ and $y - \overline{y} \in \mathcal{L}^\perp$, that is, by (i), for properly selected w one has $A^T[y - \overline{y}] + P^T w = 0$. This, due to the origin of $\overline{y}$ implies that

$$A^T y + P^T w = A^T \overline{y} = c - P^T \overline{z},$$

so that y can be augmented by $z = \overline{z} + w$ to yield a feasible solution to (D). (ii) is proved. ■

The summary of items 1 and 2 is as follows:

- Primal-dual pair (P), (D) of conic problems reduces to pair of problems $(\mathcal{P})$, $(\mathcal{D})$, "reduces" meaning that feasible solutions x and (y,z) to (P), (D) induce feasible solutions $\xi = Ax - b$ and y to $(\mathcal{P})$, $(\mathcal{D})$, and every pair of feasible solutions to the latter problems can be obtained, in the fashion just described, from a pair of feasible solutions to (P), (D);
- Geometrically, $(\mathcal{P})$, $(\mathcal{D})$ are as follows:
  - Problems' data are (a) primal-dual pair of regular cones $\mathbf{K}$, $\mathbf{K}_*$ in some $\mathbf{R}^N$, (b) pair of linear subspaces $\mathcal{L}_\mathcal{P}$, $\mathcal{L}_\mathcal{D}$ in $\mathbf{R}^N$ which are orthogonal complements to each other, and (c) pair of vectors $\overline{y}, \overline{\xi}$ in $\mathbf{R}^N$.
  - $(\mathcal{P})$ is the problem of minimizing linear objective $\overline{y}^T \xi$ over the intersection of the *primal feasible plane* $\mathcal{M}_\mathcal{P} := \mathcal{L}_\mathcal{P} - \overline{\xi}$ with the cone $\mathbf{K}$, while $(\mathcal{D})$ is the problem of maximizing the linear objective $\overline{\xi}^T y$ over the intersection of the *dual feasible plane* $\mathcal{M}_\mathcal{D} := \mathcal{L}_\mathcal{D} + \overline{y}$ with the dual cone $\mathbf{K}_*$.

Pay attention to the "nearly perfect" primal-dual symmetry; the only asymmetry is that in the primal feasible plane the shift vector is $-\overline{\xi}$ – minus the vector of coefficients of the objective in $(\mathcal{D})$, while in the dual feasible plane the shift vector is $\overline{y}$ – the vector of coefficients of the objective in $(\mathcal{P})$. This minor asymmetry stems from the fact that by tradition one of the problems (in our presentation, $(\mathcal{P})$) is written as a minimization program, and the other problem from the pair as a maximization one.

In fact, the symmetry can be made perfect, and the objectives – eliminated at all.

3. Consider pairs of problems (P), (D) along with problems $(\mathcal{P})$, $(\mathcal{D})$, and let x, (y,z) be feasible solutions to (P), (D), and ξ, y – the feasible solutions to $(\mathcal{P})$, $(\mathcal{D})$ induced by x and (y,z), respectively. Prove that the *duality gap*

$$\mathrm{DualityGap}(x; y, z) := c^T x - [b^T y + p^T z]$$

– the difference between the objective of primal problem (P) evaluated at primal feasible solution x and the objective of the dual problem (D) evaluated at the dual feasible solution (y,z) – is nothing but the inner product $\xi^\top y$ of ξ and y.

Solution: Here is the computation: Let x, (y,z) be feasible for (P), (D), and $\xi = Ax - b$, y be the induced by x, (y,z) feasible solutions to $(\mathcal{P})$, $(\mathcal{D})$. Then

$$\begin{array}{rcl} 0 & = & [y-\overline{y}]^\top[\xi+\overline{\xi}] \text{ [since } y-\overline{y}\in\mathcal{L} \text{ and } \xi+\overline{\xi}\in\mathcal{L}^\perp] \\ \Longrightarrow\ y^\top\xi & = & [\overline{y}^\top\xi - \overline{\xi}^\top y] + \overline{y}^\top\overline{\xi} \\ & = & [c^\top x - [b^\top y + p^\top z]] - b^\top\overline{y} - p^\top\overline{z} + c^\top\overline{x} + \overline{y}^\top\overline{\xi} \text{ [by } (A) \text{ and } (B)] \\ & = & \mathrm{DualityGap}(x;y,z) - b^\top\overline{y} - p^\top\overline{z} + [A^\top\overline{y} + P^\top\overline{z}]^\top\overline{x} + \overline{y}^\top[b - A\overline{x}] \\ & & \qquad\qquad \text{[by origin of } \overline{y},\overline{z},\overline{\xi}] \\ & = & \mathrm{DualityGap}(x;y,z) - p^\top\overline{z} + [A^\top\overline{y} + P^\top\overline{z}]^\top\overline{x} - \overline{y}^\top A\overline{x} \\ & = & \mathrm{DualityGap}(x;y,z) \text{ [since } P\overline{x} = p] \end{array}$$

24.3 Around $\mathcal{S}$-Lemma

Exercise IV.8. Recall that $\mathcal{S}$-Lemma guarantees that the validity of the implication

$$x^TAx \geq 0 \implies x^TBx \geq 0 \qquad [A,B\in\mathbf{S}^n]$$

is the same as the existence of $\lambda \geq 0$ such that $B \succeq \lambda A$ only under the assumption that the inequality $x^TAx \geq 0$ is strictly feasible. Does the lemma remain true when this assumption is lifted?

Solution: The answer is negative. When $n=2$, $x^TAx = -x_2^2$ and $x^TBx = 2x_1x_2$, the above implication holds true, but the quadratic form $x^T(B-\lambda A)x = 2x_1x_2 + \lambda x_2^2$ is not everywhere nonnegative whatever be $\lambda\in\mathbf{R}$.

Exercise IV.9. Given $A\in\mathbf{S}^n$, consider the set $Q_A = \{x\in\mathbf{R}^n : x^\top Ax \leq 0\}$.

1. Let $B\in\mathbf{S}^n$ be such that $B\neq A$ and $Q_B = Q_A$. Then, is it always true that there exists $\rho > 0$ such that $B = \rho A$?
2. Suppose that $A\in\mathbf{S}^n$ satisfies $A_{ij}\geq 0$ for all i,j. Under this condition, does your answer to item 1 change?
3. Suppose that $A\in\mathbf{S}^n$ satisfies $\lambda_{\min}(A) < 0 < \lambda_{\max}(A)$. Under this condition, does your answer to item 1 change?

Solution:

1 : A counter-example is given by $A = -I$ and $B = \mathrm{Diag}\{-1,-2,\ldots,-n\}$ where $Q_A = Q_B = \mathbf{R}^n$.

2 : A counter-example is given by $A = 0$ and $B = -I$, where $Q_A = Q_B = \mathbf{R}^n$.

3 : Suppose $x^\top Ax \leq 0 \iff x^\top Bx \leq 0$. Then $x^\top(-A)x \geq 0 \iff x^\top(-B)x \geq 0$. Since $\lambda_{\min}(A) < 0 < \lambda_{\max}(A)$, $Q_A \neq \mathbf{R}^n$ and $Q_A \neq \{0\}$. Therefore, $\lambda_{\min}(B) < 0 < \lambda_{\max}(B)$ also. Furthermore, the same eigenvalue condition holds for both $-A, -B$, which means $x^\top(-A)x \geq 0$ and $x^\top(-B)x \geq 0$ are both strictly feasible. By the $\mathcal{S}$-lemma, this implies that there exist $\lambda_1,\lambda_2\geq 0$ such that

$$-B \succeq -\lambda_1 A \implies \lambda_1 A \succeq B$$
$$-A \succeq -\lambda_2 B \implies \lambda_2 B \succeq A.$$

Note that $\lambda_1,\lambda_2 > 0$, otherwise one of Q_A, Q_B will be $\mathbf{R}^n$, which we have already established is not true. Therefore, we can multiply the first inequality by $1/\lambda_1 > 0$ to get

$$A \succeq \frac{1}{\lambda_1}B \implies \lambda_2 B \succeq \frac{1}{\lambda_1}B \implies (\lambda_1\lambda_2 - 1)B \succeq 0.$$

Since B is not positive semidefinite or negative semidefinite, we must have $\lambda_1\lambda_2 = 1 \implies \lambda_2 = 1/\lambda_1$. But this means

$$\lambda_1 A \succeq B$$
$$\lambda_2 B = \frac{1}{\lambda_1}B \succeq A \implies B \succeq \lambda_1 A$$

which combines with $\lambda_1 A \succeq B$ to imply that $B = \lambda_1 A$ with $\lambda_1 > 0$. ■

Exercise IV.10. For two nonzero reals a, b, one has $2|ab| = \min_{\lambda>0}[\lambda^{-1}a^2 + \lambda b^2]$, implying by the Schur Complement Lemma that $2|ab| \le c$ if and only if there exists $\lambda > 0$ such that $\left[\begin{array}{c|c} c-\lambda b^2 & a \\ \hline a & \lambda \end{array}\right] \succeq 0$. Assuming $b \neq 0$, we have also $2|ab| \le c$ if and only if there exists $\lambda \ge 0$ such that $\left[\begin{array}{c|c} c-\lambda b^2 & a \\ \hline a & \lambda \end{array}\right] \succeq 0$. Note also that $c \ge 2|ab|$ is the same as $c \ge 2a\delta b$ for all $\delta \in [-1, 1]$.

Prove the following matrix analogy of the above observation:

Let $A \in \mathbf{R}^{p\times r}$, $B \in \mathbf{R}^{p\times s}$, let $B \neq 0$, and let $\mathcal{D} = \{\Delta \in \mathbf{R}^{r\times s} : \|\Delta\| \le 1\}$, where $\|\cdot\|$ is the spectral norm. Then $C \succeq [A\Delta B^\top + B\Delta^\top A^\top]$ for all $\Delta \in \mathcal{D}$ if and only if there exists $\lambda \ge 0$ such that $\left[\begin{array}{c|c} C-\lambda BB^\top & A \\ \hline A^\top & \lambda I_r \end{array}\right] \succeq 0$. In particular, when $a, b \in \mathbf{R}^p$ and $b \neq 0$, one has $C \succeq \pm[ab^\top + ba^\top]$ if and only if there exists $\lambda \ge 0$ such that $\left[\begin{array}{c|c} C-\lambda bb^\top & a \\ \hline a^\top & \lambda \end{array}\right] \succeq 0$.

Solution: We have

$$\begin{array}{l} C \succeq [A\Delta B^\top + B\Delta^\top A^\top] \Longleftrightarrow x^\top Cx - 2x^\top A[\Delta B^\top x] \ge 0 \ \forall (x \in \mathbf{R}^p, \Delta \in \mathcal{D}) \\ \Longleftrightarrow x^\top Cx - 2x^\top A\xi \ge 0 \ \forall (x \in \mathbf{R}^p, \xi : \exists \Delta \in \mathcal{D} : \xi = \Delta B^\top x) \\ \Longleftrightarrow x^\top Cx - 2x^\top A\xi \ge 0 \ \forall (x \in \mathbf{R}^p, \xi \in \mathbf{R}^r : \xi^\top\xi \le x^\top BB^\top x) \\ \Longleftrightarrow \exists \lambda \ge 0 : x^\top Cx - 2x^\top A\xi \ge \lambda[x^\top BB^\top x - \xi^\top\xi] \ \forall (x, \xi) \ [\mathcal{S}\text{-Lemma}] \\ \Longleftrightarrow \exists \lambda \ge 0 : \left[\begin{array}{c|c} C-\lambda BB^\top & A \\ \hline A^\top & \lambda I_r \end{array}\right] \succeq 0. \end{array}$$

Note that the assumption $B \neq 0$ implies that the quadratic form $x^\top BB^\top x - \xi^\top\xi$ of x, ξ is positive at certain point, thus making $\mathcal{S}$-Lemma applicable. ■

Exercise IV.11. [Robust TTD] Let us come back to TTD problem (5.2). Assume we have solved this problem and have at our disposal the resulting *nominal truss* withstanding best of all, the total truss volume being a given $W > 0$, the load of interest f. Now, we cannot ignore the possibility that "in real life" the truss can be affected, aside of the load of interest f, by perhaps small, but still nonzero, occasional load composed of forces acting at the free nodes utilized by the nominal truss (think of railroad bridge and wind). In order for our truss to be useful, it should withstand well all small enough occasional loads of this type. Note that our design gives no guarantees of this type – when building the nominal truss, we took into account just one loading scenario f.

1. To get impression of potential dangers of "small occasional loads," run numerical study as follows:

   - Compute the optimal console t^* (see "Console design" in Exercise I.16)
   - Looking one by one at the free nodes $p^1, ..., p^\mu$ actually used by the nominal console, associate with every one of them single-force occasional load, the corresponding force acting at node under consideration, generate this force as random 2D vector of Euclidean length 0.01 (that is, 1% of the magnitude of the single nonzero force in the load of interest), and compute the compliance of the nominal truss w.r.t. to the resulting occasional load. Conclude that the nominal console can be crushed by small occasional load and is therefore completely impractical.

   Solution: Were the nominal truss be able to withstand occasional loads as well as it withstands the load of interest, we could expect the compliances w.r.t. occasional loads to be of order of 10^{-5} (the nominal compliance is ≈ 0.191, and reducing the load by factor α, we reduce the compliance by factor α^2). In our experiments, the actual compliance w.r.t. the worst small occasional external

force was as large as 0.344; the corresponding equilibrium displacement is shown on Figure S2IV.1:

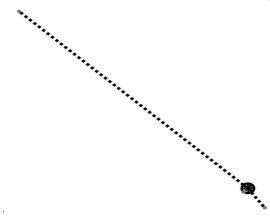


Figure S2IV.1. Deformation of nominal truss under small occasional load.

The dotted line on Figure S2IV.1 is the equilibrium displacement under small – just 1% of the force of interest – "badly placed" occasional external force. In the scale of this displacement, we merely do not see the original truss – it is represented by the black area on the figure. Thus, for all practical purposes, *the nominal truss can be completely crushed by a small occasional load and as such is completely impractical.*

2. Proposed cure is, of course, to use Robust Optimization methodology – to immunize the truss against small occasional loads, that is, to control its compliance w.r.t. the load of interest *and* all small occasional loads. An immediate question is where the occasional loads should be applied. There is no sense to allow them to act at all free nodes from the original set of tentative nodes – we have all reasons to believe that some, if not most, of these nodes will not be used in the optimal truss, so that we should not bother about forces acting at these nodes. On the other hand, we should take into account occasional loads acting at the nodes actually used by the optimal robust truss, and we do *not* know in advance what these nodes are. A reasonable compromise here as follows. After the nominal optimal truss is built, we can reduce the nodal set to the nodes actually used in this truss, allow for all pair connection of these nodes and resolve the TTD problem on this reduced sets of tentative nodes and tentative bars, now taking into account not only the load of interest, but all small occasional loads distributed along the nodes of our new nodal set. This approach can be implemented as follows.
   - We specify $\overline{\mathcal{V}}$ as the set of virtual displacements of nodes of our reduced nodal set, preserving the original status ("fixed" – "free") of these nodes, and denote by $\overline{f}$ the natural projection of the load of interest on $\overline{\mathcal{V}}$; note that all nonzero blocks in f – those representing nonzero physical forces from the collection specifying f – are inherited by $\overline{f}$, since the free nodes where these nonzero forces are applied should clearly be used by the nominal truss.
   - We specify $\mathcal{F}$ as the "ellipsoidal envelope" of $\overline{f}$ and all small in magnitude (measured in $\|\cdot\|_2$-norm) loads from $\overline{\mathcal{V}}$. Specifically, we use $\overline{f}$ as one of the half-axes of $\mathcal{F}$; the other $\overline{M}-1$ half-axes of $\mathcal{F}$ ($\overline{M}=\dim\overline{\mathcal{V}}$) are orthogonal to each other and to $\overline{f}$ vectors from $\overline{\mathcal{V}}$ of $\|\cdot\|_2$-norm $\rho\|\overline{f}\|_2$, where the "uncertainty level" $\rho\in[0,1]$ is a parameter of our construction. Note that
   $$\mathcal{F}=\{g=Ph:h^\top h\le 1\}$$
   for properly selected $\overline{M}\times\overline{M}$ matrix P.
   - We define the *robust compliance* $\overline{C}(\bar t)$ of a truss $\bar t\in\mathbf{R}^{\overline{N}}_+$ ($\overline{N}$ is the number of bars in our new – reduced – set of tentative bars), as the supremum, over $g\in\mathcal{F}$, of the usual compliances (computed for the new nodal set) of $\bar t$ w.r.t. load g, and pose the Robust Counterpart of the TTD problem as the problem of minimizing this robust compliance over trusses $\bar t\ge 0$ of total volume W. Solving this problem, we arrive at the *robust truss*.

   An immediate question is how to solve the Robust Counterpart. Those who solved Exercise I.16.3 know that as stated right now, the Robust Counterpart is the *semiinfinite* – with infinitely many convex constraints – optimization program
   $$\overline{\mathrm{Opt}}=\min_{\bar t,\tau}\left\{\tau:\bar t\in\mathbf{R}^{\overline{N}}_+,\sum_{i=1}^{\overline{N}}\bar t_i=W,\left[\begin{array}{c|c}\overline{B}\,\mathrm{Diag}\{\bar t\}\overline{B}^\top & g\\ \hline g^\top & 2\tau\end{array}\right]\succeq 0,\forall g\in\mathcal{F}\right\}\qquad(\#)$$

where $\overline{B}$ is the matrix built for the new TTD data in the same fashion as the matrix B was built for the original data.
Here go your tasks:

1. Reformulate (#) as a "normal" convex optimization problem – one with efficiently computable convex objective and finitely many explicitly verifiable convex constraints.
2. Solve the Console design version of the latter problem and subject the resulting robust truss to the same tests as those proposed above for quantifying the "real-life" quality of the nominal truss.

Solution: As we know from the solution to Exercise I.16.2-3, a real τ is an upper bound on the robust compliance of truss t iff

$$\forall g \in \mathcal{F}: \quad 2\tau \geq 2g^\top v - v^\top \overline{A}(\bar{t})v \ \forall v \in \mathbf{R}^{\overline{M}} \qquad [\overline{A}(\bar{t}) = \overline{B}\,\mathrm{Diag}\{\bar{t}\}\overline{B}^\top]$$

Note that when h runs through the $\|\cdot\|_2$-unit ball in $\overline{\mathcal{V}} = \mathbf{R}^{\overline{M}}$, vector $g = -Ph$ runs through the entire ellipsoid $\mathcal{F}$, so that the above relation is equivalent to

$$[u;h]^\top \overline{Q}[u;h] := -2h^\top P^\top u - u^\top \overline{A}(\bar{t})u \leq 2\tau \,\forall([u;h] \in \mathbf{R}^{2\overline{M}} : [u;h]^\top \overline{P}[u;h] := h^\top h \leq 1)$$

Applying Inhomogeneous $\mathcal{S}$-Lemma (Lemma IV.18.9), the latter relation takes place iff

$$\exists \lambda \geq 0 : \left[\begin{array}{c|c|c} \overline{A}(\bar{t}) & P & \\ \hline P^\top & \lambda I_{\overline{M}} & \\ \hline & & 2\tau - \lambda \end{array}\right] \succeq 0,$$

or, which is clearly the same, if and only if

$$\left[\begin{array}{c|c} \overline{A}(\bar{t}) & P \\ \hline P^\top & 2\tau I_{\overline{M}} \end{array}\right] \succeq 0.$$

The bottom line is that problem (#) is equivalent to the "normal" convex optimization problem

$$\overline{\mathrm{Opt}} = \min_{\bar{t},\tau} \left\{ \tau : \bar{t} \geq 0, \sum_i \bar{t}_i = W, \left[\begin{array}{c|c} \overline{B}\,\mathrm{Diag}\{\bar{t}\}\overline{B}^t & P \\ \hline P^\top & 2\tau I_{\overline{M}} \end{array}\right] \succeq 0 \right\}.$$

We solved the latter problem with $\rho = 0.1$ and tested the resulting robust truss against the load of interest $\overline{f}$ and 100 randomly selected occasional loads of magnitude 1% of the nominal load. The results are presented at Figure S2IV.2. Pay attention to the low cost of robustness: optimal *robust* compliance corresponding to the rather high (10%) uncertainty level is just by 10% larger than the optimal nominal compliance; compliance of the robust truss w.r.t. the load of interest is just by 0.6% larger than the compliance of the nominal truss.

24.4 Miscellaneous exercises

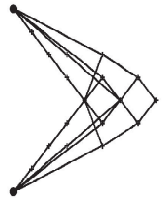
Exercise IV.12. Find the minimizer of a linear function

$$f(x) = c^\top x$$

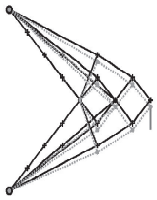
on the set

$$V_p = \{x \in \mathbf{R}^n \mid \sum_{i=1}^n |x_i|^p \leq 1\};$$

here p, $1 < p < \infty$, is a parameter. What happens with the solution when the parameter becomes 0.5?

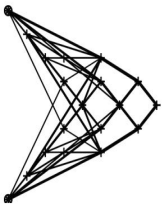


nominal truss, 38 bars
compliance 0.1917

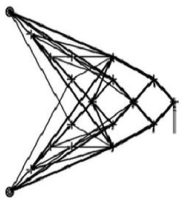


nominal truss,
displacement under load
of interest f, $\|f\|_2 = 1$

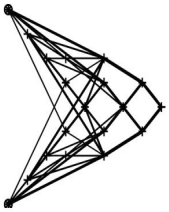
nominal truss,
displacement under badly placed
occasional load g, $\|g\|_2 = 0.01$



robust truss, 152 bars
robust compliance 0.1992



robust truss,
displacement under load
of interest f, $\|f\|_2 = 1$



robust truss,
displacement under badly placed
occasional load g, $\|g\|_2 = 0.01$

Figure S2IV.2. Nominal and robust consoles: positions of the bars and nodes before and after (in gray) deformation. The vertical segment starting at the right-most node: the external force.

Solution: Let us find a KKT point of the problem where the constraint is active. The KKT condition reads

$$c_i + \lambda p |x_i|^{p-1}\mathrm{sign}(x_i) = 0,\ i = 1, ..., n$$
$$\sum_i |x_i|^p = 1$$

whence

$$x_i = -\frac{|c_i|^{q-1}\mathrm{sign}(c_i)}{\|c\|_q^{q-1}},\ q = \frac{p}{p-1}$$

(we have assumed that $c \neq 0$, otherwise every feasible point is optimal). When $1 < p < \infty$, the problem is convex, so that the KKT point we have found is global optimal solution to the problem.
When $p = 0.5$, the solution is as follows. W.l.o.g. assume that $c_i \leq 0$; then at the optimum one clearly has $x_i \geq 0$; we lose nothing by adding these inequalities to the list of constraints. Assuming $x \geq 0$ and passing to new variables $y_i = \sqrt{x_i}$, our problem becomes

$$\min \sum_i c_i y_i^2 \text{ s.t. } y \geq 0, \sum_i y_i \leq 1.$$

Since $c_i \leq 0$, this is the problem of minimizing a *concave* function over the standard simplex; the solution is at the vertex of the simplex, and clearly this vertex is $y_{i_*} = 1$, $y_i = 0$, $i \neq i_*$, where i_* is the index of the most negative c_i. Thus, an optimal solution is the basic orth corresponding to the most negative c_i. In general (that is, without the assumption $c_i \leq 0$), the solution is ϵe_{i_*}, where i_* is the index of the maximal, in absolute value, coordinate of c, $\epsilon = \pm 1$ is the minus sign of this coordinate, and e_i are standard basic orths. The optimal value is $-\|c\|_\infty$.

Exercise IV.13 Every one of 3 random variables ξ_1, ξ_2, ξ_3 takes values 0 and 1 with probabilities 0.5, and every two of these 3 variables are independent. Is it true that all 3 variables are mutually independent? If not, how large could be probability of the event $\xi_1 = \xi_2 = \xi_3 = 1$?

Solution: We know that 8-dimensional probabilistic vector p representing the probability distribution of the random 3D Boolean vector $\xi = [\xi_1; \xi_2; \xi_3]$ satisfies a bunch of linear equalities and inequalities (specifically, is nonnegative and induces uniform on $\{0,1\}^2$ distributions of pairs of entries; we could add also "induces uniform on $\{0,1\}$ marginal distributions of entries," but this requirement is covered by the one on marginal distributions of pairs of entries), and ask what is under the circumstances the maximin allowed value of a particular entry. This is a simple LP program, and its optimal value turns out to be $1/4$ – twice the value of this entry in the distribution corresponding to the case of ξ_i independent across $i = 1, 2, 3$.

This simple example illustrates potential difficulties in recovering multivariate distributions from samples – with no a priori information on a probability distribution on, say, $\{0,1\}^d$ - a priori, it could be a whatever probabilistic vector p of dimension 2^d – statistically reliable recovery of p by sampling the corresponding random vector would require exponential in d, and thus unrealistic already for moderate d's, sample sizes. An alternative could be to try to "reconstruct" p from something we can estimate by sampling reliably, e.g., from low-dimensional marginal distributions induced by p. Our example is the simplest illustration of the difficulties which cold be met along this road.

Exercise IV.14. [computational study] Consider situation as follows: at discrete time instants $t = 1, 2, ..., T$ we observe the states $y_t \in \mathbf{R}^\nu$ of dynamical system; our observations are

$$y_t + \sigma\xi_t, t = 1, 2, ..., T,$$

where $\sigma > 0$ is a given noise intensity and ξ_t are independent across t zero mean Gaussian noises with unit covariance matrix. All we know about the trajectory of the system is that

$$\|y_{t+1} - 2y_t + y_{t-1}\|_2 \leq dt^2\alpha, \tag{!}$$

where $dt > 0$ is the continuous time interval between consecutive discrete time instants; in other words, the Euclidean norm of the (finite-difference approximation of the) acceleration of the system is $\leq \alpha$. Given time delay d, we want to estimate the linear form $f^\top y_{T+d}$ of the system's state at time $T + d \geq 1$, and we intend to use a linear estimate

$$\widehat{y} = \sum_{t=1}^{T} h_t^\top \omega_t.$$

1. Write down optimization problem specifying the minimum risk linear estimate, with the risk of an estimate defined as

$$\mathrm{Risk}[\widehat{y}] = \sqrt{\sup_{y \in \mathcal{Y}} \mathbf{E}\{|\widehat{y} - f^\top y_{T+d}|^2\}},$$

where $\mathcal{Y}$ is the set of all trajectories $y = \{y_t, -\infty < t < \infty\}$ satisfying all constraints (!).

Solution: Let $E_t y = y_{t+1} - 2y_t + y_{t-1}$, so that (!) reads $\|E_t y\|_2 \leq \beta = dt^2\alpha$, $t = 0, \pm 1, \pm 2, ...$ For a linear estimate, we clearly have

$$\mathbf{E}\{|\widehat{y} - y_{T+d}|_2^2\} = |\sum_{t=1}^{T} h_t^\top y_t - f^\top y_{T+d}|^2 + \sigma^2 \sum_{t=1}^{T} \|h_t\|_2^2$$

Consequently,

$$\begin{array}{rcl} \mathrm{Risk}^2[\widehat{y}] & = & \sigma^2 \sum_{t=1}^{T} \|h_t\|_2^2 + \Phi^2(h), \\ \Phi(h) & := & \max\limits_{y: \|E_t y\|_2 \leq \beta \, \forall t} |\sum_{t=1}^{T} h_t^\top y_t - f^\top y_{T+d}| \\ & = & \max\limits_{y: \|E_t y\|_2 \leq \beta \, \forall t} [\sum_{t=1}^{T} h_t^\top y_t - f^\top y_{T+d}] \end{array}$$

where the concluding equality follows from the fact that trajectories y and $-y$ simultaneously satisfy/do not satisfy the acceleration bound. Next, when computing $\sup_y$, we clearly can restrict ourselves with $\sup_{y: \|E_t y\|_2 \leq \beta, 1 \leq t \leq \overline{T}}$, where $\overline{T} = \max[T, T + d]$, and in this case we lose nothing when

thinking of y as of finite sequence: $y = [y_0; y_1; ...; y_{\overline{T}+1}]$. Thus, the optimization problem responsible for the minimum risk linear estimate is

$$\mathrm{Opt} = \min_{h_1,...,h_T} \left[F(h) = \sigma^2 \sum_{t=1}^{T} \|h_t\|_2^2 + \left[\max_{\substack{y=[y_0;y_1;...;y_{\overline{T}+1}]: \\ \|E_t y\|_2 \le \beta, 1 \le t \le \overline{T}}} [\sum_{t=1}^{T} h_t^\top y_t - f^\top y_{T+d}] \right]^2 \right], \qquad (*)$$

This is a convex optimization problem with albeit implicitly given, but efficiently computable objective. The optimal value Opt of the problem is the squared risk of the minimum risk linear estimate.

2. Use Conic Duality to convert the problem from the previous item into a Conic Quadratic problem.

Solution: By Conic Duality,

$$\begin{array}{rcl} \Phi(h) & := & \max\limits_{y=[y_0;y_1;...;y_{\overline{T}+1}]} \left\{ G^\top[h]y := \sum_{t=1}^{T} h_t^\top y_t - f^\top y_{T+d} : \|E_t y\|_2 \le \beta, 1 \le t \le \overline{T} \right\} \\ & = & \min\limits_{\lambda_1,...,\lambda_{\overline{T}} \in \mathbf{R}^\kappa} \left\{ \beta \sum_{t=1}^{\overline{T}} \|\lambda_t\|_2 : \sum_{t=1}^{\overline{T}} E_t^\top \lambda_t = G[h], \right\} \end{array}$$

making $(*)$ the Conic Quadratic problem

$$\sqrt{\mathrm{Opt}} = \min_{\lambda_1,...,\lambda_{\overline{T}},h_1,...,h_T,r,s,\gamma} \left\{ \gamma : \begin{array}{rcl} \|[r;s]\|_2 & \le & \gamma \\ \|h_t\|_2 & \le & r_t/\sigma, 1 \le t \le T \\ \|\lambda_t\|_2 & \le & s_t/\beta, 1 \le t \le \overline{T} \\ \sum_{t=1}^{\overline{T}} E_t^\top \lambda_t & = & [0_{\nu\times 1}; h_1; ...; h_T; 0_{\nu(\overline{T}-T)\times 1}] \\ & & -[0_{\nu(T+d)\times 1}; f; 0_{\nu(\overline{T}-T-d)\times 1}] \end{array} \right\}$$

3. Carry out numerical experimentation with minimum risk linear estimate.

Solution: In our experiments, we used $\nu = 3$, $T = 100$, $dt = 0.25$, $\alpha = 1$, $\sigma = 0.1$, $d = 10$. We built linear estimates for every one of the 3 coordinates in y_{T+d}. The risks of these estimates and their empirical, over 300 simulations, risks are as follows:

3.73	3.73	3.73
1.80	1.75	1.76

Top row: Theoretical risk bounds; bottom row: empirical risks

The plot of a sample simulation is presented on Figure S2IV.3.

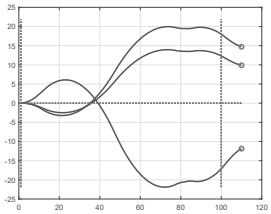


Figure S2IV.3. Coordinates of system's state vs. time. Circles: forecasts; dotted segment between dotted vertical lines: instants where observations are taken.

Exercise IV.15. [computational study] Consider the following problem:

> A particle is moving through $\mathbf{R}^d$. Given positions and velocities of the particle at times $t = 0$ and $t = 1$, find the trajectory of the particle on $[0, 1]$ with minimum possible (upper bound) on acceleration.

1. Formulate the (discretized in time version of the) problem as a Conic Quadratic problem and write down its conic dual. Are the problems solvable? Are the optimal values equal to each other? What is said by optimality conditions?
2. Run numerical experiments in 2D and 3D and look at the results.

Solution: 1: Let us discretize time interval $[0,1]$, splitting it into $n+1$ consecutive segments of length $dt = 1/(n+1)$ each, and set $t_i = i/(n+1)$. We discretize a candidate trajectory by looking at the sequence $u = \{u_1, ..., u_n\}$ of positions of the particle at time instants t_i, $i = 1, ..., n$, and augment this sequence with two initial terms, u_{-1}, u_0, and two concluding terms u_{n+1}, u_{n+2} to model the boundary conditions; specifically, we set u_0 to be the given position of the particle at time 0, and set $u_{-1} = u_0 - dtv^0$, where v^0 is the velocity of the particle at time 0. Similarly, we set u_{n+1} to be the given position of particle at time 1, and set $u_{n+2} = u_{n+1} + dtv^1$, where v^1 is the velocity of the particle at time 1. Finally, we approximate the acceleration of the particle at time t_i by the finite difference $[u_i - 2u_{i-1} + u_{i-2}]/dt^2$. As a result, the discretized model of our problem becomes

$$\mathrm{Opt}(P) = \min_{\tau,u} \left\{\tau : \|u_i - 2u_{i-1} + u_{i-2}\|_2 \le dt^2\tau, 1 \le i \le n+2\right\} \qquad (P)$$

Note that in this problem, the variables are $u_1, ..., u_n$, while $u_{-1}, u_0, u_{n+1}, u_{n+2}$ are data.
(P) is a Conic Quadratic problem; its "canonical" form is

$$\min_{\tau,u} \left\{\tau : [2u_i - u_{i-1} + u_{i+2}; dt^2\tau] \in \mathbf{L}^{d+1},\ 1 \le i \le n+2\right\}$$

Equipping the conic constraints with Lagrange multipliers $[y_i; s_i] \in \mathbf{L}^{d+1}$, $1 \le i \le n+2$, the conic dual of (P) is built as follows: we aggregate the constraints with the "weights" $[y_i; s_i]$, thus arriving at the relation

$$\sum_{i=1}^{n} [dt^2\tau s_i + y_i^\top [u_i - 2u_{i-1} + u_{i=2}]] \ge 0$$

which, due to its origin, is a consequence of the constraints of (P), rewrite this relation equivalently as

$$[\text{homogeneous linear function of } \tau, u_1, ..., u_n] \ge [\text{linear function of } y_i, s_i,\ 1 \le i \le n+2] \qquad (*)$$

and impose on the Lagrange multipliers, in addition to the constrains $[y_i; s_i] \in \mathbf{L}^{d+1}$, the restriction that the left hand side linear function in $(*)$ is identically in $\tau, u_1, ..., u_n$ equal to the objective of (P). The dual problem is to maximize under these restrictions the right hand side of $(*)$.
Executing this strategy (which is a fully mechanical process) results in the dual problem

$$\begin{array}{rcl}\mathrm{Opt}(D) & = & \max_{y_i,s_i} \left\{ y_1^\top [2u_0 - u_{-1}] - y_2^\top u_0 - y_{n+1}^\top u_{n+1} + y_{n+2}^\top [2u_{n+1} - u_{n+2}] : \right. \\ & & \left. \left\{ \begin{array}{ll} y_{i+2} - 2y_{i+1} - y_i = 0 & , i = 1, ..., n \quad (a) \\ \sum_{i=1}^{n+2} s_i = 1/dt^2 & , \quad (b) \\ [y_i; s_i] \in \mathbf{L}^{d+1} & , 1 \le i \le n+2 \quad (c) \end{array} \right. \right\} \end{array} \qquad (D)$$

Clearly, (P) is strictly feasible and bounded, implying that (D) is solvable and that the optimal values are equal to each other. Besides this, (D) clearly is essentially strictly feasible, so that (P) is solvable as well. Optimality conditions in their complementary slackness form say that a pair $(u_i^*, i \le n, \tau^*; y_i^*, s_i^*, i \le n+2)$ of primal-dual feasible solutions is composed of optimal solutions iff

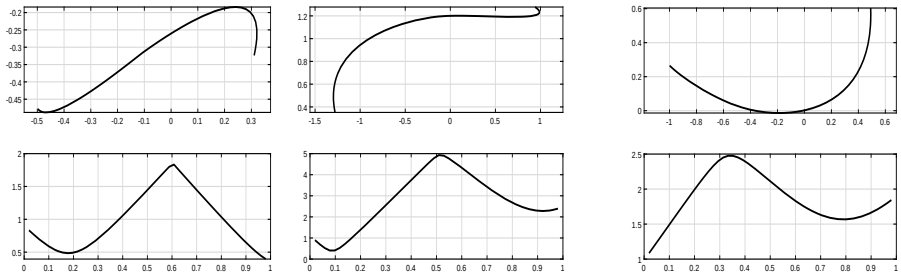
$$[u_i^* - 2u_{i-1}^* + u_{i-2}^*; dt^2\tau^*]^\top [y_i^*; s_i^*] = 0,\ 1 \le i \le n+2. \qquad (**)$$

Observe that the equality constraints (a) in (D) say that entries in y_i are linear functions of i: $y_i = g + (i-1)h$ for some g, h. As a result, (D) simplifies to

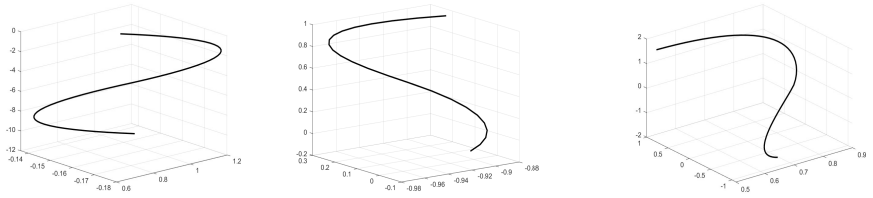
$$\begin{array}{rl}\mathrm{Opt} = \min_{g,h,s_i} \Big\{ & [-u_{-1} + u_0 + u_{n+1} - u_{n+2}]^\top g + [-u_0 + (n+2)u_{n+1} - (n+1)u_{n+2}]^\top h : \\ & \|g + (i-1)h\|_2 \le s_i, 1 \le i \le n+2, \textstyle\sum_i s_i = 1/dt^2 \Big\}, \end{array} \qquad (D')$$

and $\mathrm{Opt} = \mathrm{Opt}(P) = \mathrm{Opt}(D)$. This combines with complementary slackness $(**)$ to conclude that in the only nontrivial case $\mathrm{Opt} > 0$ an optimal solution to (P) is readily given by an optimal solution $g^*, h^*, s_i^*, i \le n+2$ to (D') via the relation

$$u_i^* - 2u_{i-1}^* + u_{i-2}^* = -dt^2 \mathrm{Opt} \frac{g^* + (i-1)h^*}{\|g^* + (i-1)h^*\|_2} \qquad (!)$$



2D trajectoris (top) and magnitudes of velocity vs. time (bottom)



3D trajetories

Figure S2IV.4. Sample trajectories (bold) in Exercise IV.15 and their 2D projections (dotted).

which holds true for all i, $1 \le i \le n+2$, such that $g^* + (i-1)h^* \neq 0$; in this relation, by definition $u^*_{-1} = u_{-1}$, $u^*_0 = u_0$, $u^*_{n+1} = u_{n+1}$, $u^*_{n+2} = u_{n+2}$. Note that if there is no $i \le n+2$ such that $g^* + (i-1)h^* = 0$, recurrence (!) fully determines u^*_i, $1 \le i \le n$. If $g^* + (i-1)h^* = 0$ (if such an $i = i_*$ exists, it is unique, since otherwise we would have $g^* = h^* = 0$ and therefore Opt $= 0$, which is not the case), we could specify u^*_i for $1 \le i < i_*$ via recurrence (!), and for $i = i_*, i_*+1, ..., n$ - running the same recurrence backward, starting with $i = n+2$.

2: In or computations, we used $n = 50$. Sample trajectories in 2D and 3D are shown at Figure S2IV.4.

Exercise IV.16. [computational study] The study offered to you in this Exercise is as follows:

> *A steel rod is heated at time $t = 0$, the magnitude of the temperature being $\le R$, and is left to cool, the temperature at the endpoints being all the time kept 0. We measure the temperature of the rod at locations s_i and times $t_i > 0$, $1 \le i \le m$; the measurements are affected by Gaussian noise with zero mean and covariance matrix $\sigma^2 I_m$. Given the measurements, we want to recover the distribution of temperature of the rod at time $\bar{t} > 0$.*

Building the model. With properly selected units of temperature and length (so that the rod becomes the segment $[0,1]$), evolution of the temperature $u(t,s)$ ($t \ge 0$ is time, $s \in [0,1]$ is location) is governed by the *Heat equation*

$$\frac{\partial}{\partial t}u(t,s) = \frac{\partial^2}{\partial s^2}u(t,s) \qquad [u(t,0) = u(t,1) \equiv 0]$$

It is convenient to represent functions on $[0,1]$ as

$$f(s) = \sum_{k=1}^{\infty} f_k \phi_k(s),\ \phi_k(s) = \sqrt{2}\sin(\pi k s).$$

Functions ϕ_k form an orthonormal basis in the space $L_2 = L_2[0,1]$ of square summable real-valued functions on $[0,1]$ equipped with the inner product

$$\langle f, g\rangle = \int_0^1 f(s)g(s)ds,$$

the corresponding norm being $\|f\|_2 = \sqrt{\int_0^1 f^2(s)ds}$.

Functions ϕ_k form an orthonormal basis in L_2, meaning that for every $f \in L_2$ the series

$$\sum_{k=1}^{\infty} f_k \phi_k(s),\ f_k = \langle f, \phi_k \rangle$$

converges in $\|\cdot\|_2$ to f, $f \in L_2$ if and only if $\sum_k f_k^2 < \infty$, and

$$\langle \sum_k f_k \phi_k(\cdot), \sum_k g_k \phi_k(\cdot) \rangle = \sum_k f_k g_k \ \forall f, g \in L_2.$$

In particular,

$$u(t,s) = \sum_{k=1}^{\infty} u_k(t)\phi_k(s),\ u_k(t) = \int_0^1 u(t,s)\phi_k(s)ds.$$

Assuming $|u(0,\cdot)| \le R$, we have

$$\sum_k u_k^2(0) \le R^2, \tag{23.1}$$

and in terms of the coefficients $u_k(t)$ of the rod's temperature, the Heat equation becomes very simple:

$$\frac{d}{dt}u_k(t) = -\pi^2 k^2 u_k(t) \implies u_k(t) = \exp\{-\pi^2 k^2 t\} u_k(0).$$

As a result, when $t > 0$, the coefficients $u_k(t)$ go to 0 exponentially fast as $k \to \infty$, so that the series

$$\sum_k u_k(t)\phi_k(s)$$

converges to the solution (t,s) of the heat equation not only in $\|\cdot\|_2$, but uniformly on $[0,1]$ as well, implying, due to $\phi_k(0) = \phi_k(1) = 0$, that the series does satisfy the boundary conditions $u(t,0) = u(t,1) = 0$, $t > 0$.

Now our problem can be posed as follows:

The sequence of coefficients $\{u_k^t\}_{k=1}^{\infty}$ of $u(t,\cdot)$ in the orthonormal basis $\{\phi_k(\cdot)\}_{k\ge1}$ of L_2 evolves according to

$$u_k^t = \exp\{-\pi^2 k^2 t\} u_k^0,$$

with

$$u^0 := \{u_k^0\}_{k\ge1} \in \mathbf{B} := \{\{c_k\}_{k\ge1} : \sum_k c_k^2 \le R^2\}.$$

Given m noisy observations

$$\omega_i = \Omega_i[u^0] + \sigma\xi_i,\ \Omega_i[u^0] = \sum_{k=1}^{\infty} \exp\{-\pi^2 k^2 t_i\} u_k^0 \phi_k(s_i),$$

where $\xi_1, ..., \xi_m$ are independent of each other $\mathcal{N}(0,1)$ observation noises, and $t_i > 0$, $s_i \in [0,1]$ are given, we want to recover the sequence $\{u_k^{\bar t}\}_{k\ge1}$.
We quantify the performance of a candidate estimate $\omega := (\omega_1, ..., \omega_m) \mapsto \widehat{u} = \{\widehat{u}_k(\omega)\}_{k\ge1}$ by the risk

$$\mathrm{Risk}[\widehat{u}] = \sqrt{\max_{u^0 \in \mathbf{B}} \mathbf{E}_\xi \left\{ \sum_{k\ge1} [\widehat{u}_k(\Omega_1[u^0] + \xi_1, ..., \Omega_m[u^0] + \xi_m) - \exp\{-\pi^2 k^2 \bar t\} u_k^0]^2 \right\}}$$

that is, Risk^2 is the worst, w.r.t. the distribution of temperature at time $t = 0$ of $\|\cdot\|_2$-norm not exceeding R, expected squared norm $\|\cdot\|_2^2$ of the recovery error.

Our last modeling step is to replace infinite sequences $\{u_k^0\}_{k\geq 1}$ with their finite initial segments $\{u_k^0\}_{1\leq k\leq K}$, that is, to approximate the situation by the one where $u_0^k = 0$ when $k > K$. The simplest way to do it is as follows. Let $\underline{t} = \min[\min_i t_i, \bar{t}]]$, so that $\underline{t} > 0$. For $u_0 \in \mathbf{B}$ and $K \geq 1$, the magnitude of the total contribution of the coefficients u_0^k, $k > K$, to $u(t,s)$ with $t \geq \underline{t}$ does not exceed

$$\sum_{k=K+1}^{\infty} \max_s |\phi_k(s)| \exp\{-\pi^2 k^2 \underline{t}\} |u_0^k| \leq \delta := \sqrt{2}R \sum_{k=K+1}^{\infty} \exp\{-\pi^2 k^2 \underline{t}\}.$$

Given a "really small" tolerance $\bar{\delta} > 0$, say, $\bar{\delta} = 10^{-10}$, we can easily find $K = K(\bar{\delta})$ such that $\delta \leq \bar{\delta}$. Thus, as far as the temperatures we measure and the temperatures we want to recover are concerned, zeroing out coefficients u_0^k with $k > K(\bar{\delta})$ changes these temperatures by at most $\bar{\delta}$. Common sense (which can be easily justified by formal analysis) says, that with $\bar{\delta}$ as small as 10^{-10}, these changes have no effect on the quality of our recovery, at least when $\sigma \gg \bar{\delta}$.

Now goes your task:

1. Assuming $u_0^k = 0$ for $k > K$, model the problem of interest as the following estimation problem:

   "In the nature" there exists K-dimensional signal u known to belong to the centered at the origin Euclidean ball $B^R = \{u \in \mathbf{R}^K : u^\top u \leq R^2\}$ of a given radius R. Given noisy observations

   $$\omega = Au + \sigma\xi, \qquad [A : m \times K, \xi \sim \mathcal{N}(0, I_m)]$$

   we want to recover Bu, quantifying the recovery error of a candidate estimate $\omega \mapsto \widehat{u}(\omega)$ by its risk

   $$\mathrm{Risk2}[\widehat{u}] = \sqrt{\sup_{u\in B^R} \mathbf{E}_{\xi\sim\mathcal{N}(0,I_m)} \{[\widehat{u}(Au+\sigma\xi) - Bu]^\top [\widehat{u}(Au+\sigma\xi) - Bu]\}}$$

   where B is a given $K \times K$ matrix.

   Write down the expressions for the matrices A and B.
2. Build convex optimization problem responsible for the minimum risk *linear estimate* – estimate of the form $\widehat{u}(\omega) = H^\top \omega$.
3. Compute the minimum risk linear estimate and run simulations to test its performance. Recommended setup:
   - $\bar{t} \in \{0.01, 0.001, 0.0001, 0.00001\}$
   - $m = 100$, t_i are drawn at random from the uniform distribution on $[\bar{t}, 2\bar{t}]$, s_i are drawn at random from the uniform distribution on $[0, 1]$;
   - $R = 10^4$, $\sigma = 10$, $\bar{\delta} = 10^{-10}$;
   - To accelerate computations, truncate $K(\bar{\delta})$ at the level 100.

Solution: 1: $A_{ik} = \sqrt{2}\exp\{-\pi^2 k^2 t_i\}\sin(\pi k s_i)$, $1 \leq i \leq m, 1 \leq k \leq K$. B is diagonal $K \times K$ matrix with diagonal entries $\exp\{-\pi^2 k^2 \bar{t}\}$, $1 \leq k \leq K$.

2: The problem is

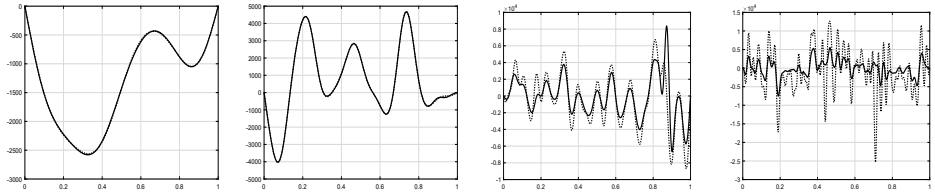
$$\mathrm{Opt} = \min_{H\in\mathbf{R}^{m\times K}} \sqrt{R^2 \|B - H^\top A\|_{2,2}^2 + \sigma^2 \mathrm{Tr}(H^\top H)},$$

where $\|\cdot\|_{2,2}$ is the spectral norm (the largest singular value) of a matrix.

3: Our results are as follows:

			empirical errors		
$\bar{t}$	K	Risk2	mean	median	max
0.01	18	6.47	6.35	11.24	7.17
0.001	58	15.60	14.06	13.84	22.71
0.0001	100	1186.8	1154.3	1638.9	2192.3
0.00001	100	4332.2	4332.2	4762.9	9294.7

Risks and empirical recovery errors,
data over 100 simulations



$\bar{t} = 0.01$ $\bar{t} = 0.001$ $\bar{t} = 0.0001$ $\bar{t} = 0.00001$

Figure S2IV.5. Sample recoveries: dotted line - true; solid line – recovery

To put the results in proper perspective, pay attention to the range of true signals on Figure S2IV.5.

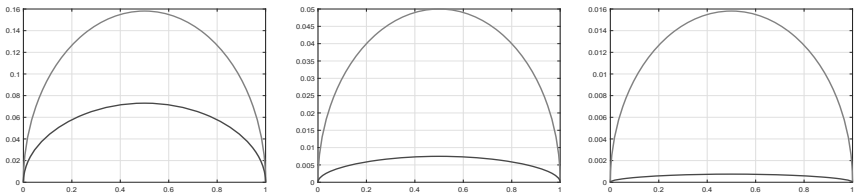
Exercise IV.17. Given positive definite $A \in \mathbf{S}^n$, let us set

$$P[A] = \{X \in \mathbf{S}^n : X \succeq 0, X^2 \preceq A\},\ Q[A] = \{X \in \mathbf{S}^n : X \succeq 0, X \preceq A^{1/2}\}.$$

From $\succeq$-monotonicity of the matrix square root on $\mathbf{S}^n_+$ (Example IV.20.5 in section 20.2) it follows that $P[A] \subseteq Q[A]$. Your task is to answer the following question:

> Are $P[A]$ and $Q[A]$ "comparable," meaning that for some c independent of A (but perhaps depending on n) one has
>
> $$Q[A] \subset c \cdot P[A] \qquad ?$$

Solution: The answer is negative, unless $n = 1$. To justify the claim, it suffices to consider the case of $n = 2$. Given $\epsilon \in (0, 1)$, let us set $A = \left[\begin{array}{c|c} \epsilon & \\ \hline & 1 \end{array}\right]$, so that $A^{1/2} = \left[\begin{array}{c|c} \epsilon^{1/2} & \\ \hline & 1 \end{array}\right]$, implying that the matrix $\overline{X} = \frac{1}{2}\left[\begin{array}{c|c} \epsilon^{1/2} & \epsilon^{1/4} \\ \hline \epsilon^{1/4} & 1 \end{array}\right]$ belongs to $Q[A]$. On the other hand, for $X = [x_{ij}]_{i,j=1,2} \in P[A]$ we should have $[X^2]_{1,1} = X_{1,1}^2 + X_{1,2}^2 \le A_{1,1} = \epsilon$, that is, $X_{1,2} \le \sqrt{\epsilon}$. By looking at off-diagonal entries, we conclude that if $Q[A] \subset c \cdot P[A]$, so that $\overline{X} = cX$ for some $X \in P[A]$, we should have $c \ge \frac{1}{2}\epsilon^{-1/4}$, and the right hand side here tends to $+\infty$ when $\epsilon \to +0$. ■



$\epsilon = 10^{-2}$ $\epsilon = 10^{-4}$ $\epsilon = 10^{-6}$

Figure S2IV.6. Results for Exercise IV.17

For a given ϵ, setting $A = \left[\begin{array}{c|c} \epsilon & \\ \hline & 1 \end{array}\right]$, the matrices from $P[A]$ are of the form $\left[\begin{array}{c|c} t\sqrt{\epsilon} & \delta \\ \hline \delta & s \end{array}\right]$ with $0 \le t, s \le 1$ and $-\gamma_\epsilon(t, s) \le \delta \le \gamma_\epsilon(t, s)$, and the matrices from $Q[A]$ are of the form $\left[\begin{array}{c|c} t\sqrt{\epsilon} & \delta \\ \hline \delta & s \end{array}\right]$ with $0 \le t, s \le 1$ and $-\theta_\epsilon(t, s) \le \delta \le \theta_\epsilon(t, s)$. What you see on Figure S2IV.6 are the plots of the functions $\gamma_\epsilon(t, 1-t)$ (lower curves) and $\theta_\epsilon(t, 1-t)$ (upper curves) vs. $t \in [0, 1]$.

Exercise IV.18. Find the optimal value in the convex optimization problem

$$\mathrm{Opt}(a) = \min_x \left\{\sum\nolimits_{i=1}^n [-(1 + a_i)x_i + x_i \ln x_i] : x \ge 0, \sum\nolimits_i x_i \le 1\right\}$$

where $0 \ln 0 = 0$ by definition, so that the function $x \ln x$ is well defined and continuous on the nonnegative ray $x \ge 0$.

Solution: The problem is of the form $\min_{x\in X} f(x)$ with $X = \{x \in \mathbf{R}^n : x \geq 0, \sum_i x_i \leq 1\}$ and

$$f(x) = \sum\nolimits_i [-(1+a_i)x_i + x_i \ln x_i];$$

f is convex on X and is differentiable on the part $x > 0$ of X. Let us make an educated guess that there is an optimal solution x to the problem with $x > 0$. The necessary and sufficient condition for such a solution x to be optimal is that

1. either $x \in \operatorname{int} X$ and $\nabla f(x) = 0$, resulting in

   $$x_i = \exp\{a_i\}, 1 \leq i \leq n;$$

   this solution indeed belongs to the interior of X provided that

   $$\sum\nolimits_{i=1}^n \exp\{a_i\} < 1 \qquad (!)$$

2. or $x > 0$ belongs to the face $\sum_i x_i = 1$ of X, in which case x is optimal if and only if $\nabla f(x)$ has nonnegative inner products with all directions leading from x to points of X, or, which is the same, $\nabla f(x)$ is a nonpositive multiple of the all-ones vector. Thus, what we want of x now is to be positive, to have $\sum_i x = 1$, and to have

   $$\nabla f(x) = -\lambda[1; \ldots; 1]$$

   with some $\lambda \geq 0$. Looking at what $\nabla f(x)$ is, this boils down to

   $$x_i = \exp\{a_i - \lambda\}$$

   and $\sum_i x_i = 1$, implying

   $$\lambda = \ln(\sum\nolimits_j \exp\{a_j\}).$$

   Thus, $\lambda \geq 0$ whenever (!) fails to be true, and in this case

   $$x_i = \frac{\exp\{a_i\}}{\sum_j \exp\{a_j\}}$$

The bottom line is that an optimal solution is given by

$$x_i = \frac{\exp\{a_i\}}{\max[1, \sum_j \exp\{a_j\}]}, \; 1 \leq i \leq n,$$

and the optimal value is

$$\begin{cases} -\sum_i \exp\{a_i\}, & \sum_i \exp\{a_i\} \leq 1, \\ -\ln\left(\sum_i \exp\{a_i\}\right) - 1, & \text{otherwise} \end{cases}$$

Exercise IV.19. Given $m \times n$ matrix A with trivial kernel, consider the matrix-valued function $F(X) = [A^T X^{-1} A]^{-1} : \operatorname{Dom} F := \{X \in \mathbf{S}^m, X \succ 0\} \to \mathbf{S}^n_+$. Prove that F is $\succeq$-concave on its domain.

Solution: We should prove that the $\succeq$-hypograph of F – the set

$$\mathcal{E} = \{X, Y : X \in \mathbf{S}^m, X \succ 0, Y \in \mathbf{S}^n, Y \preceq F(X)\}$$

is convex. To this end it suffices to show that the set

$$\mathcal{F} = \{X, Y : X \in \mathbf{S}^m, X \succ 0, 0 \prec Y \preceq F(X)\}$$

is convex, since $\mathcal{E}$ is the sum of $\mathcal{F}$ and the convex set $\{0_{m\times m}\} \times [-\mathbf{S}^n_+]$. We have

$$\begin{array}{rl} & Y \succ 0 \;\&\; X \succ 0 \;\&\; Y \preceq [A^T X^{-1} A]^{-1} \\ \Longleftrightarrow & Y \succ 0 \;\&\; X \succ 0 \;\&\; A^T X^{-1} A \preceq Y^{-1} \text{ [Exercise D.5]} \\ \Longleftrightarrow & Y \succ 0 \;\&\; X \succ 0 \;\&\; \left[\begin{array}{c|c} Y^{-1} & A^T \\ \hline A & X \end{array}\right] \succeq 0 \text{ [Schur Complement Lemma]} \\ \Longleftrightarrow & Y \succ 0 \;\&\; X \succ 0 \;\& X \succeq AYA^T \text{ [Schur Complement Lemma]} \end{array}$$

and the resulting description of $\mathcal{F}$ clearly states that $\mathcal{F}$ is convex. ■

Note: the result we have just proved is the special case of the one stated in Exercise III.2.5.

Exercise IV.20. [cone-constrained semidefinite problems]

1. Let $X, Y \in \mathbf{S}^m_+$. Prove that $\mathrm{Tr}(XY) = 0$ is and only if $XY = YX = 0$.
2. Given an ordered collection $\nu = \{n_1, ..., n_k\}$ of positive integers, let $\mathbf{S}^\nu$ be the space of block-diagonal symmetric matrices with k diagonal blocks of sizes $n_1 \times n_1$,...,$n_k \times n_k$, and let $\mathbf{S}^\nu_+$ be the cone of positive semidefinite matrices from $\mathbf{S}^\nu$. Equipping $\mathbf{S}^\nu$ with the Frobenius inner product, $\mathbf{S}^\nu_+$ clearly is a self-dual regular cone in the resulting Euclidean space.
   Convex cone-constrained problem on the cone $\mathbf{S}^\nu_+$ is of the generic form

$$\mathrm{Opt(SDP)} = \min_{x \in X} \left\{ f(x) : \overline{g}(x) := Ax - b \leq 0,\, \widehat{g}(x) := \mathrm{Diag}\{g_1(x), ..., g_k(x)\} \leq_{\mathbf{S}^\nu_+} 0 \right\} \tag{SDP}$$

   where X is a nonempty convex set in some $\mathbf{R}^n$, the function $f : X \to \mathbf{R}$ is convex, and the mapping $\widehat{g} : X \to \mathbf{S}^\nu$ is $\mathbf{S}^\nu_+$-convex.
   Prove that in the case of convex cone-constrained semidefinite problem (SDP) Theorem IV.19.7 reads

> **Theorem IV.19.7.SDP** *Consider a convex cone-constrained semidefinite problem* (SDP), *let $x^* \in X$ be a feasible solution to the problem, and let f and $\widehat{g}$ be differentiable at x^*.*
> (i) *If x^* is a KKT point of* (SDP), *the Lagrange multipliers being $\overline{\lambda}^* \geq 0$ and $\widehat{\lambda}^* \in \mathbf{S}^\nu_+$, meaning that*
>
> $$\overline{\lambda}^*_i [\overline{g}(x^*)]_i = 0 \,\forall i \;\&\; \widehat{\lambda}^* \widehat{g}(x^*) = 0 \qquad \text{[sdp complementary slackness]}$$
> $$\nabla_x \left[f(x) + [\overline{\lambda}^*]^\top \overline{g}(x) + \mathrm{Tr}(\widehat{\lambda}^* \widehat{g}(x)) \right] \big|_{x=x^*} \in -N_X(x^*) \qquad \text{[KKT equation]}$$
>
> *(here, as always, $N_X(x)$ is the normal cone of X, see (11.5)), then x^* is an optimal solution to* (SDP).
> (ii) *If x^* is optimal solution to* (SDP) *and, if addition to the above premise,* (SDP) *satisfies the cone-constrained Relaxed Slater condition, then x^* is an sdp KKT point, as defined in item* (i).

Solution: 1. This claim is Exercise IV.3.2.

2. Straightforward application of Theorem IV.19.7 to the convex cone-constrained problem (SDP) differs from Theorem IV.19.7.SDP in the only point: in the complementary slackness part of the former Theorem, the $\widehat{g}$-related equality reads $\mathrm{Tr}(\widehat{\lambda}^* \widehat{g}(x^*)) = 0$, while in Theorem IV.19.7.SDP the corresponding equality is $\widehat{\lambda}^* \widehat{g}_j(x^*) = 0$. Taking into account that we are in the case $\widehat{\lambda}^* \succeq 0$,, $\widehat{g}(x^*) \preceq 0$ and invoking item 1 of Exercise, in the situation in question both equalities are satisfied/not satisfied simultaneously.

Exercise IV.21. [follow-up to Exercise IV.20] In the sequel, we fix the dimension n of the embedding space and denote by $E_C = \{x \in \mathbf{R}^n : x^\top C x \leq 1\}$ the centered at the origin ellipsoid associated with positive definite $n \times n$ matrix C. Given positive K and K ellipsoids E_{A_k}, $k \leq K$, consider two optimization problems:

— $\mathcal{O}$: find the smallest volume centered at the origin ellipsoid containing $\cup_{k \leq K} E_{A_k}$,
— $\mathcal{I}$: find the largest volume centered at the origin ellipsoid contained in $\cap_{k \leq K} E_{A_k}$.

1. Pose $\mathcal{O}$ as a solvable convex cone-constrained semidefinite program
2. Prove that problems $\mathcal{O}$ and $\mathcal{I}$ reduce to each other at the cost of appropriate modification of the data
3. Prove that there exist matrices $\Lambda_k \succeq 0$ such that $\Lambda := \sum_k \Lambda_k \succ 0$ and

$$\Lambda_k = \Lambda_k A_k \Lambda,\; k \leq K.$$

Solution: Let X be the set of positive definite $n \times n$ matrices, and $\nu = \{\underbrace{n, ..., n}_{K}\}$.

1: By the result of Exercise I.14.2, we have $E_P \subset E_Q$ if and only if $P \succeq Q$. Specifying a candidate

solution to $\mathcal{O}$ as E_U, the constraint $E_U \supset \cup_k E_{A_k}$, that is, $E_{A_k} \subset E_U$ for all k, becomes $U \preceq A_k$ and $U \in X$. By the result of Exercise I.14.3, $\mathrm{Vol}(E_U) = \mathrm{Det}^{-1/2}(U)$, so that $\mathcal{O}$ can be posed as the optimization problem

$$\min_{U\in X}\left\{-\ln\mathrm{Det}(U) : \widehat{g}(U) := \mathrm{Diag}\{U-A_1,...,U-A_K\} \leq_{\mathbf{S}^\nu_+} 0\right\} \qquad (\mathcal{O})$$

By Fact III.14.6 applied to the convex symmetric function $g(t) = -\sum_i \ln t_i : \mathrm{int}\,\mathbf{R}^n_+ \to \mathbf{R}$, the objective in $(\mathcal{O})$ is convex on X; thus, $(\mathcal{O})$ is a convex cone-constrained semidefinite program,

It is immediately seen that the problem is solvable. Indeed, its feasible set F is nonempty and bounded, and the sublevel sets $\{U \in F : -\ln\mathrm{Det}(U) \leq a\}$ of the objective on this set clearly are closed, so that on the feasible set the objective attains its minimum.

2: As we know from Example II.6.12, $\mathrm{Polar}\,(E_P) = E_{P^{-1}}$, and from Exercise II.38

$$\mathrm{Vol}(\mathrm{Polar}\,(E_P)) = 1/\mathrm{Vol}(E_P).$$

Besides this, passing to polars reverses inclusions. It follows that an ellipsoid E_U is a feasible solution to problem $\mathcal{O}$ with data $A_1,...,A_K$ if and only if the ellipsoid $E_{U^{-1}}$ is a feasible solution to problem $\mathcal{I}$ with the data $A_1^{-1},...,A_K^{-1}$, and the volumes Vol of these two ellipsoids are reciprocals of each other. The bottom line is that problem $\mathcal{O}$ with data $A_1,...,A_K$ reduces straightforwardly to problem $\mathcal{I}$ with data $A_1^{-1},...,A_K^{-1}$, and vice versa.

3: The objective $-\ln\mathrm{Det}(U)$ in the cone-constrained Problem $(\mathcal{O})$ is differentiable everywhere on X with the gradient $-U^{-1}$, see Example C.9. Besides this, $(\mathcal{O})$ is strictly feasible due to $A_k \succ 0$, $k \leq K$. Let U_* be an optimal solution to the problem (it exists by item 1). The cone-constrained Lagrange function of $(\mathcal{O})$ is

$$-\ln\mathrm{Det}(U) + \sum_{k=1}^K \mathrm{Tr}(\Lambda_k[U-A_k]).$$

Invoking Theorem IV.19.7.SDP.ii and taking into account that $U_* \in \mathrm{int}\,X = X$, there exist Lagrange multipliers $\Lambda_k \in \mathbf{S}^n_+$, $k \leq K$, such that

$$-U_*^{-1} + \sum_k \Lambda_k = 0 \;\&\; \Lambda_k(U_*-A_k) = 0,\ k \leq K.$$

Augmenting $\Lambda_1,...,\lambda_K$ with $\Lambda := \sum_k \Lambda_k = U_*^{-1} \succ 0$, we meet the requirements from item 3. ■

Exercise IV.22. Recall convex cone-constrained problem in Example IV.18.1, section 18.1

$$\mathrm{Opt}(P) = \min_{x=(t,y)\in\mathbf{R}\times\mathbf{S}^n}\left\{t : \underbrace{t \geq \mathrm{Tr}(y)}_{\Longleftrightarrow\ \langle y, I_n\rangle - t \leq 0},\ y^2 \preceq B\right\} \qquad (18.1)$$

where B is a positive definite matrix.

1. Verify (18.2)

Solution: We have

$$\begin{array}{rl}
& L(t,y;\lambda,\Lambda) = t + \overline{\lambda}[\mathrm{Tr}(y)-t] + \mathrm{Tr}(\widehat{\lambda}[y^2-B]) : [\mathbf{R}\times\mathbf{S}^n]\times[\mathbf{R}_+\times\mathbf{S}^n_+] \to \mathbf{R}\\
\Longrightarrow & \underline{L}(\overline{\lambda},\widehat{\lambda}) := \inf_{t\in\mathbf{R},y\in\mathbf{S}^n}\left[t + \overline{\lambda}[\mathrm{Tr}(y)-t] + \mathrm{Tr}(\widehat{\lambda}[y^2-B])\right]\\
& = \begin{cases} -\infty & ,\overline{\lambda}\neq 1\\ \inf_{y\in\mathbf{S}^n}\left[\mathrm{Tr}(y) + \mathrm{Tr}(\widehat{\lambda}[y^2-B])\right] & ,\overline{\lambda}=1\end{cases}\\
& = \begin{cases} -\infty & ,\overline{\lambda}\neq 1\\ -\infty & ,\overline{\lambda}=1\ \&\ \widehat{\lambda}\in\mathrm{bd}\,\mathbf{S}^n_+\\ -\frac{1}{4}\mathrm{Tr}(\widehat{\lambda}^{-1}) - \mathrm{Tr}(\widehat{\lambda}B) & ,\overline{\lambda}=1\ \&\ \widehat{\lambda}\in\mathrm{int}\,\mathbf{S}^n_+\end{cases}
\end{array}$$

where the concluding equality stems from the following observations:

- when $\widehat{\lambda} \in \mathrm{bd}\,\mathbf{S}^n_+$, $\widehat{\lambda}$ has a nontrivial kernel; taking as f a nonzero vector from this kernel and setting $y(s) = -sff^\top$, we get

$$L(\mathrm{Tr}(y(s)), y(s); 1, \widehat{\lambda}) = -sf^\top f + \mathrm{Tr}(s^2[f^\top f]\underbrace{\widehat{\lambda}ff^\top}_{=0} - \widehat{\lambda}B) = -sf^\top f - \mathrm{Tr}(\widehat{\lambda}B) \to -\infty,\ s \to \infty,$$

  that is, $\underline{L}(1, \widehat{\lambda}) = -\infty$;
- when $\widehat{\lambda} \succ 0$, the minimum in t, y of the convex in (t, y) function

$$L(t, y; 1, \widehat{\lambda}) = \mathrm{Tr}(y) + \mathrm{Tr}(\widehat{\lambda}[y^2 - B])$$

  can be found from the Fermat equation

$$0 = \nabla_y \left[\mathrm{Tr}(y) + \mathrm{Tr}(\widehat{\lambda}[y^2 - B])\right] = I_n + \widehat{\lambda}y + y\widehat{\lambda}$$

  resulting in $y = -\frac{1}{2}\widehat{\lambda}^{-1}$ and

$$\underline{L}(1, \widehat{\lambda}) = -\frac{1}{4}\mathrm{Tr}(\widehat{\lambda}^{-1}) - \mathrm{Tr}(\widehat{\lambda}B),$$

  as claimed in (18.2). ■

2. Find Lagrange multipliers certifying that $t_* = -\mathrm{Tr}(B^{1/2})$, $y_* = -B^{1/2}$ is a cone-constrained KKT point of problem (18.1) (and thus, by Theorem IV.19.7, is an optimal solution to the problem).

Solution: As it should be, the Lagrange multipliers, if any, certifying that (t_*, y_*) is a KKT point of (18.1) should be an optimal solution to the cone-constrained Lagrange dual (18.3) of (18.1). This solution

$$\overline{\lambda} = 1, \widehat{\lambda} = \frac{1}{2}B^{-1/2}$$

was found in section 18.3, and augmenting (t_*, y_*) with these Lagrange multipliers, we clearly meet the complementary slackness

$$\overline{\lambda}[t_* - \mathrm{Tr}(y_*)] = 0,\ \mathrm{Tr}(\widehat{\lambda}[y_*^2 - B]) = 0$$

and the KKT equation

$$\nabla_t L(t, y; \overline{\lambda}, \widehat{\lambda})\Big|_{t=t_*, y=y_*} = 0,\ \nabla_y L(t, y; \overline{\lambda}, \widehat{\lambda})\Big|_{t=t_*, y=y_*} = I_n + [y_*\widehat{\lambda} + \widehat{\lambda}y_*] = 0$$

3. Consider the parametric family

$$\mathrm{Opt}(p := (v, w)) = \min_{t \in \mathbf{R}, y \in \mathbf{S}^n} \left\{t : t \geq \mathrm{Tr}(y), yv^{-1}y \preceq w\right\} \qquad (\mathrm{P}[p])$$

   of convex cone-constrained problems, with $p \in P = \{p = (v, w) : v \in \mathrm{int}\,\mathbf{S}^n_+, w \in \mathrm{int}\,\mathbf{S}^n_+\}$, so that (18.1) is problem $(\mathrm{P}[\overline{p}])$ corresponding to

$$\overline{p} = (I_n, B).$$

   Prove that $\mathrm{Opt}(p)$ is convex function of $p \in P$ and find a subgradient of this function at the point $\overline{p}$.

Solution: Observe that setting $x = (t, y)$, the scalar function $\overline{g}(x, p) = \mathrm{Tr}(y) - t$ clearly is $\mathbf{R}_+$-convex, and the $\mathbf{S}^n$-valued function $\widehat{g}(x, p) = yv^{-1}y - w$ is $\succeq$-convex in $(x, p) \in [\mathbf{R} \times \mathbf{S}^n] \times P$; indeed, by Fact IV.20.1 $\succeq$-convexity of $\widehat{g}(x, p)$ on the indicated domain is the same as the convexity of the $\succeq$-epigraph of $\widehat{g}$, and this epigraph is, by the Schur Complement Lemma, the set

$$\{(t, y, v, w, z) \in \left[\mathbf{R} \times \mathbf{S}^n \times [\mathrm{int}\,\mathbf{S}^n_+] \times \mathbf{S}^n\right] \times \mathbf{S}^n : \left[\begin{array}{c|c} z + w & y \\ \hline y & v \end{array}\right] \succeq 0\}$$

which clearly is convex. Applying Proposition IV.19.8, we conclude that $\mathrm{Opt}(p)$ is convex on P. Recalling that by item 2 of Exercise, $(t_* = -\mathrm{Tr}(B^{1/2}), y_* = -B^{1/2})$ is a cone-constrained KKT point of $(\mathrm{P}[\overline{p}])$,

the corresponding Lagrange multipliers being $\overline{\lambda}=1$, $\widehat{\lambda}=\frac{1}{2}B^{-1/2}$, and taking into account that in the case in question in the notation from Proposition IV.19.8 we have

$$F_p=0,\ G_p[\delta v,\delta w]=-y_*v^{-1}\delta v v^{-1}y_*\Big|_{v=I_n}-\delta w$$

(see Example C.8 in section C.1.6), the pair $[-\frac{1}{2}B^{1/2},-\frac{1}{2}B^{-1/2}]$ is a subgradient of $\mathrm{Opt}(\cdot)$ at $p=\overline{p}=[I_n,B]$:

$$\begin{array}{rcl}\mathrm{Opt}(p=(v,w)) & \geq & \underbrace{-\mathrm{Tr}(B^{1/2})}_{\mathrm{Opt}(I_n,B)}+\mathrm{Tr}(\widehat{\lambda}[G_p[v-I_n,w-B]])\\ & = & -\mathrm{Tr}(B^{1/2})-\frac{1}{2}\mathrm{Tr}([v-I_n]B^{1/2})-\frac{1}{2}\mathrm{Tr}(B^{-1/2}[w-B]).\end{array}$$

Exercise IV.23. [follow-up to Exercise IV.4] Given positive integers m,n, consider two parametric families of convex sets:

- $S_1[P]=\{(X,Y)\in\mathcal{R}_1:=\mathbf{S}^m\times\mathbf{S}^n:\left[\begin{array}{c|c}X&P\\\hline P^\top&Y\end{array}\right]\succeq 0\}$, where the "parameter" P runs through the space $\mathbf{R}^{m\times n}$ of $m\times n$ matrices, let it be temporarily denoted $\mathcal{P}_1$;
- $S_2[P]=\{(X,Y)\in\mathcal{R}_2:=\mathbf{S}^m\times\mathbf{R}^{m\times n}:\left[\begin{array}{c|c}X&Y\\\hline Y^\top&P\end{array}\right]\succeq 0\}$, where the "parameter" P runs through the positive semidefinite cone $\mathbf{S}^n_+$, let it be temporarily denoted $\mathcal{P}_2$.

Prove that for $\chi=1,2$ the set-valued mappings $P\to S_\chi[P]$ are super-additive on their domains:

$$P,Q\in\mathcal{P}_\chi\implies P+Q\in\mathcal{P}_\chi\ \&\ \underbrace{S_\chi[P]+S_\chi[Q]\subset S_\chi[P+Q]}_{(*)}.$$

and that the concluding inclusion
— not necessarily is equality for $\chi=1$, and
— is equality for $\chi=2$.

Solution: The super-additivity is evident due to the following evident fact:

> *Let $\mathbf{K}\subset\mathbf{R}^M$, $\mathcal{P}\subset\mathbf{R}^N$ be cones, and $\mathcal{A}(U,P)$ be a linear (linear, not affine!) mapping from $\mathbf{R}^K_U\times\mathbf{R}^N_P$ into $\mathbf{R}^M$ such that the set*
>
> $$S[P]=\{U\in\mathbf{R}^K_U:\mathcal{A}(U,P)\in\mathbf{K}\}$$
>
> *is nonempty when $P\in\mathcal{P}$. Then the set-valued mapping $P\mapsto S[P]$ is super-additive on $\mathcal{P}$.*
> Indeed, when $P_i\in\mathcal{P}$ and $U_i\in S[P_i]$, $i=1,2$, we have $\mathbf{K}\ni\mathcal{A}(U_1,P_1)+\mathcal{A}(U_2,P_2)=\mathcal{A}(U_1+U_2,P_1+P_2)$, implying that $U_1+U_2\in S[P_1+P_2]$.

The fact that inclusion $(*)$ can be strict when $\chi=1$ is nearly evident: take a nonzero $P\in\mathbf{R}^{m\times n}$ and note that all pairs $(X,Y)\in S_1[P]$, same as all pairs $(X,Y)\in S_1[-P]$, have nonzero positive semidefinite components X,Y, while $S[0_{m\times n}]$ contains the pair $(X=0_{m\times m},Y=0_{n\times n})$ which clearly cannot be represented as the sum of two pairs of matrices with nonzero positive semidefinite components in every pair. Thus, when $Q=-P\neq 0$ and $\chi=1$, inclusion $(*)$ is strict.
The only nontrivial part of the exercise is to prove that when $\chi=2$, the inclusion $(*)$ is equality whenever P,Q are positive semidefinite. In other words, we want to demonstrate that when

$$P\in\mathbf{S}^n_+\ \&\ Q\in\mathbf{S}^n_+\ \&\ R\in\mathbf{S}^m,S\in\mathbf{R}^{m\times n}\ \&\ D:=\left[\begin{array}{c|c}R&S\\\hline S^\top&P+Q\end{array}\right]\succeq 0 \tag{1}$$

the matrix D can be decomposed into the sum of two positive semidefinite matrices, one of the form $\left[\begin{array}{c|c}U&V\\\hline V^\top&P\end{array}\right]$, and the other – of the form $\left[\begin{array}{c|c}W&Z\\\hline Z^\top&Q\end{array}\right]$. This is the same as to verify that setting

$$\begin{array}{c}A_+(U,V)=\left[\begin{array}{c|c}U&V\\\hline V^\top&\end{array}\right],\ A_-(U,V)=\left[\begin{array}{c|c}-U&-V\\\hline -V^\top&\end{array}\right],\\ b_+=-\left[\begin{array}{c|c}&\\\hline &P\end{array}\right],\ b_-=-\left[\begin{array}{c|c}R&S\\\hline S^\top&Q\end{array}\right]\end{array}$$

the conic constraint

$$\underbrace{(A_+(U,V),A_-(U,V))}_{=:A(U,V)}-\underbrace{(b_+,b_-)}_{=:b}\in\underbrace{\mathbf{S}^{m+n}_+\times\mathbf{S}^{m+n}_+}_{=:\mathbf{K}} \tag{2}$$

in variables U, V is feasible.

Assume that the constraint is infeasible, and let us lead this assumption to contradiction. The image space of the linear mapping $(U,V)\mapsto A(U,V)$ is composed of pairs

$$\left(\left[\begin{array}{c|c} U & V\\ \hline V^\top & \end{array}\right],\left[\begin{array}{c|c} -U & -V\\ \hline -V^\top & \end{array}\right]\right)$$

of symmetric $(m+n)\times(m+n)$ matrices with $U\in\mathbf{S}^m$ and $V\in\mathbf{R}^{m\times n}$; such a pair can belong to $\mathbf{K}$ (that is, has both components positive semidefinite) if and only if $V=0$ and both U and $-U$ are positive semidefinite, that is, if and only if $U=0$ and $V=0$. With this in mind, the results of Exercise IV.4 say that infeasibility of (2) implies existence of $\lambda=(\lambda_+,\lambda_-)\in\mathbf{K}_*$, that is, $\lambda_\pm\in\mathbf{S}^{m+n}_+$, such that

$$\mathrm{Tr}(\lambda_+A_+(U,V))+\mathrm{Tr}(\lambda_-A_-(U,V))=0\,\forall(U\in\mathbf{S}^m,V\in\mathbf{R}^{m\times n})\ \&\ \mathrm{Tr}(\lambda_+b_+)+\mathrm{Tr}(\lambda_-b_-)>0. \tag{3}$$

Representing

$$\lambda_\pm=\left[\begin{array}{c|c} E_\pm & F_\pm\\ \hline F_\pm^\top & G_\pm\end{array}\right]\quad[E_\pm\in\mathbf{S}^m,G_\pm\in\mathbf{S}^n)$$

the first relation in (3) boils down to

$$\mathrm{Tr}(U[E_+-E_-])+2\mathrm{Tr}(V[F_+-F_-]^\top)=0\,\forall(U\in\mathbf{S}^m,V\in\mathbf{R}^{m\times n}),$$

that is, $E_+=E_-=:E$, $F_+=F_-=:F$. Now the second relation in (3) reads

$$\mathrm{Tr}(G_+P)+\mathrm{Tr}(ER)+2\mathrm{Tr}(FS^\top)+\mathrm{Tr}(G_-Q)<0. \tag{4}$$

and, besides this $\left[\begin{array}{c|c} E & F\\ \hline F^\top & G_\pm\end{array}\right]\succeq 0$. When replacing E in (4) with $E'=E+\epsilon I_m$ with small enough positive ϵ, the strict inequality remains valid:

$$\mathrm{Tr}(G_+P)+\mathrm{Tr}(E'R)+2\mathrm{Tr}(FS^\top)+\mathrm{Tr}(G_-Q)<0. \tag{5}$$

On the other hand, we have $\left[\begin{array}{c|c} E' & F\\ \hline F^\top & G_\pm\end{array}\right]\succeq 0$ and $E'\succ 0$, whence, by Schur Complement Lemma, $G_\pm\succeq G:=F^\top[E']^{-1}F$. When replacing $G_\pm$ with $G\preceq G_\pm$ in the left hand side of (5), we can only decrease the value of the left hand side due to $P\succeq 0$, $Q\succeq 0$, so that

$$\mathrm{Tr}(GP)+\mathrm{Tr}(E'R)+2\mathrm{Tr}(FS^\top)+\mathrm{Tr}(GQ)<0.$$

The left hand side here is nothing but

$$\mathrm{Tr}\left(\left[\begin{array}{c|c} R & S\\ \hline S^\top & P+Q\end{array}\right]\left[\begin{array}{c|c} E' & F\\ \hline F^\top & G\end{array}\right]\right)$$

i.e., it is the Frobenius inner product of two positive semidefinite (by (1) and due to the origin of G) matrices, and this product cannot be negative, which is the desired contradiction. ■

Exercise IV.24. In the simplest Steiner problem, one is given m distinct points $a_1,\dots,a_m$ in $\mathbf{R}^n$ and is looking for a point x_* such that the sum of Euclidean distances between the points and x_* is as small as possible (think, e.g., about m oil wells on 2D plane and the problem of locating collector to be linked to the wells by pipes in a way minimizing the total length of the pipes).

1. Pose the problem as conic problem, the cone being direct product of m Lorentz cones.
2. Build the dual problem. Are the primal and the dual problems solvable? Are the primal and the dual optimal values equal to each other?
3. Write down optimality conditions and see what they say
   Hint: You are advised to consider separately the cases where optimal solution differs from all of the points $a_1,\dots,a_m$, and the case when it is one of the points.

4. Solve the problem in the case when $n = 2$, $m = 3$ and a_1, a_2, a_3 are vertices of triangle on 2D plane.

Solution: 1: The "maiden" form of the problem is

$$\min_{x\in\mathbf{R}^n} \sum_{i=1}^m \|x - a_i\|_2, \tag{$P_{\rm ini}$}$$

the conic reformulation is

$$\min_{t,x} \left\{ \sum_i t_i : \|x - a_i\|_2 \le t_i, i \le m \right\};$$

the constraints in this reformulation say that the vectors $[x - a_i; t_i]$ belong to the Lorentz cone $\mathbf{L}^{n+1}$, and the objective is linear. To obey our standards on writing down conic problems, we should rewrite the above problem as

$$\begin{array}{c}\mathrm{Opt}(P) = \min_{x,t} \left\{ \sum_i t_i : A\,[x;t] - b := [[x - a_1; t_1]; [x - a_2; t_2]; \ldots; [x - a_m; t_m]] \ge_{\mathbf{K}} 0 \right\} \\ \mathbf{K} = \underbrace{\mathbf{L}^{n+1} \times \ldots \times \mathbf{L}^{n+1}}_{m \text{ times}}\end{array} \tag{P}$$

2: To build the dual problem, we equip the conic constraint with Lagrange multiplier restricted to belong to the dual to $\mathbf{K}$ cone $\mathbf{K}_*$. This dual cone is $\mathbf{K}$ itself due to self-duality of the Lorentz cone, so that the Lagrange multiplier is $[[y_1; s_1]; \ldots; [y_m; s_m]]$ with $[y_i; s_i] \in \mathbf{L}^{n+1}$. We then take the sidewise inner product of the conic constraint of the primal problem with the Lagrange multiplier, thus arriving at the scalar linear inequality

$$\sum_i [t_i s_i + y_i^\top x] \ge \sum_i y_i^\top a_i; \tag{$*$}$$

whenever $[y_i; s_i] \in \mathbf{L}^{n+1}$, $i \le m$, this inequality is consequence of the constraints of the primal problem. To build the dual problem, we impose on the Lagrange multipliers, aside of the restrictions $[y_i; s_i] \in \mathbf{L}^{n+1}$, the restriction that the left hand side in $(*)$ is equal to the objective of (P) identically in $x, t_1, \ldots, t_m$, which in our case reads

$$s_i = 1, i \le m, \sum_i y_i = 0.$$

The dual problem is to maximize the right hand side of $(*)$ in $[y_i; s_i]$ under the resulting constraints, so that the dual problem reads

$$\mathrm{Opt}(D) = \max_{y_i, s_i} \left\{ \sum_i a_i^\top y_i : \|y_i\|_2 \le s_i = 1, ., i \le m, \sum_i y_i = 0 \right\} \tag{D}$$

The primal and the dual problems clearly satisfy the Slater and the Relaxed Slater conditions, respectively, so that by Conic Duality Theorem problems are solvable with equal optimal values.
3: Optimality conditions from Theorem IV.19.9 in their "complementary slackness" form state that the (under the circumstances, necessary and sufficient) condition for feasible solutions $(x, \{t_i\})$ to (P) and $\{y_i, s_i\}$ to (D) to be optimal for the respective problems is

$$\sum_i [x - a_i; t_i]^\top [y_i; s_i] = 0,$$

or, which is the same due to $s_i = 1$ (by dual feasibility) and $[x - a_i; t_i] \in \mathbf{L}^{n+1}$, $[y_i; s_i] \in \mathbf{L}^{n+1}$,

$$y_i^\top [a_i - x] = t_i,\ i \le m$$

Since $\|y_i\|_2 \le 1$ and $\|a_i - x\|_2 \le t_i$, the above equality implies that $y_i^\top [a_i - x] \ge \|y_i\|_2 \|a_i - x\|_2$. This, taken together with what we know about equality case of Cauchy inequality (Theorem B.1) means that for every i,
— either $x \ne a_i$, and then $y_i = \frac{a_i - x}{\|a_i - x\|_2}$ and $t_i = \|a_i - x\|_2$
— or $x = a_i$, $t_i = 0$, and $\|y_i\|_2 \le 1$.

We conclude that there are two possible cases:
(A): the x-component of optimal solution to (P) differs from every one of a_i. In this case we should have $y_i = \frac{a_i - x}{\|a_i - x\|_2}$ for all i;
(B): the x-component of optimal solution to (P) is some a_{i_*}. In this case $y_i = \frac{a_i - x}{\|a_i - x\|_2}$ for all $i \neq i_*$ and y_{i_*} can be arbitrary vector of $\|\cdot\|_2$-norm ≤ 1.
Taking into account that we should have $\sum_i y_i = 0$, we arrive at the following conclusions:

> (A): If x is different from all a_i and satisfies the relation $\sum_i \frac{a_i - x}{\|a_i - x\|_2} = 0$, then x and $t_i = \|x - a_i\|_2$, $i \leq m$, form an optimal solution to (P).
> (B): If $x = a_{i_*}$ for i_* such that $\|\sum_{i \neq i_*} \frac{a_i - a_{i_*}}{\|a_i - a_{i_*}\|_2}\|_2 \leq 1$, then x and $t_i = \|x - a_i\|_2$, $i \leq m$, form an optimal solution to (P).
> Moreover, every optimal solution to (P) is either given by (A), or by (B), and optimal solutions do exist.

4: When $m = 3$ and a_1, a_2, a_3 are the vertices of a triangle in 2D plane ($n = 2$), (A) says that a point distinct from a_1, a_2, a_3 solves $(P_{\rm ini})$ if and only if all three sides of $\triangle a_1 a_2 a_3$ are seen from the point x under angles 120^o — the unit vectors "looking" from x at the vertices of the triangle should sum up to 0. Elementary geometry says that such a point does exist when all three angles of $\triangle a_1 a_2 a_3$ are $< 120^o$. If the latter is not the case, optimal x is given by (B) and is just the vertex of $\triangle a_1 a_2 a_3$ with angle at the vertex $\geq 120^o$.

Exercise IV.25. Consider a primal-dual pair of conic problems

$$\mathrm{Opt}(P) = \min_x \left\{ c^T x : \ Ax \geq_{\mathbf{K}} b \right\} \qquad (P)$$
$$\mathrm{Opt}(D) = \max_y \left\{ b^T y : \ y \geq_{\mathbf{K}_*} 0, \ A^T y = c \right\} \qquad (D)$$

($\mathbf{K} \subset \mathbf{R}^n$ is a regular cone) and assume that both problems are feasible.

1. Find the recessive cones $\mathrm{Rec}(P)$ and $\mathrm{Rec}(D)$ of the primal and the dual feasible sets.
2. Prove that the feasible set of at least one of the problems is unbounded.

Solution: The feasible sets of (P) and (D) are nonempty, convex, and clearly closed. Let $\bar{x}$ be primal feasible and $\bar{y}$ be dual feasible. Then, we have

$$\begin{aligned}
\mathrm{Rec}(P) &= \{h : \ A[\bar{x} + th] - b \in \mathbf{K}, \forall t \geq 0\} = \left\{h : \ Ah - t^{-1}[A\bar{x} - b] \in \mathbf{K}, \forall t > 0\right\} \\
&= \{h : \ Ah \in \mathbf{K}\} \\
\mathrm{Rec}(D) &= \left\{g : \ \bar{y} + tg \in \mathbf{K}_*, tA^T g = 0, t \geq 0\right\} = \left\{g \in \mathbf{K}_* : \ A^T g = 0\right\}.
\end{aligned}$$

Now assume that the feasible set of (D) is bounded, and let us prove that the feasible set of (P) is unbounded. As $\mathbf{K}$ is a regular cone, we can select $f \in \mathrm{int}\,\mathbf{K}$. Consider two convex sets $S = \left\{y : \ A^T y = 0\right\}$ and $T = \left\{y \in \mathbf{K}_* : \ f^T y = 1\right\}$; note that $T \neq \varnothing$ due to $f \in \mathrm{int}\,\mathbf{K}$. We are in the situation where the dual feasible set is nonempty, closed, and bounded, implying by Fact II.6.18 that $\mathrm{Rec}(D) = \{0\}$, that is, $S \cap \mathbf{K}_* = \{0\}$, whence the closed nonempty convex sets S and T do not intersect. Recall also that as $\mathbf{K}$ is a regular cone so is its dual $\mathbf{K}_*$, and thus by Fact II.6.38 T is compact. Then, the convex sets S, T are at positive distance from each other and one of them is compact, hence they can be strongly separated: there exists a vector a such that

$$\sup_{y \in S} a^T y < \min_{y \in T} a^T y.$$

As S is a linear space and T is a base of $\mathbf{K}_*$ (Fact II.6.38), this relationship is equivalent to

$$a \in S^\perp = [\mathrm{Ker} A^T]^\perp = \mathrm{Im}\, A, \quad \text{and} \quad a \in \mathrm{int}(\mathbf{K}_*)_* = \mathrm{int}\,\mathbf{K}.$$

Here the inclusion $a \in \mathrm{int}(\mathbf{K}_*)_*$ is given by Fact II.6.38.ii as applied to the regular cone $\mathbf{K}_*$ in the role of M combined with the fact that $a^T y > 0$ for all $y \in T$ and therefore for all $y \in \mathbf{K}_* \setminus \{0\}$. Thus, we conclude that $a = Ah$ for some h such that $Ah \in \mathrm{int}\,\mathbf{K}$ and thus $h \neq 0$, which clearly shows that

$\mathrm{Rec}(P) = \{h : Ah \in \mathbf{K}\} \neq \{0\}$ and thus the primal feasible set is unbounded. Note that we have shown $\{h : Ah \in \mathrm{int}\,\mathbf{K}\} \neq \varnothing$, which is indeed something even stronger than what was desired. ■

Exercise IV.26. [semidefinite duality] A *semidefinite program* is a conic program involving the positive semidefinite cone. As a matter of fact, *Semidefinite programming* – the family of semidefinite programs – possesses extremely powerful "expressive abilities." It is prudent to say that *for all practical purposes*, whatever it means, Semidefinite programming is "the same" as the entire Convex programming. In this exercise we would like to acquaint the reader with the specific form taken by Conic duality when the cone involved is the positive semidefinite cone.

Formally, a semidefinite program is of the form

$$\mathrm{Opt}(P) = \min_{x\in\mathbf{R}^n}\left\{c^\top x \;:\; \begin{array}{l} Ax-b := \sum_j a_j x_j - b \geq 0 \\ \mathcal{A}x - B := x_1A_1 + \ldots + x_nA_n - B \succeq 0 \end{array}\right\}, \qquad (P)$$

where a_j, b are vectors from some $\mathbf{R}^p$, and A_j, B are matrices from some $\mathbf{S}^q$. "Real life" form of a semidefinite program usually is a bit different, namely,

$$\mathrm{Opt}(\mathcal{P}) = \min_{x\in\mathbf{R}^n}\left\{c^\top x : \begin{array}{l} Ax-b := \sum_j x_j a_j - b \geq 0 \\ \mathcal{A}_i x - B^i := x_1A_1^i + \ldots + x_nA_n^i - B^i \succeq 0,\ \forall i \leq m \end{array}\right\}, \qquad (\mathcal{P})$$

where $A_j^i, B^i \in \mathbf{S}^{q_i}$. In the formulation $(\mathcal{P})$ as opposed to the formulation (P) we have a bunch of positive semidefinite cone constraints, i.e., $\mathcal{A}_i x - B^i \succeq 0$, $i \leq m$, instead of a single constraint $\mathcal{A}x - B \succeq 0$. We can always rewrite $(\mathcal{P})$ in the form of (P) by assembling A_j^i, B^i into block-diagonal matrices $A_j = \mathrm{Diag}\{A_j^1, \ldots, A_j^m\}$, $B = \mathrm{Diag}\{B^1, \ldots, B^m\}$. Taking into account that a block-diagonal symmetric matrix is positive semidefinite if and only if all the diagonal blocks are positive semidefinite, we deduce that $(\mathcal{P})$ is equivalent to the problem

$$\min_{x\in\mathbf{R}^n}\left\{c^\top x : \begin{array}{l} Ax-b := \sum_j x_j a_j - b \geq 0 \\ \mathcal{A}x - B := \sum_j x_j A_j - B \succeq 0 \end{array}\right\}$$

of the form (P). When proving theorems, it is usually better to work with program in the form of (P) – it saves notation; in contrast, when working with "real life" semidefinite programs, it is usually better to operate with problems in more detailed form $(\mathcal{P})$.

Your task is as follows:

1. Verify that the conic dual of $(\mathcal{P})$ is the semidefinite program

$$\max_{\lambda,\{\Lambda_i, i\leq m\}}\left\{b^\top\lambda + \sum_{i=1}^m \mathrm{Tr}(\Lambda_i B^i) : \begin{array}{l} \lambda \in \mathbf{R}^p_+, \Lambda_i \in \mathbf{S}^{q_i}_+, i \leq m \\ A^\top\lambda + \sum_{i=1}^m \mathcal{A}_i^*\Lambda_i = c, \end{array}\right\}, \qquad (\mathcal{D})$$

   where for the linear mapping $x \mapsto \sum_j x_j A_j : \mathbf{R}^n \to \mathbf{S}^q$ its *conjugate* linear mapping $X \mapsto \mathcal{A}^*X : \mathbf{S}^q \to \mathbf{R}^n$ is given by the identity

$$\mathrm{Tr}(X[\mathcal{A}x]) \equiv [\mathcal{A}^*X]^\top x \quad \forall (x \in \mathbf{R}^n, X \in \mathbf{S}^q),$$

   or, which is the same,

$$\mathcal{A}^*X = [\mathrm{Tr}(A_1X); \ldots; \mathrm{Tr}(A_nX)].$$

Solution: $(\mathcal{P})$ is the cone-constrained problem

$$\begin{array}{l} \min\limits_{x\in\mathbf{R}^n}\left\{f(x) := c^\top x :\ \overline{g}(x) := b - Ax \leq 0,\ \widehat{g}(x)\right. \\ \left. := \mathrm{Diag}\left\{B^1 - \sum_j x_j A_j^1, \ldots, B^m - \sum_j x_j A_j^m\right\} \in -\mathbf{K}\right\} \end{array} \qquad (*)$$

where $\mathbf{K}$ is the cone composed of the positive semidefinite block-diagonal symmetric matrices with m diagonal blocks of sizes $q_1, \ldots, q_m$. Then, the cone $\mathbf{K}$ lives in the space $\mathbf{S}^{\{q_1,\ldots,q_m\}}$ of block-diagonal symmetric matrices with m diagonal blocks of sizes $q_1, \ldots, q_m$. Equipping $\mathbf{S}^{\{q_1,\ldots,q_m\}}$ with Frobenius inner product and taking into account that positive semidefinite cone is self-dual, we immediately conclude that $\mathbf{K}$ is self-dual as well. As a result,

- the Lagrange multipliers $\Lambda \in \mathbf{K}_*$ are exactly block-diagonal matrices $\Lambda = \mathrm{Diag}\{\Lambda_1, \ldots \Lambda_m\}$ with diagonal blocks $\Lambda_i \in \mathbf{S}^{q_i}_+$, for all $i \le m$;
- the cone-constrained Lagrange function of $(*)$ is the function

$$L(x; \lambda, \Lambda) = f(x) + \lambda^\top \overline{g}(x) + \mathrm{Tr}(\widehat{g}(x)\Lambda), \qquad (!)$$

where the last term in the right hand side is precisely what is prescribed by our general description of cone-constrained Lagrange function, i.e., it is the inner product of the Lagrange multiplier Λ for the cone constraint $\widehat{g}(x) \le_{\mathbf{K}} 0$ and the left hand side of this constraint[12]. In other words,

$$L(x; \lambda, \Lambda) = c^\top x + \lambda^\top [b - Ax] + \textstyle\sum_{i=1}^m \mathrm{Tr}\left(\Lambda_i [B^i - \sum_j A^i_j x_j]\right) : \mathbf{R}^n_x \times \left[\mathbf{R}^p_+ \times \mathbf{S}^{q_1}_+ \times \ldots \times \mathbf{S}^{q_m}_+\right] \to \mathbf{R}.$$

Consequently, the cone-constrained Lagrange dual of $(*)$ is the problem

$$\max_{\lambda \in \mathbf{R}^p_+, \Lambda = \{\Lambda_i \in \mathbf{S}^{q_i}_+\}} \left\{ \begin{array}{l} \underline{L}(\lambda, \Lambda) \\ \qquad := \inf\limits_{x \in \mathbf{R}^n} \left[\lambda^\top b + \sum_i \mathrm{Tr}(\Lambda_i B^i) + \sum_j x_j [c_j - \lambda^\top a_j - \sum_i \mathrm{Tr}(\Lambda_i A^i_j)]\right]. \end{array} \right\}$$

Note also that

$$\underline{L}(\lambda, \Lambda) = \left\{ \begin{array}{ll} \lambda^\top b + \sum_i \mathrm{Tr}(\Lambda_i B_i), & \text{if } A^\top \lambda + \sum_i \left[\mathrm{Tr}(A^i_1 \Lambda_i); \ldots; \mathrm{Tr}(A^i_n \Lambda_i)\right] = c \\ -\infty, & \text{otherwise} \end{array} \right. .$$

Therefore, the conic dual of $(\mathcal{P})$ is given by

$$\max_{\lambda, \{\Lambda_i, i \le m\}} \left\{ \lambda^\top b + \sum_i \mathrm{Tr}(\Lambda_i B_i) : \begin{array}{l} \lambda \in \mathbf{R}^p_+,\ \Lambda_i \in \mathbf{S}^{q_i}_+,\ i \le m \\ A^\top \lambda + \sum_i \left[\mathrm{Tr}(A^i_1 \Lambda_i); \ldots; \mathrm{Tr}(A^i_n \Lambda_i)\right] = c \end{array} \right\} \qquad (\mathcal{D})$$

In words, the recipe for building the dual to the semidefinite program $(\mathcal{P})$ is as follows:

1. We equip the constraints of $(\mathcal{P})$ with Lagrange multipliers, specifically, the linear constraints $Ax - b \ge 0$ with the multiplier $\lambda \in \mathbf{R}^p$ such that $\lambda \ge 0$, and the semidefinite constraints $\mathcal{A}_i x - B_i := \sum_j x_j A^i_j - B^i \succeq 0$ with the multipliers $\Lambda_i \in \mathbf{S}^{q_i}$ such that $\Lambda_i \succeq 0$.
2. We take the inner products of the left hand sides of the constraints in $(\mathcal{P})$ and the associated Lagrange multipliers (the standard inner product for the linear constraint $Ax - b \ge 0$, and the Frobenius inner products for the semidefinite constraints $\mathcal{A}_i x - B^i \succeq 0$) and sum up the results, arriving at the aggregated constraint

$$\left[A^\top \lambda + \sum_i \mathcal{A}^*_i \Lambda_i\right]^\top x \ge b^\top \lambda + \sum_i \mathrm{Tr}(B^i \Lambda_i),$$
$$\text{where } \mathcal{A}^*_i X = [\mathrm{Tr}(A^i_1 X); \ldots; \mathrm{Tr}(A^i_n X)].$$

By its origin, this constraint is a consequence of the system of constraints in $(\mathcal{P})$.

[12] in our general description of cone-constrained Lagrange function, the cone in the cone constraint lived in some $\mathbf{R}^N$, and the product of the Lagrange multiplier and the body of the constraint was the standard inner product in $\mathbf{R}^N$. Our present situation can be reduced to the standard one by identifying $\mathbf{S} = \mathbf{S}^{\{q_1, \ldots, q_m\}}$ equipped with the Frobenius inner product with appropriate $\mathbf{R}^N$ equipped with the standard inner product, identification being given by selecting orthonormal, w.r.t. the Frobenius inner product, basis in $\mathbf{S}$ and identifying $X \in \mathbf{S}$ with the vector of coordinates of X in this basis. There, however, is no necessity to carry out this identification explicitly, since all we are interested in is what will be the standard inner product of vectors of coordinates of Λ and of $\widehat{g}(x)$ in this orthonormal basis, and we know the answer in advance – this will be the Frobenius inner product of Λ and $\widehat{g}(x)$, the entity we see in (!).

3. We impose on the Lagrange multipliers, aside of the restrictions mentioned in item 1, the restriction that the left hand side in the aggregated constraint is equal to $c^\top x$ identically in $x \in \mathbf{R}^n$, so that the right hand side in this constraint is a lower bound on $\mathrm{Opt}(\mathcal{P})$, The dual program $(\mathcal{D})$ is nothing but the problem of maximizing this lower bound over Lagrange multipliers satisfying the restrictions just listed.

Exercise IV.27. [example of semidefinite relaxation] Let $T_k \succeq 0, k \le K$, be positive semidefinite $m \times m$ matrices such that $\sum_k T_k \succ 0$, $\mathcal{T} \subset \mathbf{R}^K_+$ be a convex compact set intersecting the interior of $\mathbf{R}^K_+$, and A be a symmetric $m \times m$ matrix. Let also $\phi_{\mathcal{T}}(z) = \max_{t \in \mathcal{T}} z^\top t$ be the support function of $\mathcal{T}$. Prove that

$$\begin{array}{rcll} \mathrm{Opt} & := & \min_z \left\{\phi_{\mathcal{T}}(z) : z \ge 0, A \preceq \sum_k z_k T_k\right\} & (a) \\ & = & \max_{\Lambda,t} \left\{\mathrm{Tr}(A\Lambda) : \Lambda \succeq 0, t \in \mathcal{T}, \mathrm{Tr}(T_k\Lambda) \le t_k, k \le K\right\} & (b) \end{array}$$

and that both minimization and maximization problems above are solvable.

Solution: Since $\mathcal{T}$ is bounded, $\phi_{\mathcal{T}}$ is real-valued and continuous, and since $\mathcal{T} \subset \mathbf{R}^K_+$ contains a positive vector, the sets $\{z \ge 0 : \phi_{\mathcal{T}}(z) \le a\}$ are closed and bounded for every $a \in \mathbf{R}$. The problem specifying Opt is cone constrained problem which is strictly feasible (due to $\sum_k T_k \succ 0$), and by the above, denoting by $\mathcal{Z}$ the feasible set of the problem, the feasible sublevel sets $\{z \in \mathcal{Z} : \phi_{\mathcal{T}}(z) \le a\}$ of $\phi_{\mathcal{T}}$ are closed and bounded for every a; since the objective is continuous, it follows that the problem is solvable (Theorem B.32). The minimization problem specifying Opt is cone constrained strictly feasible and below bounded problem. Thus, by cone constrained version of Convex Programming Duality Theorem (Theorem IV.18.1), the cone constrained Lagrange dual of problem (a) is solvable with optimal value Opt. The cone constrained Lagrange function of (a) is

$$L(z;\lambda,\Lambda) = \phi_{\mathcal{T}}(z) - \lambda^\top z + \mathrm{Tr}(\Lambda[A - \sum_k z_k T_k]) : \mathbf{R}^K_z \times [\mathbf{R}^K_+ \times \mathbf{S}^m_+] \to \mathbf{R},$$

so that the objective in the dual problem is

$$\underline{L}(\lambda,\Lambda) = \mathrm{Tr}(A\Lambda) + \inf_{z \in \mathbf{R}^K} \Big[\phi_{\mathcal{T}}(z) - z^\top \underbrace{[\mathrm{Tr}(\Lambda T_1) + \lambda_1; ...; \mathrm{Tr}(\Lambda T_K) + \lambda_K]}_{=:\ell(\lambda,\Lambda)}\Big],$$

that is, $\underline{L}(\lambda,\Lambda) - \mathrm{Tr}(\Lambda A)$ is the minus Legendre transform $\phi^*_{\mathcal{T}}$ of $\phi_{\mathcal{T}}(\cdot)$ as evaluated at $\ell(\lambda,\Lambda)$. Since $\mathcal{T}$ is convex, nonempty, and closed, $\phi^*_{\mathcal{T}}$ is just the characteristic function of $\mathcal{T}$ (Exercise III.10), that is,

$$\underline{L}(\lambda,\Lambda) = \left\{ \begin{array}{ll} \mathrm{Tr}(A\Lambda) & , \ell(\lambda,\Lambda) \in \mathcal{T} \\ -\infty & , \text{otherwise} \end{array} \right\}$$

so that the cone constrained Lagrange dual of (a), which is the problem of maximizing $\underline{L}$ over the set $\{\lambda \ge 0, \Lambda \succeq 0\}$ is equivalent to (b). ■

Exercise IV.28. What follows is the concluding exercise in the "Truss Topology Design" series. We have already used TTD problem to present instructive "real life" illustrations of the power of several results of Convex Analysis, specifically, Caratheodory Theorem (Exercise I.18), epigraph description of convexity and Helly Theorem (Exercise III.9) and $\mathcal{S}$-Lemma (Exercise IV.11), not speaking about the Schur Complement Lemma which was instrumental in all these exercises. Now it is time to illustrate the power of conic duality.

In the sequel, we assume that the reader is reasonably well acquainted with Truss Topology Design story as told in Exercise I.16 and use without additional comments the notions, notation, and the results presented in this Exercise, including the default assumption $\mathfrak{R}$ which remains in force below. In addition, we assume from now on that the load of interest f is nonzero – this is the only nontrivial case in TTD.

Recall that the TTD problem as posed in Exercise I.16.2 reads

$$\mathrm{Opt} = \min_{\tau,r} \left\{ \tau : \left[\begin{array}{c|c} B\,\mathrm{Diag}\{t\}B^\top & f \\ \hline f^\top & 2\tau \end{array}\right] \succeq 0, t \ge 0, \sum_i t_i = W \right\} \qquad (P)$$

In our present language, this is a semidefinite program, and we know from Exercise I.16 that this problem is solvable.

Your first task is easy:

1. Build the semidefinite dual of (P) and prove that the dual problem is solvable with the same optimal value Opt as the primal problem (P).

Since passing from a semidefinite problem to its dual is a purely mechanical process, on one hand, and the subsequent tasks will be formulated in terms of the dual problem, here is the dual as given by Conic Duality:

$$\max_{V,g,\theta,\lambda,\mu}\left\{-2f^\top g - W\mu : 2\theta = 1, \mathfrak{b}_i^\top V\mathfrak{b}_i + \lambda_i - \mu = 0\,\forall i, \lambda \geq 0, \left[\begin{array}{c|c} V & g \\ \hline g^\top & \theta \end{array}\right] \succeq 0\right\}$$

Eliminating variable θ (which is fixed by the corresponding constraint), we rewrite the dual as

$$\max_{V,g,\lambda,\mu}\left\{-2f^\top g - W\mu : \mathfrak{b}_i^\top V\mathfrak{b}_i + \lambda_i - \mu = 0\,\forall i, \lambda \geq 0, \left[\begin{array}{c|c} V & g \\ \hline g^\top & \frac{1}{2} \end{array}\right] \succeq 0\right\} \qquad (D)$$

What is left to you, is to verify the derivation and to prove that (D) is solvable with the same optimal value Opt as (P).

Solution: Assumption $\mathfrak{R}$ states that every $t > 0$ satisfying the linear equality $\sum_i t_i = W$ results in positive definite matrix $B\,\mathrm{Diag}\{t\}B^\top$, implying by the Schur Complement Lemma that augmenting t with large enough τ, we get a feasible solution to (P) which strictly satisfies all $\geq$- and $\succeq$-constraints of (P). Thus, (P) is essentially strictly feasible (and of course bounded – the objective is nonnegative on the feasible set, not speaking about already known to us solvability of (P)). Applying Conic Duality Theorem, we conclude that (D) is solvable with the same optimal value Opt as the primal problem (P). ■

Your next task still is easy:

2. Verify that eliminating, by partial optimization, variables V and λ, problem (D) reduces to the problem

$$\max_{g,\mu}\left\{-2f^\top g - W\mu : \left[\begin{array}{c|c} \mu & \mathfrak{b}_i^\top g \\ \hline \mathfrak{b}_i^\top g & \frac{1}{2} \end{array}\right] \succeq 0\,\forall i\right\} \qquad (\overline{D})$$

   and the latter problem is solvable with the same optimal value Opt as (P) and (D).

Pay attention to the first surprising fact: semidefinite constraints in $(\overline{D})$ involve the cone $\mathbf{S}^2_+$ of 2×2 positive semidefinite matrices, and this cone, as we know, is, up to one-to-one linear transformation, just the Lorentz cone $\mathbf{L}^3$. Thus, $(\overline{D})$ is a conic quadratic problem.

Solution: Eliminating variables λ_i is immediate – all we need is to replace the linear equality constraints $\mathfrak{b}_i^\top V\mathfrak{b}_i + \lambda_i - \mu = 0$ with inequality constraints

$$\mathfrak{b}_i^\top V\mathfrak{b}_i \leq \mu,\ i \leq N,$$

reducing (D) to the problem

$$\max_{V,g,\mu}\left\{-2f^\top g - W\mu : \underbrace{\mathfrak{b}_i^\top V\mathfrak{b}_i \leq \mu\,\forall i}_{(*)}, \left[\begin{array}{c|c} V & g \\ \hline g^\top & \frac{1}{2} \end{array}\right] \succeq 0\right\} \qquad (D')$$

Next, by the Schur Complement Lemma, semidefinite constraint in (D') is equivalent to the constraint $V \succeq \overline{V} := 2gg^\top$, and replacing V with $\overline{V}$, we clearly preserve validity of constraints $(*)$. It follows that if (V, g, μ) is feasible for (D'), so is $(\overline{V}, g, \mu)$. As a result, (D') is equivalent to the problem

$$\max_{g,\mu}\left\{-2f^\top g - W\mu : 2(\mathfrak{b}_i^\top g)^2 \leq \mu\,\forall i\right\}. \qquad (D'')$$

Due to its origin, (D'') is solvable along with (D) and shares with (D) and with (P) the optimal value Opt. It remains to note that by the Schur Complement Lemma (D'') is exactly the same as $(\overline{D})$. ■

Your next task is

3. Pass from problem $(\overline{D})$ to its semidefinite dual $(\overline{P})$ and prove that the latter problem is solvable with optimal value Opt.

At the first glance, the task seems crazy: the dual of the dual is the primal! Note, however, that $(\overline{D})$ is *not* the plain conic dual to (P) problem (D) – it is obtained from (D) by eliminating part of variables, and nobody told us that this elimination keeps the dual to $(\overline{D})$ equivalent to the dual of (D), that is, to (P).

By the same reasons as in item 1, we take upon ourselves writing down $(\overline{P})$:

$$\min_{s,t,q}\left\{\frac{1}{2}\sum_i s_i : \sum_i t_i = W, \sum_i q_i\mathfrak{b}_i = f, \left[\begin{array}{c|c} t_i & q_i \\ \hline q_i & s_i \end{array}\right] \succeq 0\,\forall i\right\} \qquad (\overline{P})$$

What is left to you is to prove that $(\overline{P})$ is solvable with optimal value Opt.

Solution: Problem $(\overline{D})$ clearly is strictly feasible, and we already know that it is solvable (and thus bounded) with optimal value Opt. By Conic Duality, $(\overline{P})$ is solvable with the same optimal value. ■

Now – the main surprise:

4. Verify that $(\overline{P})$ allows eliminating, by partial optimization, variables t_i and s_i, which reduces $(\overline{P})$ to solvable optimization problem

$$\min_q\left\{\frac{1}{2W}\left(\sum_i |q_i|\right)^2 : \sum_i q_i\mathfrak{b}_i = f\right\} \qquad (\#.1)$$

with the same optimal value Opt as all preceding problems, (P) included.

This indeed is a great surprise – $(\#.1)$ is equivalent to *Linear Programming* problem

$$\min_q\left\{\|q\|_1 : \sum_i q_i\mathfrak{b}_i = f\right\}. \qquad (\#.2)$$

Solution: Let s, t, q be a feasible solution to $(\overline{P})$, and let I be the set of indexes i with nonzero q_i; note that $I \neq \varnothing$ since, as we have assumed from the very beginning, $f \neq 0$. Note that zeroing out s_i and t_i with $i \notin I$ and increasing somehow t_i with $i \in I$ to keep $\sum_i t_i$ intact, we preserve feasibility and do not spoil the value of the objective. In the resulting feasible solution q, t', s' we have $t'_i = s'_i = 0$, $i \notin I$, $t'_i > 0$ for $i \in I$ (due to $\left[\begin{array}{c|c} t_i & q_i \\ \hline q_i & s_i \end{array}\right] \succeq 0$) and $s'_i \geq q_i^2/t'_i$ for $i \in I$ (Schur Complement Lemma); when replacing in s' entries with indexes from I with q_i^2/t'_i, we again preserve feasibility and do not spoil the objective. The bottom line is that partial optimization over s, t-components of a feasible solution (q, t, s) reduces to solving the optimization problem

$$\min_{t_i, i\in I}\left\{\frac{1}{2}\sum_{i\in I} q_i^2/t_i : t_i > 0, i \in I, \sum_{i\in I} t_i = W\right\}$$

This problem is easy to solve (see Exercise III.30); its optimal solution is given by

$$t_i = W|q_i|/\sum_{j\in I}|q_j|,\ i \in I,$$

and optimal value is $\frac{1}{2}W^{-1}(\sum_{i\in I}|q_i|)^2$. Thus, problem $(\overline{P})$ reduces to the optimization problem

$$\min_q\left\{\frac{1}{2W}(\sum_i |q_i|)^2 : \sum_i q_i\mathfrak{b}_i = f\right\}.$$

As follows from our analysis, the latter problem is solvable with optimal value Opt. ■

The challenge is, of course, to extract from optimal solution to $(\#.2)$ an optimal truss t^* – one with total bar volume W and compliance, w.r.t. load f, equal to Opt, and this is your final task:

5.1. Prove the following

Observation Let $t \geq 0$ be a nontrivial ($t \neq 0$) truss and $I = \{i : t_i > 0\}$. Consider the convex optimization problem

$$\min_q \left\{ \frac{1}{2} \sum_{i \in I} q_i^2/t_i : q_i = 0, i \notin I, \sum_i q_i \mathfrak{b}_i = f \right\} \qquad (\#.3)$$

and assume that the problem is feasible. Then

1. The problem is solvable
2. A feasible solution q to the problem is optimal if and only if for some nodal displacement $v \in \mathcal{V}$ one has
$$q_i = t_i \mathfrak{b}_i^\top v \, \forall i \qquad (\#.4)$$
3. The optimal value in the problem is nothing but the compliance of truss t w.r.t. load f.

Solution: Solvability of (#.3) is evident - the problem is feasible with bounded sublevel sets of the objective. By optimality conditions in convex minimization under linear equality constraints (see the second example after Proposition III.11.3) a feasible solution q is optimal if and only if for some $v \in \mathbf{R}^M$ one has

$$q_i/t_i = \mathfrak{b}_i^\top v, i \in I,$$

which is the same as (#.4). Assuming q optimal, (#.4) combines with $\sum_i q_i \mathfrak{b}_i = f$ to imply that

$$\sum_i t_i \mathfrak{b}_i \mathfrak{b}_i^\top v = f.$$

We see that v is the equilibrium displacement of truss t loaded by f, implying that the compliance of this truss under the load f is (see Exercise I.16.1)

$$\begin{array}{l} \frac{1}{2} v^\top f = \frac{1}{2} \sum_i t_i (\mathfrak{b}_i^\top v)^2 \\ = \frac{1}{2} \sum_{i \in I} t_i (\mathfrak{b}_i^\top v)^2 \text{ [since } t_i = 0 \text{ for } i \notin I] \\ = \frac{1}{2} \sum_{i \in I} q_i^2/t_i \text{ [since } \mathfrak{b}_i^\top v = q_i/t_i, \text{ for } i \in I] \end{array}$$

and the concluding quantity is the optimal value of (#.3). ■

5.2. Extract from optimal solution to (#.2) an optimal truss.

Solution: From our preceding considerations (#.1) is solvable with the same optimal value Opt as (P) and ($\overline{P}$) and is obtained from ($\overline{P}$) by partial optimization in s, t-variables. Let q^* be an optimal solution to (#.2), or, which is the same, to (#.1). Due to the origin of (#.1), the value Opt of its objective at q^* satisfies

$$\text{Opt} = \min_{s,t} \left\{ \frac{1}{2} \sum_i s_i : \sum_i t_i = W, \left[\begin{array}{c|c} t_i & q_i^* \\ \hline q_i^* & s_i \end{array} \right] \succeq 0 \, \forall i \leq N \right\}$$

and we know what is an optimal solution s^*, t^* to the right hand side problem: setting $I = \{i : q_i^* \neq 0\}$, we have

$$t_i^* = \left\{ \begin{array}{ll} 0, & i \notin I \\ W \frac{|q_i^*|}{\sum_{j \in I} |q_j^*|}, & i \in I \end{array} \right. , \; s_i^* = \left\{ \begin{array}{ll} 0, & i \notin I \\ \frac{(q_i^*)^2}{t_i^*}, & i \in I \end{array} \right. \qquad (\#.4)$$

Thus,

$$\text{Opt} = \frac{1}{2} \sum_{i \in I} [q_i^*]^2/t_i^* \qquad (!)$$

Now consider the optimization problem (#.3) stemming from $t = t^*$. q^* is a feasible solution to this problem with the value of the objective Opt (by (!)). By Observation, the optimal value in this problem is the compliance of t^* w.r.t. f, and since the total bar volume of t^* is W, this optimal value is $\geq$ Opt due to the origin of Opt. Thus, q^* is a feasible solution to the stemming from $t = t^*$problem (#.3), the value of the problem's objective at this solution is Opt, and the optimal value in the problem is $\geq$ Opt.

We conclude that q^* is an optimal solution to the problem in question with the value of the objective Opt, implying by Observation that Opt is the compliance of truss t^* w.r.t. f. Recalling that Opt is the optimal value in (P) and the total bar volume of t^* is W, we conclude that t^* is the t-component of an optimal solution of (P). ■

Explanation of LP miracle. Problem $(\#.1)$ was obtained from problem $(\overline{P})$ by eliminating t- and s-variables. When eliminating in $(\overline{P})$ s-variables only, we arrive at the problem

$$\min_{q,t}\left\{\sum_i \frac{q_i^2}{2t_i} : t \geq 0, \sum_i t_i = W, \sum_i q_i \mathfrak{b}_i = f\right\} \qquad (\widetilde{P})$$

where, by definition, $\frac{q_i^2}{t_i}$ is 0 when $q_i = 0$ and $+\infty$ otherwise. LP reformulation of the problem is an immediate consequence of formulation $(\widetilde{P})$. The question we address here is: can we derive $(\widetilde{P})$ directly from the first principles of Mechanics (as was the case with our initial TTD problem (P)), thus avoiding twice passing to dual which led us from (P) to $(\widetilde{P})$? As we shall see in a while, the answer is both "yes" and "no."

To interpret $(\widetilde{P})$ in terms of Mechanics, we need first of all to interpret in this way the decision variables of the problem. Looking at $(\widetilde{P})$, we can guess that t plays the role of a tentative truss; at least the constraints on t are exactly those imposed on a truss with total bar volume W. To interpret q, consider a displacement v of nodes in truss t. As we remember from the derivation of the TTD model in the preamble to Exercise I.16, the vector

$$-\sum_i [t_i \mathfrak{b}_i^\top v]\mathfrak{b}_i$$

is the reaction (block-vector of reaction forces at different nodes) resulting from nodal displacement v, and

$$t_i \mathfrak{b}_i^\top v = -S_i \delta_i, \qquad (\#.5)$$

where S_i is the cross-sectional size of i-th bar, and δ_i is the change in the bar's length caused by the displacement v of the nodes[13]. Recall that by Hooke's Law the *tension* in a bar of (pre-deformation) length d and cross-sectional size S caused by elongation/shortening of the bar by δ (that is, the reaction force caused by this deformation at bar's endpoint) is $-S\delta/d$, so that the quantities $t_i \mathfrak{b}_i^\top v$ admit, according to $(\#.5)$, transparent mechanical interpretation – these are *scaled tensions*, products of (pre-deformation) bar lengths and tensions in bars of truss t caused by displacement v of the nodes. Moreover, Mechanics says that the potential energy capacitated in elastic bar of length d and cross-sectional size S as a result of bar's elongation/shortening by δ is $\frac{1}{2}S\delta^2/d$. It follows that given a truss t and a nodal displacement v and setting $q_i = t_i \mathfrak{b}_i^\top v$, the reaction of the truss caused by nodal displacement v is $-\sum_i q_i \mathfrak{b}_i$, and the potential energy capacitated in the truss as a result of nodal displacement v is $\frac{1}{2}\sum_i q_i^2/t_i$. We see that when t is a truss, and vector q is linked to t and to some nodal displacement v by the relations

$$q_i = t_i \mathfrak{b}_i^\top v \qquad (\#.6)$$

then q_i, $-\sum_i q_i \mathfrak{b}_i$ and $\frac{1}{2}\sum_i q_i^2/t_i$ are, respectively, the scaled tensions, the reaction, and the potential energy capacitated in the truss as a result of displacement v of its nodes. Consequently, if (q,t) is a feasible solution to $(\widetilde{P})$ *and* q, t *and some nodal displacement* v *are linked by* $(\#.6)$, then v is the equilibrium displacement of truss t under load f, and the value of the objective of $(\widetilde{P})$ at the feasible solution (q,t) is the compliance of truss t w.r.t. the load f.

Our observations suggest the following mechanical interpretation of candidate solutions to $(\widetilde{P})$: t_i are bar volumes, and q_i are scaled tensions in bars. With this interpretation, the linear constraints $\sum_i q_i \mathfrak{b}_i = f$ say that the reaction compensates the external load, and the value of the objective at a feasible solution (q,t) is the compliance of truss t w.r.t load f, so that $(\widetilde{P})$ indeed is the problem of minimizing, over trusses of total volume W, the compliance of the truss w.r.t. load f. Unfortunately, this mechanical interpretation of $(\widetilde{P})$ is *completely wrong*. Indeed, the dimension of vector q is N, and typically it is much larger than the dimension M of those vectors q which could be linked to M-dimensional nodal displacements v according to $(\#.6)$ (think about Console design where $N = 3024$ and $M = 144$). In order for our guessed mechanical interpretation of $(\widetilde{P})$ to make sense, $(\widetilde{P})$ should include additional constraints stating that q is linked to t and some nodal displacement v by relations $(\#.6)$, but $(\widetilde{P})$ does *not* include this sine qua non, from the viewpoint of Mechanics, restriction! As a result, "most" of feasible solutions to $(\widetilde{P})$ make no mechanical sense – what pretends to be the vector of scaled tensions does *not* come from any deformation of the truss! Note that a straightforward attempt to include into the problem the above sine qua non restriction by adding to the design variables t, q additional design variables v, and to the constraints – equality constraints $(\#.6)$, fails – it recovers "mechanical validity" of $(\widetilde{P})$ at the disastrous, from the computational viewpoint, price – constraints $(\#.6)$ are *nonconvex* in the design variables q, t, v!

All this being said, how happens that $(\widetilde{P})$ does allow to recover the optimal truss? The explanation is: *at the optimum*, q and t indeed are linked by relations $(\#.6)$ with certain nodal displacement v; this displacement stems from the Lagrange multipliers certifying optimality of q, t (look at the justification of Observation from item 5.1). Thus, $(\widetilde{P})$ can be treated as *precise relaxation* of the "true" TTD problem: formulating the latter problem in terms of scaled tensions, bar volumes, *and* nodal displacements, which is fully legitimate from the viewpoint of Mechanics, we then relax the problem by throwing

[13] All this corresponds to the Hooke's Law in the form "reaction force caused by elongation/shortening by δ of bar with length d and cross-sectional size S is $-S\delta/d$" – the form corresponding to the linearly elastic model of truss's deformation.

away variables v and constraints ($\#.6$), thus arriving at problem $(\widetilde{P})$. This relaxation is precise in the sense that the optimal solution to the relaxed problem provably is the (q,t)-component of optimal solution to the "true" TTD problem in variables q,t,v.

Finally, we remark that while the "LP miracle" stemming from $(\widetilde{P})$ has rather restricted scope – it disappears when passing from single-load TTD with the simplest possible constraints on tentative t's to more general problems of structural design (multi-load TTD, Shape Design, etc.), these more general problems still admit "precise relaxations" of type $(\widetilde{P})$, see [BTN], and one arrives at these reformulations by strategy similar to the one we have used – start with the "natural" conic formulation (P) of the problem, pass to the conic dual (D) of (P), process (D) "on paper" by eliminating variables allowing for easy elimination, and end up by passing from the resulting reformulation $(\overline{D})$ of (D) to the conic dual $(\overline{P})$ of $(\overline{D})$.

24.5 Cone-convexity

Exercise IV.29. [elementary properties of cone-convex functions] The goal of this Exercise is to extend elementary properties of convex functions onto cone-convex mappings.

A. Let $\mathcal{X},\mathcal{Y}$ be Euclidean spaces equipped with norms $\|\cdot\|_{\mathcal{X}}$, $\|\cdot\|_{\mathcal{Y}}$. Let, next, $\mathbf{X}$ be a closed pointed cone in $\mathcal{X}$, $\mathbf{Y}$ be a closed *and pointed* cone in $\mathcal{Y}$, and $f: X\to\mathcal{Y}$ be a mapping defined on a nonempty convex set $X\subset\mathcal{X}$. Recall that for a closed and pointed cone $\mathbf{K}$ in Euclidean space $\mathcal{K}$ and $x,x'\in\mathcal{K}$, relation $x\le_{\mathbf{K}}x'$, same as $x'\ge_{\mathbf{K}}x$, means that $x'-x\in\mathbf{K}$.

Recall that f is called

• $(\mathbf{X},\mathbf{Y})$-monotone on X, if

$$\big\{x,x'\in X \text{ and } x\le_{\mathbf{X}}x'\big\}\implies f(x)\le_{\mathbf{Y}}f(x');$$

• $\mathbf{Y}$-convex on X, if

$$f(\lambda x+(1-\lambda)x')\le_{\mathbf{Y}}\lambda f(x)+(1-\lambda)f(x')$$

for every $x,x'\in X$ and $\lambda\in[0,1]$.

For example,

— an affine mapping $f(x)=Ax+a:\mathcal{X}\to\mathcal{Y}$ is $\mathbf{Y}$-convex, whatever be pointed closed cone $\mathbf{Y}$;

— when $\mathcal{Y}=\mathbf{R}$ and $\mathbf{Y}=\mathbf{R}_+$, $\mathbf{Y}$-convex on X functions are the convex, in the standard definition, real-valued functions on X.

A.1. In the situation of $\mathbf{A}$, let $\mathbf{Y}^*$ be the cone dual to $\mathbf{Y}$. For $e\in\mathcal{Y}$, let $f_e(x)=\langle e,f(x)\rangle_{\mathcal{Y}}:X\to\mathbf{R}$. Prove that f is

— $\mathbf{Y}$-convex on X if and only if the function f_e is convex on X whenever $e\in\mathbf{Y}^*$

— $(\mathbf{X},\mathbf{Y})$-monotone on X if and only if the function f_e is $\mathbf{X}$-monotone on X (i.e., $x,x'\in X, x\le_{\mathbf{X}}x'\implies f_e(x)\le f_e(x')$) for every $e\in\mathbf{Y}^*$.

Solution: Evident due to the fact that $y\in\mathbf{Y}$ if and only if $\langle e,y\rangle\ge 0$ for all $e\in\mathbf{Y}^*$; indeed, the cone $\mathbf{Y}$ is closed and therefore is dual to $\mathbf{Y}^*$.

A.2. In the situation of $\mathbf{A}$, let f be $\mathbf{Y}$-convex. Prove that f is locally bounded and locally Lipschitz continuous on the interior of X, meaning that if $\bar{X}\subset\operatorname{int}X$ is a closed and bounded set, then there exists $M<\infty$ such that $\|f(x)\|_{\mathcal{Y}}\le M$ holds for all $x\in\bar{X}$ (this is local boundedness) and there exists $L<\infty$ such that $\|f(x)-f(z')\|_{\mathcal{Y}}\le L\|x-x'\|_{\mathcal{X}}$ holds for all $x,x'\in\bar{X}$ (this is local Lipschitz continuity).

Solution: Since $\mathbf{Y}$ is pointed closed cone, the cone $\mathbf{Y}^*$ has a nonempty interior. Selecting once for ever $N:=\dim\mathcal{Y}$ linearly independent vectors $e^1,...,e^N$ in $\operatorname{int}\mathbf{Y}^*$, let us set $y^i:=\langle y,e^i\rangle_{\mathcal{Y}}$. Then, the linear mapping $y\mapsto\bar{y}(y):=[y^1;...;y^N]$ is a one-to-one linear map from $\mathcal{Y}$ onto $\mathbf{R}^N$, so that the function $|y|_\infty:=\|\bar{y}(y)\|_\infty$ is a norm on $\mathcal{Y}$. By A.1, the real valued functions $f_{e^i}(x)$ are convex on X, and therefore are locally bounded and locally Lipschitz continuous on $\operatorname{int}X$, $1\le i\le N$, implying similar properties of f w.r.t. $|\cdot|_\infty$ on $\mathcal{Y}$, and therefore w.r.t. $\|\cdot\|_{\mathcal{Y}}$. ■

B. Now let us look at elementary operations preserving cone convexity. From now on, $\mathrm{Lin}(\mathcal{X},\mathcal{Y})$ denotes the linear space of linear mappings acting from Euclidean space $\mathcal{X}$ to Euclidean space $\mathcal{Y}$. Prove the following statements:

B.1. ["nonnegative linear combinations"] Let X be a nonempty convex subset of Euclidean space $\mathcal{X}$, $\mathcal{Y}_j$, $j \le J$, and $\mathcal{Y}$ be Euclidean spaces equipped with pointed closed cones $\mathbf{Y}_j$, $\mathbf{Y}$, and $\alpha_j \in \mathrm{Lin}(\mathcal{Y}_j, \mathcal{Y})$ be "nonnegative coefficients", meaning that $\alpha_j y_j \in \mathbf{Y}$ whenever $y_j \in \mathbf{Y}_j$. When mappings $f_j(x) : X \to \mathcal{Y}_j$. are $\mathbf{Y}_j$-convex, $j \le J$, their "linear combination with coefficients α_j" – the mapping

$$f(x) = \sum_j \alpha_j f_j(x) : X \to \mathcal{Y}$$

– is $\mathbf{Y}$-convex.

Solution: For $x, x' \in X$ and $\lambda \in [0,1]$ we have

$$\lambda f(x) + (1-\lambda) f(x') - f(\lambda x + (1-\lambda)x') = \sum_j \alpha_j \underbrace{[\lambda f_j(x) + (1-\lambda) f_j(x') - f_j(\lambda x + (1-\lambda)x')]}_{\in \mathbf{Y}_j} \ge_{\mathbf{Y}} 0,$$

where the concluding $\ge_{\mathbf{Y}}$ is due to $\alpha_j y_j \ge_{\mathbf{Y}} 0$ whenever $y_j \ge_{\mathbf{Y}_j} 0$. ■

B.2. [affine substitution of variables] In the situation of $\mathbf{A}$, let $z \mapsto Az + a : \mathcal{Z} \to \mathcal{X}$ be an affine mapping, and let f be $\mathbf{Y}$-convex on X. Then, the function $g(z) := f(Az + a)$ is $\mathbf{Y}$-convex on the set $Z = \{z : Az + a \in X\}$.

Solution: evident.

B.3. [monotone composition] Let $\mathcal{U}_j$, $j \le J$, be Euclidean spaces equipped with closed pointed cones $\mathbf{U}_j$, let $\mathcal{U} = \mathcal{U}_1 \times ... \times \mathcal{U}_J$, $\mathbf{U} = \mathbf{U}_1 \times ... \times \mathbf{U}_J$, and let $\mathcal{Y}$ be an Euclidean space equipped with closed pointed cone $\mathbf{Y}$. Next, let X be nonempty convex set in Euclidean space $\mathcal{X}$, U be a nonempty convex set in $\mathcal{U}$, let $f_j(x) : X \to \mathcal{U}_j$ be $\mathbf{U}_j$-convex functions, $j \le J$, such that $f(x) = [f_1(x); ...; f_J(x)] \in U$ whenever $x \in X$. Finally, let mapping $F : U \to \mathcal{Y}$ be $(\mathbf{U}, \mathbf{Y})$-monotone and $\mathbf{Y}$-convex on U. Then the composition

$$G(x) = F(f(x)) : X \to \mathcal{Y}$$

is $\mathbf{Y}$-convex on X.

Solution: Indeed, when $x, x' \in X$ and $\lambda \in [0,1]$, setting $\bar{x} = \lambda x + (1-\lambda)x'$, $u = f(x)$, $u' = f(x')$, $\bar{u} = f(\bar{x})$, we get $\bar{x} \in X$, $u, u', \bar{u} \in U$ (since f maps X into U) and $\bar{u} \le_{\mathbf{U}} \widehat{u} := \lambda u + (1-\lambda)u'$ due to $\mathbf{U}_j$-convexity of f_j and the origin of $\mathbf{U}$. Consequently,

$$\begin{aligned} G(\bar{x}) = F(\bar{u}) &\le_{\mathbf{Y}} F(\widehat{u}) \ [\text{since } \bar{u}, \widehat{u} \in U,\ \bar{u} \le_{\mathbf{U}} \widehat{u} \text{ and } F \text{ is } (\mathbf{U}, \mathbf{Y})\text{-monotone}] \\ &\le_{\mathbf{Y}} \lambda F(u) + (1-\lambda) F(u') \ [\text{since } F \text{ is } \mathbf{Y}\text{-convex on } U] \\ &= \lambda G(x) + (1-\lambda) G(x'). \blacksquare \end{aligned}$$

C. The gradient inequality and existence of directional derivative can be extended from the usual convex functions (i.e., $\mathbf{R}_+$-convex functions taking values in $\mathbf{R}$) to the cone-convex ones. Prove the following statements:

C.1. ["gradient inequality"] In the situation of $\mathbf{A}$, let $\bar{x} \in X$ and f be $\mathbf{Y}$-convex on X and differentiable at $\bar{x}$. Then

$$\forall y \in X : f(y) \ge_{\mathbf{Y}} f(\bar{x}) + f'(\bar{x})(y - x),$$

where $f'(\bar{x})$ is the Jacobian of f at $\bar{x}$.

Solution: it suffices to apply the standard gradient inequality to convex functions f_e, $e \in \mathbf{Y}^*$, and use the same argument as when processing A.1. ■

C.2. [existence of directional derivative] In the situation of $\mathbf{A}$, let f be $\mathbf{Y}$-convex on X, let $\bar{x} \in \mathrm{int}\, X$ and $d \in \mathcal{X}$. Then

$$\exists Df(\bar{x})[d] := \lim_{t \to +0} \frac{f(\bar{x} + td) - f(\bar{x})}{t}$$

and

$$(t \ge 0 \ \&\ \bar{x} + td \in X) \implies f(\bar{x} + td) \ge_{\mathbf{Y}} f(\bar{x}) + tDf(\bar{x})[d]. \qquad (\#)$$

Besides this, as a function of $d \in \mathcal{X}$, $Df(\bar{x})[d]$ is positively homogeneous of degree 1 (i.e., $Df(\bar{x})[td] = tDf(\bar{x})[d]$ when $t \geq 0$) and $\mathbf{Y}$-convex.

Solution: By arguments completely similar to those used when justifying A.1-3, this is immediate consequence of the standard results on directional derivatives of the usual convex functions, see section 12.3.

D. Subdifferentials of the usual convex functions admit natural extensions to the cone-convex mappings. Specifically, in the situation of $\mathbf{A}$, let $\bar{x} \in X$. Let us say that $g \in \mathrm{Lin}(\mathcal{X}, \mathcal{Y})$ is a *sub-Jacobian* of f at $\bar{x}$, if

$$\forall y \in X : f(y) \geq_{\mathbf{Y}} f(\bar{x}) + g[y - x].$$

For example, C.1 says that if f is $\mathbf{Y}$-convex on X and differentiable at $\bar{x} \in X$, then the taken at x Jacobian $f'(\bar{x})$ of f is a sub-Jacobian of f at $\bar{x}$. Clearly, for a usual convex function its sub-Jacobians at a point are exactly the linear forms on $\mathcal{X}$ given by subgradients $f'(x)$ of f at x according to

$$gh = \langle f'(x), h\rangle_{\mathcal{X}},\ h \in \mathcal{X}.$$

Let $\mathscr{J}f(x)$ be the set of all sub-Jacobians of f at $x \in X$. Prove the following statements:

D.1. In the situation of $\mathbf{A}$, for $x \in X$ one has $g \in \mathscr{J}f(x)$ if and only if for every $e \in \mathbf{Y}^*$ the vector $g^*e \in \mathcal{X}$ is a subgradient of f_e at x; here for $g \in \mathrm{Lin}(\mathcal{X}, \mathcal{Y})$, $g^* \in \mathrm{Lin}(\mathcal{Y}, \mathcal{X})$ is the conjugate of g: $\langle gu, v\rangle_{\mathcal{Y}} = \langle u, g^*v\rangle_{\mathcal{X}}$ for all $u \in \mathcal{X}$, $v \in \mathcal{Y}$.

Solution: Evident due to the same argument as used when processing A.1. ■

D.2. In the situation of $\mathbf{A}$, let f be $\mathbf{Y}$-convex on X. Then
— D.2.1. For every $x \in X$, the set $\mathscr{J}f(x)$ is a closed convex subset of $\mathrm{Lin}(\mathcal{X}, \mathcal{Y})$;
— D.2.2. The mapping $x \mapsto \mathscr{J}f(x)$ is locally bounded on the interior of X, that is, for every closed and bounded set $\bar{X} \subset \mathrm{int}\, X$, the induced norms $\|g\|_{\mathcal{X},\mathcal{Y}} = \max_z\{\|gz\|_{\mathcal{Y}} : \|z\|_{\mathcal{X}} \leq 1\}$ of linear mappings $g \in \mathscr{J}f(x)$, $x \in \bar{X}$ are bounded away from $+\infty$;
— D.2.3. The multivalued mapping $x \mapsto \mathscr{J}f(x)$ is closed on $\mathrm{int}\, X$: if $x_i \in \mathrm{int}\, X$ converge as $i \to \infty$ to $\bar{x} \in \mathrm{int}\, X$ and linear mappings $g_i \in \mathscr{J}f(x_i)$ converge as $i \to \infty$ to some $\bar{g} \in \mathrm{Lin}(\mathcal{X}, \mathcal{Y})$, then $\bar{g} \in \mathscr{J}f(\bar{x})$.

Solution: D.2.1 is immediate consequence of the fact that $\mathbf{Y}$ is a closed cone; D.2.2-3 are readily given by local Lipschitz continuity of f on $\mathrm{int}\, X$, see A.2. ■

The most attractive property of subgradients of the usual convex function is their existence, at least at interior points of the function's domain. This fact extends to the cone-convex mappings. Prove the following statements:

D.3. [existence of sub-Jacobians] In the situation of $\mathbf{A}$, let $\bar{x} \in \mathrm{int}\, X$ and f be $\mathbf{Y}$-convex on X. Then $\mathscr{J}f(\bar{x})$ is nonempty.

Solution: This is the only claim which seemingly cannot be extracted more or less automatically from standard facts about the usual convex functions. Moreover, the Separation Theorem underlying the existence of subgradients of the usual convex functions at interior points of their domains seemingly does not help now. Fortunately, there is an easily implementable alternative as follows.
For $\epsilon > 0$, let $X_\epsilon = \{x \in X : \|y - x\|_{\mathcal{X}} \leq \epsilon \implies y \in \mathrm{int}\, X\}$ and $\delta_\epsilon(x)$ be a nonnegative C^∞ function such that $\delta_\epsilon(x) = 0$ when $\|x\|_{\mathcal{X}} \geq \epsilon$ and $\int_{\mathcal{X}} \delta_\epsilon(x)dx = 1$. Clearly, for small $\epsilon > 0$ X_ϵ is a nonempty open convex set, and the function

$$f_\epsilon(x) := \int_{\mathcal{X}} f(x - y)\delta_\epsilon(y)dy$$

with the domain X_ϵ is well defined, continuously differentiable and $\mathbf{Y}$-convex on its domain. Besides this, for a convex compact set $\bar{X} \subset \mathrm{int}\, X$ such that $\bar{x} \in \mathrm{int}\, \bar{X}$ we have $\bar{X} \subset X_\epsilon$ for all small enough positive ϵ, and for those ϵ the functions f_ϵ are uniformly in ϵ Lipschitz continuous on $\bar{X}$. From this latter observation it follows that the Jacobians $f'_\epsilon(\bar{x})$ are uniformly in ϵ bounded, which in turn implies that for a properly selected $\epsilon_i \to +0$, $i \to \infty$, the linear mappings $g_i := f'_{\epsilon_i}(\bar{x})$ converge as $i \to \infty$ to some $\bar{g} \in \mathrm{Lin}(\mathcal{X}, \mathcal{Y})$. Let us prove that $\bar{g} \in \mathscr{J}f(\bar{x})$, implying that $\mathscr{J}f(\bar{x})$ is nonempty. Indeed, in view of A.2 the

functions $f^i := f_{\epsilon_i}$ converge as $i \to \infty$, uniformly on compact subsets of $\operatorname{int} X$, to f. Then, by C.1, we have

$$y \in X_{\epsilon_i} \implies f_i(y) \geq_{\mathbf{Y}} f(\bar{x}) + g_i(y - \bar{x}),$$

implying in view of the outlined convergencies that

$$f(y) \geq_{\mathbf{Y}} f(\bar{x}) + g[y - \bar{x}] \ \forall y \in \operatorname{int} X.$$

The only remaining task is to extend the latter relation from $y \in \operatorname{int} X$ to $y \in X$. Passing from $f : X \to \mathcal{Y}$ to $\overline{f}(x) := f(x) - [f(\bar{x}) + g[x - \bar{x}]]$, which, of course, is $\mathbf{Y}$-convex on X along with f, we get

$$\overline{f}(y) \geq_{\mathbf{Y}} \overline{f}(\bar{x}) = 0, \ \forall y \in \operatorname{int} X, \tag{!}$$

and what we need to prove is that $\overline{f}(y) \geq_{\mathbf{Y}} \overline{f}(\bar{x})$ for all $y \in Y$. Let $y \in X$. Using the definition of the directional derivative, we observe that (!) implies that $D\overline{f}(\bar{x})[y - \bar{x}] \geq_{\mathbf{Y}} 0$, whence by (#) one has $\overline{f}(y) \geq_{\mathbf{Y}} \overline{f}(\bar{x})$. ■

For a real-valued convex function f and $x \in \operatorname{int} \operatorname{Dom} f$, $d \in \mathcal{X}$, one has $Df(x)[d] = \max_{y \in \partial f(x)} \langle y, d \rangle_{\mathcal{X}}$. A similar fact holds true for cone-convex functions:

D.4. In the situation of $\mathbf{A}$, let f be $\mathbf{Y}$-convex on X. Let also $\bar{x} \in \operatorname{int} X$ and $d \in \mathcal{X}$. Then for properly selected $g \in \mathscr{J}f(\bar{x})$ one has

$$Df(\bar{x})[d] = gd,$$

while for every $g' \in \mathscr{J}f(\bar{x})$ one has

$$Df(\bar{x})[d] \geq_{\mathbf{Y}} g'd.$$

Solution: For $g' \in \mathscr{J}f(\bar{x})$ and $t > 0$ such that $\bar{x} + td \in X$ we have

$$f(\bar{x} + td) - f(\bar{x}) \geq_{\mathbf{Y}} tg'd,$$

whence, dividing by t and passing to limit as $t \to +0$, we get

$$Df(\bar{x})[d] \geq_{\mathbf{Y}} g'd.$$

On the other hand, let $t_0 > t_1 > t_2 > ... > 0$ be such that $x_i := \bar{x} + t_i d \in \operatorname{int} X$ and $t_i \to 0$, $i \to \infty$. By D.3, there exists $g_i \in \mathscr{J}f(x_i)$; by D.2.2 the sequence g_i is bounded, so that passing to a subsequence, we can assume that $g_i \to g$ as $i \to \infty$; by D.2.3, $g \in \mathscr{J}f(\bar{x})$. Since $g_i \in \mathscr{J}f(x_i)$, we have

$$f(x_i) - f(\bar{x}) \leq_{\mathbf{Y}} g_i[x_i - \bar{x}],$$

whence

$$g_i d \geq_{\mathbf{Y}} t_i^{-1}[f(x_i) - f(\bar{x})].$$

As $i \to \infty$, the left hand side in this $\geq_{\mathbf{Y}}$-inequality tends to gd, and the right hand side to $Df(\bar{x})[d]$. Thus, $g'd \leq_{\mathbf{Y}} Df(\bar{x})[d]$ for all $g' \in \mathscr{J}f(\bar{x})$ and $gd \geq_{\mathbf{K}} Df(\bar{x})[d]$ for some $g \in Df(\bar{x})[d]$; in particular, $gd = Df(\bar{x})[d]$ (recall that $\mathbf{Y}$ is pointed). ■

There is a natural relation between sub-Jacobians of $\mathbf{Y}$-convex function f and subgradients of functions $f_e = \langle e, f \rangle_{\mathcal{Y}}$, $e \in \mathbf{Y}^*$:

D.5. In the situation of $\mathbf{A}$, let f be $\mathbf{Y}$-convex on X and $\bar{x} \in \operatorname{int} X$. For $e \in \mathbf{Y}^*$, $h \in \partial f_e(\bar{x})$ (that is, $f_e(y) \geq f_e(\bar{x}) + \langle h, y - \bar{x} \rangle_{\mathcal{X}}$ for all $y \in X$) if and only if $h = g^*e$ for some $g \in \mathscr{J}f(\bar{x})$.

Solution: In one direction: when $e \in \mathbf{Y}^*$ and $h = g^*e$ for $g \in \mathscr{J}f(\bar{x})$, we have for every $y \in X$:

$$f(y) \geq_{\mathbf{Y}} f(\bar{x}) + g[y - x] \implies \langle e, f(y) \rangle_{\mathcal{Y}} \geq \langle e, f(\bar{x}) \rangle_{\mathcal{Y}} + \langle e, g[y - \bar{x}] \rangle_{\mathcal{Y}} \iff f_e(y) \geq f_e(\bar{x}) + \langle g^*e, y - \bar{x} \rangle_{\mathcal{X}}.$$

In the opposite direction: let $e \in \mathbf{Y}^*$ and $h \in \partial f_e(\bar{x})$. By D.2 and D.3, the set $\mathcal{J} = \mathscr{J}f(\bar{x})$ is a nonempty closed and bounded convex set in $\operatorname{Lin}(\mathcal{X}, \mathcal{Y})$. Thus, the set $\mathcal{I} := \{g^*e : g \in \mathcal{J}\}$ is a nonempty closed

and bounded convex set in $\mathcal{X}$. Assume for contradiction that $h \notin \mathcal{I}$. Then, by Separation Theorem there exists $d \in \mathcal{X}$ such that

$$\langle h, d\rangle_{\mathcal{X}} > \max_{g\in\mathcal{J}} \langle g^* e, d\rangle_{\mathcal{X}} = \max_{g\in\mathcal{J}} \langle e, gd\rangle_{\mathcal{Y}}. \qquad (*)$$

As $h \in \partial f_e(\bar{x})$, for all small enough $t > 0$ we have

$$f_e(\bar{x} + td) - f_e(\bar{x}) \geq t\langle h, d\rangle_{\mathcal{X}},$$

whence $Df_e(\bar{x})[d] \geq \langle h, d\rangle_{\mathcal{X}}$. We clearly have $Df_e(\bar{x})[d] = \langle e, Df(\bar{x})[d]\rangle_{\mathcal{Y}}$, and we arrive at

$$\langle h, d\rangle_{\mathcal{X}} \leq \langle e, Df(\bar{x})[d]\rangle_{\mathcal{Y}}.$$

By D.4, we have $Df(\bar{x})[d] = \bar{g}d$ for some $\bar{g} \in \mathcal{J}f(\bar{x})$, so that $\langle h, d\rangle_{\mathcal{X}} \leq \langle e, \bar{g}d\rangle_{\mathcal{Y}}$, contradicting $(*)$. ■

Finally, the chain rule:

D.6. [chain rule] Let $\mathcal{U}_j$, $j \leq J$, be Euclidean spaces equipped with closed pointed cones $\mathbf{U}_j$, let $\mathcal{U} = \mathcal{U}_1 \times ... \times \mathcal{U}_J$, $\mathbf{U} = \mathbf{U}_1 \times ... \times \mathbf{U}_J$, and let $\mathcal{Y}$ be an Euclidean space equipped with closed pointed cone $\mathbf{Y}$. Next, let X be nonempty convex set in Euclidean space $\mathcal{X}$, U be a nonempty convex set in $\mathcal{U}$, let $f_j(x) : X \to \mathcal{U}_j$ be $\mathbf{U}_j$-convex on X functions, $j \leq J$, such that $f(x) = [f_1(x); ...; f_J(x)] \in U$ whenever $x \in X$. Finally, let mapping $F : U \to \mathcal{Y}$ be $(\mathbf{U}, \mathbf{Y})$-monotone and $\mathbf{Y}$-convex on U. As we know from B.3, the composition

$$G(x) = F(f(x)) : X \to \mathcal{Y}$$

is $\mathcal{Y}$-convex on X. Now let $\overline{x} \in \operatorname{int} X$, $\overline{u}_j = f_j(\overline{x})$ be such that $\overline{u} = [\overline{u}_1; ...; \overline{u}_J] \in \operatorname{int} U$. Finally, let $g_j \in \mathcal{J}f_j(\overline{x})$, $j \leq J$, and $g \in \mathcal{J}F(\overline{u})$. Then the linear mapping $[u_1; ...; u_J] \mapsto g[u_1; ...; u_J]$ is $(\mathbf{U}, \mathbf{Y})$-monotone, and the linear mapping

$$h \mapsto \widehat{g}h := g[g_1 h; ...; g_J h] : \mathcal{X} \to \mathcal{Y}$$

is sub-Jacobian of G at $\overline{x}$.

Solution: Indeed, let $\overline{V}$ be a convex neighborhood of $\overline{x}$ such that the images of $\overline{V}$ under the mapping $f(\cdot)$ and under the linear mapping

$$\overline{f}(x) := f(\overline{x}) + [g_1[x - \overline{x}]; ...; g_J[x - \overline{x}]]$$

belong to $\operatorname{int} U$ (such a neighborhood exists due to $f(\overline{x}) = \overline{f}(\overline{x}) \in \operatorname{int} U$ combined with continuity of $\overline{f}$ (evident) and f (by A.2) at $\overline{x}$). Let also V^j, $j = 1, ..., J$, be convex neighborhoods of origins in $\mathcal{U}_j$ such that $\overline{u} + V^1 \times ... \times V^J \subset U$. For $d = [d_1; ...; d_J]$ with $d_j \in V^j \cap \mathbf{U}_j$ for $j \leq J$ we have

$$-gd + F(\overline{u}) \leq_{\mathbf{Y}} F(\overline{u} - d),$$

whence $gd \geq_{\mathbf{Y}} 0$ by $(\mathbf{U}, \mathbf{Y})$-monotonicity of F. Thus, $gd \geq_{\mathbf{Y}} 0$ for all $d \in \mathbf{U}$ of small enough norm, implying that $gd \geq_{\mathbf{Y}} 0$ for all $d \in \mathbf{U}$, as claimed.

When $x \in \overline{V}$, we have $\overline{f}(x) \leq_{\mathbf{U}} f(x)$, since g_j are sub-Jacobians of f_j at $\overline{x}$. Due to the $(\mathbf{U}, \mathbf{Y})$-monotonicity of F, we conclude that

$$\begin{aligned} G(x) = F(f(x)) &\geq_{\mathbf{Y}} F(\overline{f}(x)) = F\left(\overline{u} + [g_1[x - \overline{x}]; ...; g_J[x - \overline{x}]]\right) \\ &\geq_{\mathbf{Y}} F(f(\overline{x})) + g[g_1[x - \overline{x}]; ...; g_J[x - \overline{x}]] \text{ [since } g \in \mathcal{J}F(\overline{u})] \\ &= G(\overline{x}) + \widehat{g}[x - \overline{x}]. \end{aligned}$$

Thus, for x in a neighborhood $\overline{V}$ of $\overline{x} \in \operatorname{int} X$ we have

$$G(x) \geq_{\mathbf{Y}} G(\overline{x}) + \widehat{g}[x - \overline{x}].$$

It remains to prove that the latter relation holds true for all $x \in X$, not for only $x \in \overline{V}$. This can be done in the same way as when justifying D.3: the mapping $\overline{G}(x) := G(x) - [G(\overline{x}) + \widehat{g}[x - \overline{x}]] : X \to \mathcal{Y}$ which is $\mathbf{Y}$-convex along with G satisfies

$$\overline{G}(x) \geq_{\mathbf{Y}} \overline{G}(\overline{x}) = 0 \qquad (!!)$$

for x from a neighborhood of $\overline{x}$, and we want to prove that in fact (!!) holds true for all $x \in X$. Indeed, (!!) implies that for every $x \in X$ we have $D\overline{G}(\overline{x})[x-\overline{x}] \geq_{\mathbf{Y}} 0$, whence $\overline{G}(x) \geq_{\mathbf{Y}} \overline{G}(\overline{x}) = 0$ for $x \in X$ by (#), so that $G(x) \geq_{\mathbf{Y}} G(\overline{x}) + \widehat{g}[x-\overline{x}]$ for all $x \in X$. ■

Exercise IV.30. Univariate function $f(x) = x^{-1/2} : \{x > 0\} \to \mathbf{R}$ is nonincreasing and convex, and $\nabla f(x) = -x^{-3/2}/2$, $x > 0$. Now let P be $m \times n$ matrix of rank m.

1. Prove that the mapping $F(X) = [PXP^\top]^{-1/2} : \mathbf{S}^n_{++} \to \mathbf{S}^m$, where $\mathbf{S}^n_{++} = \operatorname{int} \mathbf{S}^n_+ = \{X \in \mathbf{S}^n : X \succ 0\}$, is $(\mathbf{S}^n_+, \mathbf{S}^m_+)$-antimonotone and $\mathbf{S}^m_+$-convex
2. Assuming $P = I_2$, compute numerically $F(X)$ and $dF(X)[dX]$ for $X = \begin{bmatrix} 2 & -1 \\ -1 & 1 \end{bmatrix}$ and $dX = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$. For the above X, compute also the Jacobian J of F at X – the matrix of the linear mapping $dX \mapsto DF(X)[dX] : \mathbf{S}^2 \to \mathbf{S}^2$ – in the basis $[1,0;0,0]$, $[0,0;0,1]$, $[0,1/\sqrt{2};1/\sqrt{2},0]$ of $\mathbf{S}^2$.
3. How the "Gradient inequality" (Exercise IV.29.C.1) for the $\mathbf{S}^n_+$-convex mapping F looks like?

Solution:

1: Let $G(X) = -(PXP^\top)^{1/2} : \mathbf{S}^n_{++} \to \mathbf{S}^m$ and $H(U) = [-U]^{-1} : -\mathbf{S}^m_{++} \to \mathbf{S}^m$. The mapping $U \mapsto \mathcal{P}(X) := PXP^\top$ with the domain $\mathbf{S}^n_{++}$ maps this domain into $\mathbf{S}^m_{++}$ (since $\mathrm{Ker} P^\top = \{0\}$ due to $\mathrm{rank}(P) = m$), clearly is $(\mathbf{S}^n_+, \mathbf{S}^m_+)$-monotone and, as any linear mapping, is $\mathbf{S}^m_+$-convex; the mapping $U \mapsto \mathcal{G}(U) = -U^{1/2} : \mathbf{S}^m_+ \to \mathbf{S}^m$ is $(\mathbf{S}^m_+, \mathbf{S}^m_+)$-antimonotone and $\mathbf{S}^m_+$-convex by Example IV.20.5. Consequently, the map $X \mapsto G(X) = \mathcal{G}(\mathcal{P}(X)) : \mathbf{S}^n_+ \to \mathbf{S}^m$ is $\mathbf{S}^m_+$-convex (Rule **A.3** in section 20.3) and clearly is $(\mathbf{S}^n_+, \mathbf{S}^m_+)$-antimonotone (as superposition of $(\mathbf{S}^m_+, \mathbf{S}^m_+)$-antimonotone mapping $\mathcal{G}$ and $(\mathbf{S}^n_+, \mathbf{S}^m_+)$-monotone mapping $\mathcal{P}$). Next, mapping $U \mapsto U^{-1} : \mathbf{S}^m_{++} \to \mathbf{S}^m$ is $\mathbf{S}^m_+$-convex and $(\mathbf{S}^m_+, \mathbf{S}^m_+)$-antimonotone (Example IV.20.4), whence the mapping $U \mapsto H(U) := [-U]^{-1} : [-\mathbf{S}^m_{++}] \to \mathbf{S}^m$ is $\mathbf{S}^m_+$-convex (as the superposition of $\mathbf{S}^m_+$-convex mapping $U \mapsto U^{-1} : \mathbf{S}^m_{++} \to \mathbf{S}^m$ and linear mapping $U \mapsto -U : \mathbf{S}^m \to \mathbf{S}^m$). In addition $H(U)$ is $(\mathbf{S}^m_+, \mathbf{S}^m_+)$-monotone on its domain $-\mathbf{S}^m_{++}$ in view of $(\mathbf{S}^m_+, \mathbf{S}^m_+)$-antimonotonicity of the mapping $U \mapsto -U$ and $(\mathbf{S}^m_+, \mathbf{S}^m_+)$-antimonotonicity of the mapping $U \mapsto U^{-1}$ on the domain $\mathbf{S}^m_{++}$. The indicated cone-convexity and cone-monotonicity properties of the mapping $G(\cdot)$ and $H(\cdot)$ imply, in view of Rule **B** in section 20.3, that $F(X) = H(G(X))$ is $\mathbf{S}^m_+$-convex and $(\mathbf{S}^n_+, \mathbf{S}^m_+)$-antimonotone.

2: When justifying Examples IV.20.4 and IV.20.5, we have seen that the mappings $H(\cdot)$ and $\mathcal{G}(\cdot)$ are differentiable on the domains $-\mathbf{S}^m_{++}$, resp., $\mathbf{S}^m_{++}$, and

$$\begin{array}{rcl} DH(U)[dU] & = & U^{-1}dUU^{-1},\ U \in -\mathbf{S}^m_{++}, dU \in \mathbf{S}^m, \\ D\mathcal{G}(V)[dV] & = & -\int\limits_0^\infty \exp\{-V^{1/2}t\}dV \exp\{-V^{1/2}t\}dt,\ V \in \mathbf{S}^m_{++}, dV \in \mathbf{S}^m, \end{array}$$

implying by Chain rule that for $X \in \mathbf{S}^n_{++}, dX \in \mathbf{S}^n$ one has

$$DF(X)[dX] = -\left[PXP^\top\right]^{-1/2}\left[\int\limits_0^\infty \exp\{-[PXP^\top]^{1/2}t\}PdXP^\top \exp\{-[PXP^\top]^{1/2}t\}dt\right]\left[PXP^\top\right]^{-1/2}$$

3: Our computation yields the following results (rounded to 4 digits after the dot):

$$F = \begin{bmatrix} 0.8944 & 0.4472 \\ 0.4472 & 1.3416 \end{bmatrix},\ DF(X)[dX] = \begin{bmatrix} -0.6265 & -0.9846 \\ -0.9846 & -1.1634 \end{bmatrix},$$
$$J = \begin{bmatrix} -0.4025 & -0.2683 & -0.4427 \\ -0.2683 & -1.2970 & -0.8222 \\ -0.4427 & -0.8222 & -0.9839 \end{bmatrix}.$$

4: $\forall (X, Y \in \mathbf{S}^n_{++}) : [PYP^\top]^{-1/2} \succeq [PXP^\top]^{-1/2} + DF(X)[Y-X]$ with $DF(X)[\cdot]$ as described in item 2.

24.6 Around conic representations of sets and functions

24.6.1 Conic representations: definitions

Let $\mathfrak{K}$ be a family of regular cones in Euclidean spaces which contains the nonnegative ray $\mathbf{R}_+$ and is closed with respect to taking finite direct products and passing from a cone to its dual. Instructive examples are the families $\mathfrak{R}$ of nonnegative orthants, $\mathfrak{L}$ of finite direct products of Lorentz cones, and $\mathfrak{S}$ of finite direct products of semidefinite cones.

• A $\mathfrak{K}$-*representation* ($\mathfrak{K}$-r.) of a set $X \subset \mathbf{R}^n$ is its representation of the form

$$X = \{x \in \mathbf{R}^n : \exists u : Px + Qu - r \in \mathbf{K}\} \tag{23.2}$$

with $\mathbf{K} \in \mathfrak{K}$ – representation of X as the projection of the solution set of conic inequality $Px+Qu \geq_{\mathbf{K}} r$ in variables x, u onto the plane of x-variables where X lives. A set X admitting conic representation with cone from $\mathfrak{K}$ is called $\mathfrak{K}$-*representable* ($\mathfrak{K}$-r for short).

• A $\mathfrak{K}$-*representation of a function* $f : \mathbf{R}^n \to \mathbf{R} \cup \{+\infty\}$ is, by definition, $\mathfrak{K}$-representation of the epigraph of f:

$$[t; x] \in \mathrm{epi}\{f\} := \{[x; t] : t \geq f(x)\} \iff \exists u : Px + tp + Qu - r \in \mathbf{K} \text{ with } \mathbf{K} \in \mathfrak{K}.$$

Functions admitting $\mathfrak{K}$-representation are called $\mathfrak{K}$-*representable* ($\mathfrak{K}$-r for short)

We are already acquainted with $\mathfrak{R}$-representability – it is that was called polyhedral representability. By Fourier-Motzkin elimination, polyhedral representable sets $X \subset \mathbf{R}^n$ admit polyhedral representations not involving additional variables u, and similarly for $\mathfrak{R}$-representable functions; this is not the case for more general families $\mathfrak{K}$, like families $\mathfrak{L}$ of Lorentz- and $\mathfrak{S}$ of semidefinite-representable sets.

The following exercise explains what is the rationale underlying the above restrictions on $\mathfrak{K}$ and why we are interested in $\mathfrak{K}$-representations.

Exercise IV.31. Check that

1. Every finite system $P_0 y \geq r_0,\ P_i y - r_i \in \mathbf{K}_i,\ i \leq I$, of scalar linear inequalities and conic inequalities, involving cones from $\mathfrak{K}$, in variables y is equivalent to a single conic inequality, with cone from $\mathfrak{K}$, in these variables:

$$\begin{array}{c} \{P_0 y - r_0 \geq 0, P_i y - r_i \in \mathbf{K}_i, 1 \leq i \leq I\} \\ \iff \left\{[P_0; P_1; ...; P_I]y - [r_0; r_1; ...; r_I] \in \mathbf{K} := \underbrace{\mathbf{R}_+ \times ... \times \mathbf{R}_+}_{\dim r_0 \text{ times}} \times \mathbf{K}_1 \times \mathbf{K}_2 \times ... \times \mathbf{K}_I\right\} \end{array}$$

   and $\mathbf{K} \in \mathfrak{K}$ (since $\mathbf{R}_+ \in \mathfrak{K}$ and $\mathfrak{K}$ is closed w.r.t. taking finite direct products).
   As a result, representation of a set X as

$$X = \{x : \exists u : P_0 x + Q_0 u - r_0 \geq 0, P_i x + Q_i u - r_i \in \mathbf{K}_i, 1 \leq i \leq I\} \qquad [\mathbf{K}_i \in \mathfrak{K}] \quad (!)$$

   – as the projection of the solution set of a finite system of linear and $\mathfrak{K}$-conic inequalities in variables x, u onto the plane of x-variables where X lives, can be straightforwardly converted into a $\mathfrak{K}$-r. of X.

Important: Item 1 allows us from now on to refer to representations of the form (!) as to $\mathfrak{K}$-representations of X, skipping (always straightforward and purely mechanical) conversion of such a representation into the "canonical" representation (23.2).

2. $\mathfrak{K}$-r. of a function straightforwardly induces $\mathfrak{K}$-r.'s of its sublevel sets:

$$\begin{array}{c} \left\{\{t \geq f(x)\} \iff \{\exists u : Px + tp + Qu - r \in \mathbf{K}\}\right\} \\ \implies X_a := \{x : f(x) \leq a\} = \{x : \exists u : Px + Qu - [r - ap] \in \mathbf{K}\} \end{array} \qquad [a \in \mathbf{R}, \mathbf{K} \in \mathfrak{K}]$$

3. Given $\mathfrak{K}$-representations of a set $X \subset \mathbf{R}^n$ and a function $f : \mathbf{R}^n \to \mathbf{R} \cup \{+\infty\}$:

$$\begin{array}{rcll} X & = & \{x \in \mathbf{R}^n : \exists u : P_X x + Q_X u - r_X \in \mathbf{K}_X\}, & \\ \mathrm{epi}\{f\} & = & \{[x; t] : \exists v : P_f x + tp_f + Q_f v - r_f \in \mathbf{K}_f\} & [\mathbf{K}_X \in \mathfrak{K}, \mathbf{K}_f \in \mathfrak{K}] \end{array}$$

we can straightforwardly convert the optimization problem

$$\min_{x\in X} f(x) \qquad (*)$$

into conic problem on a cone from $\mathfrak{K}$, namely, the problem

$$\min_{x,t,u,v}\Big\{t:\quad A[x;t;u;v]-b \\ :=[P_Xx+Q_Xu;P_fx+tp_f+Q_fv]-[r_X;r_f]\in\underbrace{\mathbf{K}:=\mathbf{K}_X\times\mathbf{K}_f}_{\in\mathfrak{K}}\Big\}$$

As a result, a solver $\mathcal{S}$ capable to solve conic problems on cones from $\mathfrak{K}$ can be straightforwardly utilized when solving problems $(*)$ with X and f given by $\mathfrak{K}$-r.'s.

4. Given a conic problem

$$\min_x\Big\{c^Tx: Ax-b\in\mathbf{K}, Rx\geq r\Big\} \qquad (P)$$

on a cone from $\mathfrak{K}$, its conic dual – the conic problem

$$\max_{y,z}\big\{\langle b,y\rangle+r^Tz: A^*y+R^Tz=c, y\in\mathbf{K}_*, z\geq 0\big\} \\ \left[\begin{array}{c}\langle\cdot,\cdot\rangle \text{ is the inner product in the Euclidean space where } \mathbf{K} \text{ lives, } \mathbf{K}_* \text{ is the cone dual to } \mathbf{K},\\ A^* \text{ is the conjugate of } A\text{: } \langle Ax,y\rangle\equiv x^TA^*y\ \forall x,y\end{array}\right] \qquad (D)$$

also is a conic problem on a cone from $\mathfrak{K}$ (since $\mathfrak{K}$ is closed w.r.t. passing from a cone to its dual and contains nonnegative orthants).

Solution: This is straightforward – substitute "$\mathfrak{K}$-representation" with the definition of this notion.

Note that the option mentioned in the last item of Exercise IV.31 is implemented in "CVX: MATLAB software for disciplined convex programming" due to M. Grant and S. Boyd `http://cvxr.com/cvx` – second to none in its scope and user-friendliness tool for numerical processing of well-structured convex problems, the underlying family $\mathfrak{K}$ being the semidefinite family $\mathfrak{S}$. We conclude that it makes sense to develop a kind of calculus allowing to recognize $\mathfrak{K}$-representability of sets/functions and to build, when possible, their $\mathfrak{K}$-representations. The desired calculus exists and is pretty simple, general and fully algorithmic. The goal of subsequent exercises is to make you acquainted with the most frequently used elements of this calculus; for more on this subject, see [BTN].

24.6.2 Conic representability: elementary calculus

Elementary calculus of conic representability is completely similar to calculus of polyhedral representations from section 3.3.

Exercise IV.32. [elementary calculus of $\mathfrak{K}$-representable sets] Check that basic convexity-preserving[14] operations with sets preserve $\mathfrak{K}$-representability. Specifically,

1. Finite intersection of $\mathfrak{K}$-r sets $X_i=\{x\in\mathbf{R}^n:\exists u^i: P_ix+Q_iu^i-r_i\in\mathbf{K}_i\}$, $i\leq I$ (here and in what follows all cones involved are from $\mathfrak{K}$) is $\mathfrak{K}$-r.:

$$\bigcap_{i\leq I}X_i= \quad \{x\in\mathbf{R}^n:\exists u=[u^1;...;u^I]: Px+Qu-r \\ :=[P_1x+Q_1u^1;...;P_Ix+Q_Iu^I]-[r_1;...;r_I]\in\underbrace{\mathbf{K}:=\mathbf{K}_1\times...\times\mathbf{K}_I}_{\in\mathfrak{K}}\}$$

2. Direct product of finitely many $\mathfrak{K}$-r sets $X_i=\{x\in\mathbf{R}^n:\exists u^i: P_ix+Q_iu^i-r_i\in\mathbf{K}_i\}$, $i\leq I$ is $\mathfrak{K}$-r:

$$X_1\times...\times X_I=\{x=[x^1;...;x^I]:\exists u=[u^1;...;u^I]: \\ Px+Qu-r:=[P_1x^1+Q_1u^1;...;P_Ix^I+Q_Iu^I]-[r_1;...;r_I]\in\underbrace{\mathbf{K}:=\mathbf{K}_1\times...\times\mathbf{K}_I}_{\in\mathfrak{K}}\}$$

[14] "convexity-preserving" is crucial – clearly, $\mathfrak{K}$-r sets and functions must be convex!

3. Affine image $Y = \{y = Ax + b : x \in X\}$ of $\mathfrak{R}$-r set $X = \{x \in \mathbf{R}^n : \exists u : Px + Qu - r \in \mathbf{K}\}$ is $\mathfrak{R}$-r:
$$Y = \{y : \exists [x; u] : Ax + b = y, Px + Qu - r \in \mathbf{K}\}$$
is the projection onto the y-plane of a set given by explicit finite system of linear and $\mathfrak{R}$-conic inequalities and as such admits an explicit $\mathfrak{R}$-r. by item 1 of Exercise IV.31.
4. Inverse affine image $Y = \{y : Ay + b \in X\}$ of $\mathfrak{R}$-r. set $X = \{x \in \mathbf{R}^n : \exists u : Px + Qu - r \in \mathbf{K}\}$ is $\mathfrak{R}$-r.:
$$Y = \{y : \exists u : PAy + Qy - [r - Pb] \in \mathbf{K}\}.$$
5. The arithmetic sum $X = X_1 + ... + X_I$ of $\mathfrak{R}$-r sets $X_i = \{x \in \mathbf{R}^n : \exists u^i : P_i x + Q_i u^i - r_i \in \mathbf{K}_i\}$, $i \leq I$, is $\mathfrak{R}$-r:
$$X = \{x : \exists [x^1; ...; x^I; u^1; ...; u^I] : x - \sum_i x^i = 0, P_i x^i + Q_i u^i - r_i \in \mathbf{K}_i, i \leq I\}$$
and it remains to apply item 1 of Exercise IV.31.

Solution: This is straightforward – substitute "$\mathfrak{R}$-representation" with the definition of this notion.

Exercise IV.33. [elementary calculus of $\mathfrak{R}$-representable functions] Check that the following convexity-preserving operations with functions preserve $\mathfrak{R}$-representability:

0. Restricting onto $\mathfrak{R}$-r set: $\mathfrak{R}$-r. $t \geq f(x) \iff \exists u : P_f x + t_f p + Q_f u - r_f \in \mathbf{K}_f$ of a function $f : \mathbf{R}^n \to \mathbf{R} \cup \{+\infty\}$ taken together with $\mathfrak{R}$-r. $X = \{x \in \mathbf{R}^n : \exists v : P_X x + Q_X v - r_X \in \mathbf{K}_X\}$ of a set $X \subset \mathbf{R}^n$ induce $\mathfrak{R}$-r.
$$t \geq f\big|_X(x) \iff \exists u, v : P_f x + tp_f + Q_f u - r_f \in \mathbf{K}_f,\ P_X x + Q_X v - r_X \in \mathbf{K}_X$$
of the restriction $f\big|_X(x) = \begin{cases} f(x) & , x \in X \\ +\infty & , x \notin X \end{cases}$ of f onto X
1. Taking linear combination $\sum_{i=1}^I \lambda_i f_i$ with positive coefficients:
$$\begin{array}{c} t \geq f_i(x) \iff \exists u^i : P_i x + tp_i + Q_i u^i - r_i \in \mathbf{K}_i,\ i \leq I \\ \Downarrow \\ t \geq f(x) := \sum_{i=1}^I \lambda_i f_i(x) \iff \exists [t_1; ...; t_I; u^1; ...; u^I] : t \geq \sum_i \lambda_i t_i, P_i x + t_i p_i + Q_i u^i - r_i \in \mathbf{K}_i,\ i \leq I \end{array}$$
2. Direct summation:
$$\begin{array}{c} t \geq f_i(x^i) \iff \exists u^i : P_i x^i + tp_i + Q_i u^i - r_i \in \mathbf{K}_i,\ i \leq I \\ \Downarrow \\ t \geq f(x^1, ..., x^I) := \sum_{i=1}^I f_i(x^i) \iff \exists [t_1; ...; t_I; u^1; ...; u^I] : t \geq \sum_i t_i, P_i x^i + t_i p_i + Q_i u^i - r_i \in \mathbf{K}_i,\ i \leq I \end{array}$$
3. Taking finite maxima:
$$\begin{array}{c} t \geq f_i(x) \iff \exists u^i : P_i x + tp_i + Q_i u^i - r_i \in \mathbf{K}_i,\ i \leq I \\ \Downarrow \\ t \geq f(x) := \max\limits_{i \leq I} f_i(x) \iff \exists [u^1; ...; u^i] : P_i x + tp_i + Q_i u^i - r_i \in \mathbf{K}_i,\ i \leq I \end{array}$$
4. Affine substitution of variables:
$$\begin{array}{c} t \geq f(x) \iff \exists u : Px + tp + Qu - r \in \mathbf{K} \\ \Downarrow \\ t \geq g(y) := f(Ay + b) \iff \exists u : PAu + tp + Qu - [r - Pb] \in \mathbf{K} \end{array}$$
In fact, claims in items 1–4 are special cases of the following observation:
5. Monotone superposition: let functions $f_i(x)$, $i \leq I$, be $\mathfrak{R}$-r, with the first K of the functions being affine, and let $F(y) : \mathbf{R}^I \to \mathbf{R} \cup \{+\infty\}$ be $\mathfrak{R}$-r and monotonically nondecreasing in $y_{K+1}, ..., y_I$.
$$y, y' \in \mathbf{R}^I, y \geq y', y_i = y'_i, i \leq K \implies F(y) \geq F(y').$$
Then the functions
$$g(x) = \begin{cases} F(f_1(x), ..., f_I(x)) & , f_i(x) < \infty\, \forall i \\ +\infty & , \text{otherwise.} \end{cases}$$

is $\mathfrak{K}$-r, specifically,

$$\left\{\begin{array}{c} f_i \text{ are affine}, i\le K,\ \&\ t\ge f_i(x) \iff \exists u^i: P_ix+tp_i+Q_iu^i-r_i\in\mathbf{K}_i,\ K<i\le I \\ t\ge F(y) \iff \exists u: Py+tp+Qu-r\in\mathbf{K}\end{array}\right\}$$

$$\Downarrow$$

$$t\ge g(x) \iff \exists t_i, 1\le i\le I, u^i, K<i\le I, u: \left\{\begin{array}{ll} \underbrace{t_i-f_i(x)=0}_{\text{linear equations}} & ,i\le K \\ P_ix+t_ip_i+Q_iu^i-r_i\in\mathbf{K}_i & ,K<i\le I \\ P[t_1;...;t_k]+tp+Qu-r\in\mathbf{K} & \end{array}\right.$$

Solution: This is straightforward – substitute "$\mathfrak{K}$-representation" with the definition of this notion.

24.6.3 $\mathfrak{R}/\mathfrak{L}/\mathfrak{S}$ hierarchy

Exercise IV.34.

1. Let $\mathfrak{K}$ and $\mathfrak{M}$ be two families of regular cones, each containing nonnegative rays and closed w.r.t. taking finite direct products and passing from a cone to its dual cone. Assume that every cone $\mathbf{M}\in\mathfrak{M}$ admits $\mathfrak{K}$-representation:
$$\mathbf{M}=\{y:\exists v: P_{\mathbf{M}}y+Q_{\mathbf{M}}v-r_{\mathbf{M}}\in\underbrace{\mathbf{K}_{\mathbf{M}}}_{\in\mathfrak{K}}\}.$$
Show that a $\mathfrak{M}$-r. $X=\{x\in\mathbf{R}^n:\exists u: Px+Qu-r\in\underbrace{\mathbf{M}}_{\in\mathfrak{M}}\}$ of a set X can be straightforwardly converted into $\mathfrak{K}$-r. of X.
2. [Cf. Exercise IV.35] Note that $\mathbf{R}^n_+$ belongs to $\mathfrak{L}$ (same as to every other family of cones we are considering here – all these families contain nonnegative rays and are closed w.r.t. taking finite direct products), thus, every polyhedral representable set/function is Lorentz-representable as well by item 1. Check that the Lorentz cone $\mathbf{L}^m$ is semidefinite-representable as well, specifically,
$$\begin{array}{rcl}\mathbf{L}^m & := & \{x\in\mathbf{R}^m: x_m\ge\sqrt{\sum_{i=1}^{m-1}x_i^2}\} \\ & = & \{x\in\mathbf{R}^m:\mathrm{Arrow}(x):=\left[\begin{array}{c|c|c|c} x_m & x_1 & \cdots & x_{m-1} \\ \hline x_1 & x_m & & \\ \hline \vdots & & \ddots & \\ \hline x_{m-1} & & & x_m\end{array}\right]\succeq 0\}\end{array}$$
implying by item 1 that cones from $\mathfrak{L}$ admit explicit $\mathfrak{S}$-representations and thus that Lorentz-representable sets and functions are semidefinite representable as well, with $\mathfrak{S}$-r.'s readily given by $\mathfrak{L}$-r.'s.

Solution: 1: When $X=\{x\in\mathbf{R}^n:\exists u: Px+Qu-r\in\mathbf{M}\}$ and $\mathbf{M}=\{y:\exists v: P_{\mathbf{M}}y+Q_{\mathbf{M}}v-r_{\mathbf{M}}\in\mathbf{K}_{\mathbf{M}}\in\mathfrak{K}\}$, we clearly have

$$\begin{array}{rcl}X & = & \{x:\exists u: Px+Qu-r\in\mathbf{M}\}=\{x:\exists u,y: y=Px+Qu-r, y\in\mathbf{M}\} \\ & = & \{x:\exists u,y,v: y=Px+Qy-r, P_{\mathbf{M}}y+Q_{\mathbf{M}}v-r_{\mathbf{M}}\in\underbrace{\mathbf{K}_{\mathbf{M}}}_{\in\mathfrak{K}}\},\end{array}$$

and we end up with $\mathfrak{K}$-representation of X. ■

2: See solution to Exercise IV.35.1.

■

Exercise IV.35. It is easy "to see" the nonnegative orthant $\mathbf{R}^n_+$ in the semidefinite cone $\mathbf{S}^n_+$ – $\mathbf{R}^n_+$ is nothing but the intersection of $\mathbf{S}^n_+$ with the linear subspace L of diagonal matrices from $\mathbf{S}^n$. Formally: Let A be the embedding of $\mathbf{R}^n$ into $\mathbf{S}^n$ which maps vector a into diagonal matrix $\mathrm{Diag}\{a\}$; then $z\in\mathbf{R}^n_+$ if and only if $Az\in\mathbf{S}^n_+$. Alternatively, you can get $\mathbf{R}^n_+$ as the linear image of the positive semidefinite cone, namely, its image under the linear mapping which maps a symmetric $n\times n$ matrix Z into the vector $\mathrm{Dg}\{Z\}$ composed of diagonal entries of Z. As a result, a Linear Programming problem $\min_{x\in\mathbf{R}^n}\{c^Tz: Ax\le b\}$ can be converted into equivalent semidefinite

problem $\min_{X\in\mathbf{S}^n}\{\sum_i c_i X_{ii} : X \succeq 0, A\mathrm{Dg}\{X\} \le b\}$. As it happens, similar possibilities exist for the Lorentz cone $\mathbf{L}^n$, including possibility to reformulate a conic problem involving direct products of Lorentz cones as a semidefinite program. Specifically,

1. Prove that $x \in \mathbf{L}^n$ if and only if the "arrow" matrix

$$\mathrm{Arrow}(x) = \left[\begin{array}{c|c|c|c|c} x_n & x_1 & x_2 & \dots & x_{n-1} \\ \hline x_2 & x_n & & & \\ \hline \vdots & & \ddots & & \\ \hline x_{n-1} & & & & x_n \end{array}\right]$$

   is positive semidefinite.
2. Represent $\mathbf{L}^n$ as the image of $\mathbf{S}^n_+$ under a linear mapping.

Solution: 1: The case of $n = 1$ is trivial. Now let $n \ge 2$. In one direction: Assume that $x \in \mathbf{L}^n$, and let us verify that $\mathrm{Arrow}(x) \in \mathbf{S}^n_+$. Indeed, from $x \in \mathbf{L}^n$ it follows that $x_n \ge 0$. If $x_n = 0$, then $x = 0$ due to $\sum_{i=1}^{n-1} x_i^2 \le x_n^2$, and therefore $\mathrm{Arrow}(x) = 0_{n\times n} \succeq 0$. If $x_n > 0$, then $x_n - \sum_{i=1}^{n-1} x_i^2/x_n \ge 0$, or, which is the same, $x_n - [x_1;\dots;x_{n-1}]^\top [x_n I_{n-1}]^{-1}[x_1;\dots;x_{n-1}] \ge 0$, and $\mathrm{Arrow}(x) \succeq 0$ by the Schur Complement Lemma applied with the 1×1 North-Western block. In the opposite direction: Assume that $\mathrm{Arrow}(x) \succeq 0$, and let us prove that $x \in \mathbf{L}^n$. Indeed, x_n is diagonal element in positive semidefinite matrix and as such is nonnegative. If $x_n = 0$, then the diagonal of positive semidefinite matrix $\mathrm{Arrow}(x)$ is zero, whence the matrix itself is zero[15], so that $x = 0 \in \mathbf{L}^n$. And if $x_n > 0$, then $\sum_{i=1}^{n-1} x_i^2/x_n \le x_n$ by the Schur Complement Lemma, the bottom line being that $x_n \ge \sqrt{\sum_{i=1}^{n-1} x_i^2}$, that is, $x \in \mathbf{L}^n$. ■

2: The required linear mapping is the mapping $X \mapsto A^*(X)$ conjugate to the mapping $x \mapsto \mathrm{Arrow}(x)$, that is, the mapping $A(X)$ given by the identity

$$[A^*(X)]^\top x = \mathrm{Tr}(X\mathrm{Arrow}(x))\ \forall x \in \mathbf{R}^n, X \in \mathbf{S}^n$$

or, which is the same,

$$A^*(X) = [2X_{12}; 2X_{13}; \dots; 2X_{1n}; \mathrm{Tr}(X)]$$

Indeed, let L be the linear subspace in $\mathbf{S}^n$ composed of arrow matrices, that is, the image space of the linear mapping $x \mapsto \mathrm{Arrow}(x) : \mathbf{R}^n \to \mathbf{S}^n$. Since L intersects $\mathrm{int}\,\mathbf{S}^n_+$, Dubovitski-Milutin Lemma says that restricting onto L nonnegative on $\mathbf{S}^n_+$ linear forms, we get exactly the set of all linear forms on L which are nonnegative on $\mathbf{S}^n_+ \cap L$ (see Exercise II.49.2). Taking into account that $x \mapsto \mathrm{Arrow}(x)$ is one-to-one linear mapping of $\mathbf{R}^n$ onto L, we conclude that *linear form $g^\top x$ is nonnegative whenever $x \in \mathbf{L}^n$ if and only if it is of the form* $\mathrm{Tr}(X\mathrm{Arrow}(x))$ *for some $X \in \mathbf{S}^n$ such that* $\mathrm{Tr}(XY) \ge 0$ *for all $Y \in \mathbf{S}^n_+$*. Since both $\mathbf{S}^n_+$ and $\mathbf{L}^n_+$ are self-dual, the latter observation can be reformulated as "$g \in \mathbf{L}^n$ if and only if there exists $X \in \mathbf{S}^n_+$ such that $g^\top x = \mathrm{Tr}(X\mathrm{Arrow}(x))$ identically in $x \in \mathbf{R}^n$," or, which is the same, if and only if $g = A^*(X)$ for some $X \in \mathbf{S}^n_+$. ■

24.6.4 More calculus

The calculus rules to follow are less trivial:

Exercise IV.36 [passing from a set to its support function and polar] Let $X \subset \mathbf{R}^n$ be a nonempty closed convex set given by essentially strictly feasible $\mathfrak{R}$-representation:

$$\begin{array}{c} X = \{x \in \mathbf{R}^n : \exists u : Ax + Bu - c \ge 0, Px + Qu - r \in \mathbf{K}\} \\ \& \ \exists \bar{x}, \bar{u} : A\bar{x} + B\bar{u} - c \ge 0, P\bar{x} + Q\bar{u} - r \in \mathrm{int}\,\mathbf{K}. \end{array} \qquad (*)$$

[15] due to immediate observation: *if a diagonal entry A_{ii} of $A \succeq 0$ vanishes, then all entries in i-th row and i-th column of A vanish as well*, due to the inequality $A_{ij}^2 \le A_{ii}A_{jj}$, see Remark D.28

This representation induces $\mathfrak{R}$-r. of the support function $\phi_X(y) = \sup_{x\in X} y^\top x$, specifically,

$$t \geq \phi_X(y) \iff \exists(\lambda,\xi): \begin{array}{l} A^\top\lambda + P^*\xi + y = 0, B^\top\lambda + Q^*\xi = 0 \\ c^\top\lambda + \langle r,\xi\rangle + t \geq 0, \lambda \geq 0, \xi \in \mathbf{K}_* \end{array}.$$

where $\langle\cdot,\cdot\rangle$ is the inner product in the Euclidean space where $\mathbf{K}$ lives and, as always, $\mathbf{K}_*$ is the cone dual to $\mathbf{K}$. In addition, $(*)$ induces $\mathfrak{R}$-r. of the polar $\mathrm{Polar}\,(X)$ of X:

$$\begin{array}{rcl} \mathrm{Polar}\,(X) & := & \{y : y^\top x \leq 1\, \forall x \in X\} \\ & = & \left\{y : \exists(\lambda,\xi): \begin{array}{l} A^\top\lambda + P^*\xi + y = 0, B^\top\lambda + Q^*\xi = 0 \\ c^\top\lambda + \langle r,\xi\rangle + 1 \geq 0, \lambda \geq 0, \xi \in \mathbf{K}_* \end{array}\right\} \end{array}$$

Solution: By definition, $t \geq \phi_X(y)$ if and only if the optimization problem

$$\max_{x\in X} y^\top x$$

is bounded with optimal value $\leq t$, or, which is the same under the circumstances, the conic problem

$$\max_{x,u} \left\{y^\top x : Ax + Bu - c \geq 0, Px + Qu - r \in \mathbf{K}\right\} \qquad (\#)$$

is bounded with the optimal value $\leq t$. We are in the case when the latter problem is essentially strictly feasible; applying Conic Duality Theorem, we conclude that $(\#)$ is bounded with optimal value $\leq t$ if and only if the optimization problem

$$\min_{\lambda,\xi} \left\{-c^\top\lambda - \langle r,\xi\rangle : A^\top\lambda + P^*\xi = -y, B^\top\lambda + Q^*\xi = 0, \lambda \geq 0, \xi \in \mathbf{K}_*\right\}$$

has a feasible solution with the value of the objective $\leq t$. Thus,

$$t \geq \phi_X(y) \iff \exists(\lambda,\xi): \begin{array}{l} A^\top\lambda + P^*\xi + y = 0, B^\top\lambda + Q^*\xi = 0 \\ b^\top\lambda + \langle r,\xi\rangle + t \geq 0, \lambda \geq 0, \xi \in \mathbf{K}_* \end{array}.$$

The resulting representation of the epigraph of ϕ_X, by item 1 of Exercise IV.31, straightforwardly induces a $\mathfrak{R}$-r. of ϕ_X.

Now, $\mathrm{Polar}\,(X) = \{Y : \phi_X(y) \leq 1\}$; applying item 2 of Exercise IV.31, we conclude that

$$\mathrm{Polar}\,(X) = \left\{y : \exists(\lambda,\xi): \begin{array}{l} A^\top\lambda + P^*\xi + y = 0, B^\top\lambda + Q^*\xi = 0 \\ c^\top\lambda + \langle r,\xi\rangle + 1 \geq 0, \lambda \geq 0, \xi \in \mathbf{K}_* \end{array}\right\} \qquad \blacksquare$$

Exercise IV.37. Let $f : \mathbf{R}^n \to \mathbf{R} \cup \{+\infty\}$ be a proper convex lower semiconscious function given by essentially strictly feasible $\mathfrak{R}$-representation:

$$\begin{array}{c} t \geq f(x) \iff \exists u : Ax + tq + Bu \geq c, Px + tp + Qu - r \in \mathbf{K} \\ \& \ \exists \overline{x}, \overline{t}, \overline{u} : A\overline{x} + \overline{t}q + B\overline{u} \geq c, P\overline{x} + \overline{t}p + Q\overline{u} - r \in \mathrm{int}\,\mathbf{K} \end{array}$$

Build $\mathfrak{R}$-r. of the Legendre transform

$$f^*(y) = \sup_x \left[y^\top x - f(x)\right]$$

of f.

Solution: We clearly have

$$f^*(y) = \sup_{x,t} \left\{y^\top x - t : t \geq f(x)\right\} = \sup_{x,t,u} \left\{y^\top x - t : Ax + tq + Bu \geq c, Px + tp + Qu - r \in \mathbf{K}\right\},$$

that is, $f^*(y)$ is the optimal value in the conic problem

$$\sup_{x,t,u} \left\{y^\top x - t : Ax + tq + Bu \geq c, Px + tp + Qu - r \in \mathbf{K}\right\} \qquad (P)$$

Under the circumstances, the problem is essentially strictly feasible, implying by the Conic Duality Theorem that $f^*(y) \le \tau$ if and only if the conic dual of (P) – the problem

$$\min_{\lambda,\xi}\left\{-c^T\lambda - \langle r,\xi\rangle : \begin{array}{l} A^T\lambda + P^*\xi + y = 0 \\ q^T\lambda + \langle p,\xi\rangle = 1 \\ B^T\lambda + Q^*\xi = 0 \\ \lambda \ge 0, \xi \in \mathbf{K}_* \end{array}\right\}$$

– has a feasible solution with the value of the objective $\le \tau$, that is,

$$\tau \ge f^*(y) \iff \exists \lambda,\xi : \begin{array}{l} c^T\lambda + \langle r,\xi\rangle + \tau \ge 0 \\ A^T\lambda + P^*\xi + y = 0 \\ q^T\lambda + \langle p,\xi\rangle = 1 \\ B^T\lambda + Q^*\xi = 0 \\ \lambda \ge 0, \xi \in \mathbf{K}_* \end{array}$$

which is a $\mathfrak{K}$-r. of f^*. ■

Raw materials. Rules of grammar become useful only after we have at our disposal words in "dictionary form" which we can combine using these rules. Similarly, calculus of conic representations becomes useful only after a rich enough dictionary of "raw materials," "atoms" – specific $\mathfrak{K}$-representable sets and functions – is built. In contrast to calculus rules which are, basically, independent of what is the family $\mathfrak{K}$ of cones in question, raw materials do depend on $\mathfrak{K}$. Here we restrict ourselves with few instructive examples of Lorentz- and Semidefinite-representable sets and functions; for in-depth acquaintance with this topic, we refer the reader to [BTN].

We understand well what are the "atomic" $\mathfrak{R}$-representable functions and sets – these are half-spaces and affine functions. Other polyhedrally representable sets are intersections of finite families of half-spaces, and other polyhedrally representable functions – maxima of finitely many affine functions restricted on a polyhedral domain. In other words, all $\mathfrak{R}$-representable functions and sets are obtained from the above atoms via the calculus we have just outlined.

In the next two exercises we present instructive examples of $\mathfrak{L}$-r functions and sets.

Exercise IV.38. [$\mathfrak{L}$-representability of $\|\cdot\|_2$ and $\|\cdot\|_2^2$] Check that the functions $\|x\|_2$ and x^Tx on $\mathbf{R}^n$ admits $\mathfrak{L}$-r.'s as follows:

$$\{[x;t] \in \mathbf{R}^n_x \times \mathbf{R}_t : t \ge \|x\|_2\} = \left\{[x;t] \in \mathbf{R}^n \times \mathbf{R} : [x;t] \in \mathbf{L}^{n+1}\right\}$$
$$\{[x;t] \in \mathbf{R}^n_x \times \mathbf{R}_t : t \ge x^Tx\} = \left\{[x;t] \in \mathbf{R}^n \times \mathbf{R} : [2x;t-1;t+1] \in \mathbf{L}^{n+2}\right\}$$

Solution: evident.

Exercise IV.39. [$\mathfrak{L}$-representability of power functions] Justify the following claims

1. Let k be a positive integer. Then the set

$$\mathfrak{G}_k = \left\{[t;x_1;x_2;...;x_{2^k}] \ge 0 : t \le \left[\prod_{i=1}^{2^k} x_i\right]^{1/2^k}\right\}$$

– the intersection of the hypograph of the geometric mean of 2^k nonnegative variables $x_1,...,x_{2^k}$ with the half-space $\{[t;x] \in \mathbf{R}^{2^k}_x \times \mathbf{R}_t : t \ge 0\}$ – admits $\mathfrak{L}$-representation, specifically,

$$\mathfrak{G}_k = \left\{[t;x_1;x_2;...;x_{2^k}] \ge 0 : \exists\{u_{i,\ell} \ge 0, 1 \le \ell \le k, 1 \le i \le 2^\ell\} : \begin{array}{l} u_{ik} = x_i, 1 \le i \le 2^k \\ [2u_{i\ell}; u_{2i-1,\ell+1} - u_{2i,\ell+1}; u_{2i-1,\ell+1} + u_{2i,\ell+1}] \in \mathbf{L}^3, \\ \qquad\qquad 1 \le i \le 2^\ell, 1 \le \ell < k \\ [2t; u_{1,1} - u_{2,1}; u_{1,1} + u_{2,1}] \in \mathbf{L}^3. \end{array}\right\} \quad (*)$$

Solution: For a triple of nonnegative reals u, v, w, relation $[2u; w-v; v+w] \in \mathbf{L}^3$ is equivalent to $u \le \sqrt{vw}$. Thus, the inequalities on $x, t, u_{i,\ell}$ in $(*)$ tell us the following story:

> We split 2^k nonnegative variables x_i, $i \le 2^k$ of "generation 0" into 2^{k-1} consecutive pairs and associate with i-th of these pairs its "child" – nonnegative variable $u_{i,k-1}$ "of generation 1" linked to its parents x_{2i-1}, x_{2i} by the inequality $u_{i,k-1} \le \sqrt{x_{2i-1}x_{2i}}$. Similarly, we split 2^{k-1} variables $u_{i,k-1}$ of generation 1 into 2^{k-2} consecutive pairs and associate with every pair its child, nonnegative variable $u_{i,k-2}$ of generation 2, and link it to its parents by the inequality $u_{i,k-2} \le \sqrt{u_{2i-1,k-1}u_{2i,k-1}}$.
> We proceed in the same fashion until 2 variables, $u_{1,1}$, $u_{2,1}$ of generation $k-1$ are built, and link these two variables to variable t by the inequality $t \le \sqrt{u_{1,1}u_{2,1}}$.

Note that the constraints on all our variables are the linear nonnegativity constraints and the constraints stating that specific linear images of the vector of these variables belong to $\mathbf{L}^3$, that is, the solution set S of the system of constraints specifying all our variables is given by explicit system of linear and $\mathbf{L}^3$-conic inequalities, and this system provides an explicit $\mathfrak{L}$-r. of the projection $\overline{S}$ of S onto the plane of variables t, x_i. On the other hand, it is clear that what our story says about relation between (nonnegative!) variables x_i and t is *exactly* the inequality $t \le \left[\prod_{i=1}^{2^k} x_i\right]^{1/2^k}$, so that $\overline{S}$ is nothing but the set $\mathfrak{G}_k$.

Surprisingly, item 1 paves road to $\mathfrak{L}$-representations of power functions.

2. Build explicit $\mathfrak{L}$-r's of the univariate functions as follows:

2.1. $f(x) = \max[0, x]^\theta$ with rational $\theta = p/q \ge 1$ ($p \ge q$ are positive integers).

2.2. $f(x) = \begin{cases} x^{p_+/q_+} & , x \ge 0 \\ |x|^{p_-/q_-} & , x \le 0 \end{cases}$, where $p_\pm$, $q_\pm$ are positive integers with $p_+/q_+ \ge 1$, $p_-/q_- \ge 1$

2.3. $f(x) = \begin{cases} -x^{p/q} & , x \ge 0 \\ +\infty & , x < 0 \end{cases}$ with positive integers p, q such that $p/q \le 1$

2.4. $f(x) = \begin{cases} x^{-p/q} & , x > 0 \\ +\infty & , x \le 0 \end{cases}$ with positive integers p, q

Solution: 2.1: Given positive integers $p \ge q$, let us select positive integer k such that $p + q \le 2^k$ and consider the affine mapping

$$(y, t) \to [y; \overbrace{y; ...; y}^{2^k-p}; \overbrace{t; ...; t}^{q}; \overbrace{1; ...; 1}^{p-q}] : \mathbf{R}^2 \to \mathbf{R}^{1+2^k}.$$

Our calculus of conic representations allows to convert the $\mathfrak{L}$-r. of $\mathfrak{G}_k$ built in item 1 into an explicit $\mathfrak{L}$-r. for the inverse image of the set $\mathfrak{G}_k$ under the above affine mapping, that is, for the set

$$F = \{[y; t] \in \mathbf{R}^2_+ : t \ge y^{p/q}\}.$$

The epigraph E of f is obtained from F by operations covered by our calculus:

$$E = \{[t; x] : t \ge \max[x; 0]^{p/q}\} = \{[t; x] : \exists y : [t; y] \in F, y \ge x\},$$

so that our calculus allows to convert the $\mathfrak{L}$-r. of F we have already built into $\mathfrak{L}$-r. for E.
2.2: Construction from item 2.1 allows us to build an explicit $\mathfrak{L}$-r. for the function $\max[0, x]^{p_+/q_+}$ and, after evident modification, for the function $\max[0, -x]^{p_-/q_-}$. These $\mathfrak{L}$-r.'s via calculus provide explicit $\mathfrak{L}$-r. for the sum of these two functions, that is, for our now target function f.
2.3: The epigraph of our f is obtained from the one of the function $g(z) = \max[0, z]^{q/p}$ by one-to-one linear transformation, and we can convert the explicit $\mathfrak{L}$-r. of g given in item 2.1 into $\mathfrak{L}$-r. of our current f.
2.4: Given p, q, let us find positive integer k such that $2^k \ge p + q$, and consider the affine mapping

$$[t; x] \mapsto [1; \overbrace{t; ...; t}^{q}; \overbrace{x; ...; x}^{p}; \overbrace{1; ...; 1}^{2^k-p-q}] : \mathbf{R}^2 \to \mathbf{R}^{1+2^k}.$$

The inverse affine image of $\mathfrak{G}_k$ under this mapping is exactly the epigraph of our current f, so that calculus of $\mathfrak{L}$-r.'s provides us with explicit $\mathfrak{L}$-r. of f inherited from the $\mathfrak{L}$-r. of $\mathfrak{G}_k$ built in item 1.

3. Build $\mathfrak{L}$-r's of the following sets:

   3.1. The hypograph
   $$\{[x;t] \in \mathbf{R}^n_+ \times \mathbf{R}_t : t \le f(x) := x_1^{\pi_1} x_2^{\pi_2} ... x_n^{\pi_n}\}$$
   of algebraic monomial of n nonnegative variables, where π_i are positive rationals such that $\sum_i \pi_i \le 1$ (the latter inequality for nonnegative π_i's is a necessary and sufficient for f to be concave on $\mathbf{R}^n_+$).

   3.2. The epigraph of algebraic monomial $f(x) = x_1^{-\pi_1} x_2^{-pi_2} ... x_n^{-\pi_n}$ of n positive variables, where π_i are positive rationals.

   3.3. The epigraph of $\|\cdot\|_\pi$ on $\mathbf{R}^n$ with rational $\pi \ge 1$.

Solution: 3.1: Let $\pi_i = p_i/q$ with positive integers p_i and q and k be positive integer such that $2^k \ge q$. Consider the affine mapping

$$[x;t] \mapsto [t; \overbrace{x_1;...;x_1}^{p_1}; \overbrace{x_2;...;x_2}^{p_2}; ...; \overbrace{x_n;...;x_n}^{p_n}; \overbrace{t;...;t}^{2^k-q}; \overbrace{1;...;1}^{q-p_1-...-p_n}] : \mathbf{R}^{1+n} \to \mathbf{R}^{1+2^k};$$

note that the right hand side makes sense due to $p_1+...+p_n \le q$ in view of $\sum_i \pi_i \le 1$. As is immediately seen, the inverse image of $\mathfrak{G}_k$ under this mapping is the set

$$F = \{[x;t] \ge 0 : t \le f(x)\},$$

and the $\mathfrak{L}$-r. of $\mathfrak{G}_k$ built in item 1 combines with the calculus of $\mathfrak{L}$-representations to yield an explicit $\mathfrak{L}$-r. for F. It remains to note that, similarly to what happens in item 2.1, the hypograph E of f is obtained from F by operations covered by our calculus:

$$E = \{[x;t] : \exists \tau : [x;\tau] \in F \ \&\ t \le \tau\}.$$

3.2: Representing $\pi_i = p_i/q$ with positive integers p_i, q and selecting positive integer k such that $2^k \ge q + \sum_i p_i$, consider the affine mapping

$$[x;t] \mapsto [1; \overbrace{x_1;...;x_1}^{p_1}; \overbrace{x_2;...;x_2}^{p_2}; ...; \overbrace{x_n;...;x_n}^{p_n}; \overbrace{t;...;t}^{q}; \overbrace{1;...;1}^{2^k-q-\sum_i p_i}] : \mathbf{R}^{n+1} \to \mathbf{R}^{1+2^k}.$$

as is immediately seen, the inverse image of $\mathfrak{G}_k$ under this mapping is exactly the epigraph of f, so that a $\mathfrak{L}$-r. for f is readily given by our calculus as applied to the $\mathfrak{L}$-r. of $\mathfrak{G}_k$ built in item 1.

3.3: The case of $\pi = 1$ is trivial. Now let $\pi \in (1,\infty)$. It is immediately seen (check it) that

$$t \ge \|x\|_\pi \iff t \ge 0 \ \&\ \exists u_i, v_i : \pm x_i \le u_i, u_i \le v_i^{1/\pi} t^{1-1/\pi}, \sum_i v_i \le t.$$

The sets $\{(t,u_i,v_i) \ge 0 : u_i \le v_i^{1/\pi} t^{1-1/\pi}\}$ admit explicit $\mathfrak{L}$-r.'s by item 3.1, and these $\mathfrak{L}$-r.'s via our calculus yield an explicit $\mathfrak{L}$-r. for the epigraph of $\|x\|_\pi$

By Exercise IV.34, expressive abilities of semidefinite representations are at least as strong as those of Lorentz representability. In fact, $\mathfrak{S}$-representability is strong enough to bring, "for all practical purposes," the entire Convex Optimization within the grasp of Semidefinite Optimization. In our next exercise we are just touching the tip of the "semidefinite iceberg."

Exercise IV.40.

1. For starters, build $\mathfrak{S}$-r.'s of the maximum eigenvalue of a symmetric matrix and of the spectral norm $\|\cdot\|_{2,2}$ (the maximum singular value) of a rectangular matrix.
   Hint: Note that for a $p \times q$ matrix A, the eigenvalues of the symmetric $(p+q) \times (p+q)$ matrix $\left[\begin{array}{c|c} & A \\ \hline A^T & \end{array}\right]$ are the singular values of A, minus these singular values, and perhaps a number of zeros.

Solution: $\mathfrak{S}$-r. of the maximal eigenvalue $\lambda_{\max}(X)$ of symmetric $m \times m$ matrix X is immediate:

$$t \geq \lambda_{\max}(X) \iff tI_m - X \succeq 0,$$

This observation combines with Hint to yield $\mathfrak{S}$-r. of the spectral norm of $p \times q$ matrix:

$$t \geq \|X\|_{2,2} \iff \left[\begin{array}{c|c} tI_p & X \\ \hline X^\top & tI_q \end{array}\right] \succeq 0.$$

As a matter of fact, the single most valuable $\mathfrak{S}$-representation is the one for the sums $S_k(X)$ of k largest eigenvalues of a symmetric matrix X; convexity of these sums in X was established in chapter 14.

2. Build $\mathfrak{S}$-r. of the sum $S_k(X)$ of $k \leq m$ largest eigenvalues of $m \times m$ symmetric matrix X.
   Hint: Recall the polyhedral representation, built in Exercise I.29, of the "vector analogy" of $S_k(X)$ – the sum $s_k(x)$ of k largest entries in m-dimensional vector x:

$$t \geq s_k(x) \iff \exists z \geq 0, s : x \leq z + s\mathbf{1}, \sum_i z_i + ks \leq t,$$

   where $\mathbf{1}$ is the all-ones vector.

Solution: The matrix analogy of the representation of $s_k(x)$ is

$$\exists Z \succeq 0, s : X \preceq Z + sI_m, \mathrm{Tr}(Z) + ks \leq t,$$

and we arrive at the "educated guess" stating that for symmetric $m \times m$ matrices X it holds

$$t \geq S_k(X) \iff \exists Z \succeq 0, s : X \preceq Z + sI_m, \mathrm{Tr}(Z) + ks \leq t.$$

Let us verify that this educated guess is true.
In one direction: assume that $Z \succeq 0$ and s are such that $X \preceq Z + sI_m$ and $\mathrm{Tr}(Z) + ks \leq t$, and let us prove that $S_k(X) \leq t$. Denoting by $\lambda(U)$ the vector of eigenvalues, taken with their multiplicities and written down in the non-ascending order, of a symmetric matrix U, recall that $U \succeq U'$ implies that $\lambda(U) \geq \lambda(U')$ (by Variational Characterisation of Eigenvalues). Consequently,

$$S_k(X) = s_k(\lambda(X)) \leq s_k(\lambda(Z + sI_m)) = s_k(\lambda(Z) + s\mathbf{1}) = s_k(\lambda(Z)) + sk \leq \mathrm{Tr}(Z) + sk,$$

where the last inequality is due to $Z \succeq 0$. The concluding quantity in the above chain is $\leq t$, that is , $S_k(X) \leq t$, as claimed.
In the opposite direction: let $S_k(X) \leq t$, and let $X = U\,\mathrm{Diag}\{\lambda_1, \lambda_2, ..., \lambda_m\}U^\top$ be the eigenvalue decomposition of X, $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_m$ being the eigenvalues of X. Let us set $s = \lambda_k$ and $Z = U\,\mathrm{Diag}\{\lambda_1 - \lambda_k, \lambda_2 - \lambda_k, ..., \lambda_{k-1} - \lambda_k, 0, ..., 0\}U^\top$, so that $Z \succeq 0$ and $\mathrm{Tr}(Z) = S_k(X) - k\lambda_k = S_k(X) - ks$, that is, $t \geq S_k(X) = \mathrm{Tr}(Z) + ks$. It remains to note that $X \preceq Z + sI_m$ due to

$$\begin{array}{l} U^\top[sI_m + Z - X]U = \lambda_k I_m + \mathrm{Diag}\{\lambda_1 - \lambda_k, ..., \lambda_{k-1} - \lambda_k, 0, ..., 0\} - \mathrm{Diag}\{\lambda_1, \lambda_2, ...\lambda_m) \\ = \mathrm{Diag}\{0, ..., 0, \lambda_k - \lambda_{k+1}, \lambda_k - \lambda_{k+2}, ..., \lambda_k - \lambda_m\} \succeq 0. \end{array}$$

The importance of $\mathfrak{S}$-representability of $S_k(\cdot)$ becomes clear from the following

3. Let $f(x) : \mathbf{R}^m \to \mathbf{R} \cup \{+\infty\}$ be a convex function symmetric w.r.t. permutations of entries in the argument, and let

$$F(X) = f(\lambda(X)) : \mathbf{S}^m \to \mathbf{R} \cup \{+\infty\};$$

   recall that F is convex by Proposition III.14.3. Show that $F(X)$ admits the following representation:

$$t \geq F(x) \iff \exists u \in \mathbf{R}^m : \begin{array}{ll} f(u) \leq t & (a) \\ u_1 \geq u_2 \geq ... \geq u_m & (b) \\ S_k(X) \leq u_1 + ... + u_k,\ 1 \leq k < m & (c_k) \\ \mathrm{Tr}(X) = u_1 + ... + u_m & (c_m) \end{array} \tag{23.3}$$

   Combine this fact with $\mathfrak{S}$-representability of $S_k(\cdot)$ to arrive at the following

> *Corollary* In the situation of item 3, assume that f is not just symmetric, but is $\mathfrak{S}$-representable as well. A $\mathfrak{S}$-r. of f gives rise to explicit $\mathfrak{S}$-r. of $F(X)$.

Corollary underlies $\mathfrak{S}$-representations of numerous highly important functions and sets, e.g., *Shatten norms* of rectangular matrices – p-norms of the vector of matrix's singular values, or the hypograph $t \le \mathrm{Det}^{1/m}(X)$ of the (appropriate power of the) determinant of $X \in \mathbf{S}^m_+$, or the epigraph of the function $\mathrm{Det}^{-1}(X)$ of $X \succ 0$.

Solution: All we need is to justify (23.3). In one direction: when $t \ge F(X)$, setting $u = \lambda(X)$, we satisfy $(a)-(c)$. In the opposite direction: Let u, X satisfy $(a)-(c)$. From $(b), (c)$ it follows that $s_k(\lambda(X)) \le s_k(u)$ for all $k \le m$, with $s_m(\lambda(X)) = s_m(u)$. Invoking Majorization Principle (section 7.4), we conclude that $\lambda(X) = Pu$ for a properly selected doubly stochastic matrix P. The latter relation, by permutational symmetry and convexity of f, implies that $f(\lambda(X)) \le f(u)$ (see Lemma III.14.1), which combines with (a) to imply the desired relation $F(X) \le t$. ■

Exercise IV.41. A rather interesting example of $\mathfrak{S}$-representable sets deals with matrix square and marix square root:

1. [$\succeq$-epigraph of the matrix square] Prove that the function $F(X) = X^\top X : \mathbf{R}^{m\times n} \to \mathbf{S}^n$ is $\succeq$-convex and find a $\mathfrak{S}$-r. of its $\succeq$-epigraph $\{(X,Y) \in \mathbf{R}^{m\times n} \times \mathbf{S}^n : Y \succeq X^\top X\}$.

   Solution: This is immediate: by Schur Compelent Lemma,

   $$\{(X,Y) \in \mathbf{R}^{m\times n} \times \mathbf{S}^n : Y \succeq X^\top X\} = \{(X,Y) : \left[\begin{array}{c|c} Y & X^\top \\ \hline X & I_m \end{array}\right] \succeq 0\}.$$

   In particular,

   $$\{(X,Y) \in \mathbf{S}^n \times \mathbf{S}^n : Y \succeq X^2\} = \{(X,Y) : \left[\begin{array}{c|c} Y & X \\ \hline X & I_n \end{array}\right] \succeq 0\}.$$

2. [$\succeq$-hypograph of the matrix square root] Prove that the set $\{(X,Y) \in \mathbf{S}^n \times \mathbf{S}^n : X \succeq 0, Y \preceq X^{1/2}\}$ is convex and find its $\mathfrak{S}$-r.

   Solution: The function $X^{1/2} : \mathbf{S}^n_+ \to \mathbf{S}^n_+$ is $\succeq$-concave and $\succeq$-monotone (Example IV.20.5), and therefore

   $$\begin{array}{rcl} \{(X,Y) \in \mathbf{S}^n \times \mathbf{S}^n : X \succeq 0, Y \preceq X^{1/2}\} & = & \{(X,Y) : \exists V : 0 \preceq V, V^2 \preceq X, Y \preceq V\} \\ & = & \{(X,Y) : \exists V : \left\{\begin{array}{l} X \succeq 0, V \succeq 0, Y \preceq V \\ \left[\begin{array}{c|c} X & V \\ \hline V & I_n \end{array}\right] \succeq 0 \end{array}\right. \} \end{array}$$

Note: Solutions to items 1–2 provide us with $\mathfrak{S}$-r.'s of the sets $\{(X,Y) \in \mathbf{S}^n \times \mathbf{S}^n : X \succeq 0, 0 \preceq X \preceq Y^{1/2}\}$ and $\{(X,Y) \in \mathbf{S}^n \times \mathbf{S}^n : X \succeq 0, X^2 \preceq Y\}$. These sets are different, and the second is "essentially smaller" than the first one, see Exercise IV.17.

Exercise IV.42. [important example of $\mathfrak{S}$-representation] Consider the situation as follows. Given a *basic set* $\mathcal{B} \subset \mathbf{R}^n$ which is the solution set of a strictly feasible quadratic inequality:

$$\mathcal{B} = \{u \in \mathbf{R}^n : u^\top Q u + 2q^\top u + \kappa \le 0\},$$

we consider *target set*

$$\mathcal{Q} = \{x \in \mathbf{R}^m : x^\top S x + 2s^\top x + \sigma \le 0\} \qquad [S \in \mathbf{S}^m, s \in \mathbf{R}^m, \sigma \in \mathbf{R}]$$

and affine mapping

$$u \mapsto P(x) := Pu + p : \mathbf{R}^n \to \mathbf{R}^m.$$

We are interested in the situation when the image of the basic set under the mapping $P(\cdot)$ is contained

in the target set, and want to describe this situation in terms of the parameters S, s, σ, P, p. Your task is as follows. Let us set

$$\mathcal{M}(S,s,\sigma;P,p;\lambda) = [P,p]^\top S[P,p] + \left[\begin{array}{c|c} -\lambda Q & P^\top s - \lambda q \\ \hline s^\top P - \lambda q^\top & 2s^\top p + \sigma - \lambda\kappa \end{array}\right].$$

Prove that the inclusion $P(\mathcal{B}) \subset \mathcal{Q}$ is equivalent to the existence of $\lambda \geq 0$ such that

$$\mathcal{M}(S,s,\sigma;P,p;\lambda) \preceq 0. \qquad (!)$$

Solution:

1. Observe that $P(\mathcal{B}) \subset \mathcal{Q}$ if and only if the strictly feasible quadratic inequality

   $$u^\top Q u + 2q^\top u + \kappa \leq 0$$

   on variables $u \in \mathbf{R}^n$ implies validity of the quadratic inequality

   $$[Pu+p]^\top S[Pu+p] + 2s^\top[Pu+p] + \sigma \leq 0,$$

   By Inhomogeneous $\mathcal{S}$-Lemma this is the case if and only if there exists $\lambda \geq 0$ such that

   $$\forall (u \in \mathbf{R}^n, t \in \mathbf{R}) : [Pu+tp]^\top S[Pu+tp] + 2ts^\top[Pu+tp] + \sigma t^2 - \lambda[u^\top Q u + 2tq^\top u + \kappa t^2] \leq 0,$$

   and immediate computation shows that the matrix of the left hand side homogeneous quadratic function of $[u;t]$ is exactly $\mathcal{M}(S,s,\sigma;P,p;\lambda)$. ■
2. Here are the results of our experiments with the inscribed ellipsoid method:
   - $n = 5$: # of iterations: $I = 71$, $f(x^I) = 37.36223$, cpu 68 sec
   - $n = 10$: # of iterations: $I = 131$, $f(x^I) = 41.30913$, cpu 177 sec

   Note that the convex optimization problems in question are well-structured: from the results of Exercise IV.39 it follows that the objectives are $\mathfrak{L}$-r, so that the problems can be solved via Conic Quadratic Programming. With this tool (as implemented in `CVX`), solving the instance with $n = 5$ took just 1.28 sec with reported optimal value 37.36220; similar numbers for the instance with $n = 10$ are 1.99 and 41.30908. We see that, on one hand, just exploiting convexity *per se* already allows to solve optimization problems, at lest low-dimensional ones, to high accuracy in reasonable time, and, on the other hand, utilizing problem's structure via the machinery of $\mathfrak{R}/\mathfrak{L}/\mathfrak{S}$ representations reduces dramatically the computational effort.

Exercises from Appendix A

Exercise A.1.

1. Mark in the list below those subsets of $\mathbf{R}^n$ which are linear subspaces. For the ones that are linear subspaces, find their dimensions and point out bases. For the ones that are not linear subspaces provide counterexamples.
   1. $\mathbf{R}^n$

      Solution: linear subspace, dimension is n, basis, e.g., the collection of n standard basic orth.
   2. $\{0\}$

      Solution: linear subspace, dimension is 0, basis is empty.
   3. $\varnothing$

      Solution: not a linear subspace (linear subspace by definition must be nonempty).
   4. $\left\{x \in \mathbf{R}^n : \sum_{i=1}^{n} i x_i = 0\right\}$

      Solution: linear subspace, dimension is $n-1$, basis, e.g., the collection of vectors

      $$f_i := [\underbrace{0; \ldots; 0}_{i-1}; i+1; -i; 0; \ldots; 0], \quad \text{for } 1 \le i \le n-1.$$
   5. $\left\{x \in \mathbf{R}^n : \sum_{i=1}^{n} i x_i^2 = 0\right\}$

      Solution: linear subspace, dimension is 0, basis is empty.
   6. $\left\{x \in \mathbf{R}^n : \sum_{i=1}^{n} i x_i = 1\right\}$

      Solution: not a linear subspace (e.g., does not contain the origin).
   7. $\left\{x \in \mathbf{R}^n : \sum_{i=1}^{n} i x_i^2 = 1\right\}$

      Solution: not a linear subspace (e.g., contains the first basic orth, but does not contain twice this orth).
2. Suppose that we know L is a subspace of $\mathbf{R}^n$ with exactly one basis. What is L?

   Solution: $L = \{0\}$, basis is empty.

Exercise A.2. Consider the sets given in Exercise A.1 and identify those that are affine subspaces. For those that are affine subspaces, find their affine dimensions and point out their linear subspaces that are parallel to them. For those that are not affine subspaces, provide counterexamples.

Solution: All of the sets that are marked as linear subspaces are also affine subspaces. Their affine dimension is equal to their linear dimension, and the corresponding linear subspace parallel to them is just themselves.

Among the ones that are not linear subspaces, we have the following for their affine subspace status:

- $\varnothing$: not an affine subspace since affine subspace by definition needs to be nonempty.

- $\left\{x \in \mathbf{R}^n : \sum_{i=1}^n ix_i = 1\right\}$: affine subspace, affine dimension is $n-1$, and the corresponding linear subspace parallel to it is given by $\left\{x \in \mathbf{R}^n : \sum_{i=1}^n ix_i = 0\right\}$.
- $\left\{x \in \mathbf{R}^n : \sum_{i=1}^n ix_i^2 = 1\right\}$: not an affine subspace, e.g., it contains the points $a_\pm = [\pm 1; 0; \ldots; 0]$, but does not contain their average (which is their affine combination!).

Exercise A.3.

1. Find the orthogonal complement (w.r.t. the standard inner product) of the following subspace of $\mathbf{R}^n$:
$$\left\{x \in \mathbf{R}^n : \sum_{i=1}^n x_i = 0\right\}.$$
Solution: The orthogonal complement in question is $\mathbf{R} \cdot [1; \ldots; 1]$, i.e., the one-dimensional linear subspace spanned by the all-ones vector.
2. Given vectors $a_1, \ldots, a_m \in \mathbf{R}^n$, find the orthogonal complement (w.r.t. the standard inner product) of the linear subspace $\{x \in \mathbf{R}^n : a_i^\top x = 0, \forall i = 1, \ldots, n\}$.
Solution: The orthogonal complement to the linear subspace $\{x \in \mathbf{R}^n : Ax = 0\}$ is spanned by the transposes of rows of A.
3. Find an orthonormal basis (w.r.t. the standard inner product) of the linear subspace $\{x \in \mathbf{R}^n : x_1 = 0\}$ of $\mathbf{R}^n$.
Solution: An orthonormal basis is, e.g., $\{e_2, e_3, \ldots, e_n\}$, where e_i are the standard basic orth in $\mathbf{R}^n$.

Exercise A.4. Suppose $a \in \mathbf{R}^n$ where $a_i > 0$ for all $i = 1, \ldots, n$, and consider the affine subspace
$$M = \left\{x \in \mathbf{R}^n : \sum_{i=1}^n a_i x_i = 1\right\}.$$
Point out the linear subspace parallel to M and find an affine basis in M.

Solution: The parallel linear subspace is $\{x \in \mathbf{R}^n : \sum_{i=1}^n a_i x_i = 0\}$. An example of an affine basis is the collection $\left\{\frac{1}{a_1}e_1, \ldots, \frac{1}{a_n}e_n\right\}$, where e_i is the i-th standard basic orth.

Exercise A.5. Let $\varnothing \neq C \subseteq \mathbf{R}^n$ and $x \in \mathbf{R}^n$ be given.

1. Is it always true that $\mathrm{Aff}(C - \{x\}) = \mathrm{Aff}(C) - \{x\}$?
Solution: This is always true. Let $y \in \mathrm{Aff}(C - \{x\})$. Then, there are λ_i's with $\sum_i \lambda_i = 1$ and $z_i \in C - \{x\}$, such that $y = \sum_i \lambda_i z_i$. Since $z_i \in C - \{x\}$, there are $x_i \in C$ such that $z_i = x_i - x$. Therefore, $y = \sum_i \lambda_i (x_i - x) = \sum_i \lambda_i x_i - x \in \mathrm{Aff}(C) - \{x\}$. Similarly, if $y \in \mathrm{Aff}(C) - \{x\}$, then there are λ_i's with $\sum_i \lambda_i = 1$ and $x_i \in C$, such that $y = \sum_i \lambda_i x_i - x = \sum_i \lambda_i (x_i - x) \in \mathrm{Aff}(C - \{x\})$. Therefore, $\mathrm{Aff}(C - \{x\}) = \mathrm{Aff}(C) - \{x\}$.
2. Is it always true that $\mathrm{Lin}(C - \{x\}) = \mathrm{Aff}(C) - \{x\}$?
Solution: The equality $\mathrm{Lin}(C - \{x\}) = \mathrm{Aff}(C) - \{x\}$ is not always true, because if $x \notin \mathrm{Aff}(C)$, the set $\mathrm{Aff}(C) - \{x\}$ does not contain the zero vector, but the set $\mathrm{Lin}(C - \{x\})$ always contains the zero vector.
3. Do your answers to the previous questions change if you further assume $x \in \mathrm{Aff}(C)$?
Solution: The answer to the first question does not depend on whether or not $x \in \mathrm{Aff}(C)$ holds. On the other hand, when $x \in \mathrm{Aff}(C)$, the answer to the second question changes and the relation $\mathrm{Lin}(C - \{x\}) = \mathrm{Aff}(C) - \{x\}$ always holds. This is because $\mathrm{Aff}(C) - \{x\}$ is an affine subspace that contains the zero vector, therefore it is a linear subspace. Since it also contains all the elements of $C - \{x\}$, it holds that $\mathrm{Lin}(C - \{x\}) \subseteq \mathrm{Aff}(C) - \{x\}$. For the other direction, we can use the equality

$\mathrm{Aff}(C-\{x\}) = \mathrm{Aff}(C)-\{x\}$ we have shown before. Therefore, we have $\mathrm{Aff}(C)-\{x\} = \mathrm{Aff}(C-\{x\}) \subseteq \mathrm{Lin}(C-\{x\})$. Thus, $\mathrm{Lin}(C-\{x\}) = \mathrm{Aff}(C)-\{x\}$ when $x \in \mathrm{Aff}(C)$.

Exercise A.6. Suppose that we are given n sets $E_1, E_2, \ldots, E_n$ in $\mathbf{R}^{100}$ that are distinct from each other and they satisfy

$$E_1 \subset E_2 \subset \ldots \subset E_n.$$

How large can n be, if

1. every one of E_i is a linear subspace?
2. every one of E_i is an affine subspace?
3. every one of E_i is a convex set?

Solution: The answers are: 101 in items 1 and 2 (dimensions of E_i should grow with i and be integers from the range 0 – 100); in item 3, n can be arbitrary large (take $E_i = \{x \in \mathbf{R}^{100} : \|x\|_2 \leq i\}$).

Exercise A.7. Prove that the *Triangle inequality in the Euclidean norm*, i.e., $\|x+y\|_2 \leq \|x\|_2+\|y\|_2$, holds true as an *equality* if and only if x and y are nonnegative multiples of some vector (which always can be taken to be $x+y$).

Solution: Observe, first, that x, y are nonnegative multiples of some vector iff they are nonnegative multiples of $x+y$. Next, the Triangle inequality in $\|\cdot\|_2$ holds true as equality if and only if $x^\top x+2x^\top y+y^\top y = \|x+y\|_2^2 = (\|x\|_2+\|y\|_2)^2 = \|x\|_2^2+2\|x\|_2\|y\|_2+\|y\|_2^2$, or, which is the same, $x^\top y = \|x\|_2\|y\|_2$. The latter relation clearly holds true when x, y are nonnegative multiples of some vector. Now let $x^\top y = \|x\|_2\|y\|_2$, and let us prove that x and y are nonnegative multiples of some vector. There is nothing to prove when either x, or y, or both, are zero. Now assume that $x \neq 0$, $y \neq 0$. Setting $f = x/\|x\|_2$, $g = y/\|y\|_2$, we arrive at the situation when $\|f\|_2 = \|g\|_2 = 1$, and $x^\top y = \|x\|_2\|y\|_2$ translates to $f^\top g = 1$. Consequently, $\|f-g\|_2^2 = \|f\|_2^2+\|g\|_2^2-2f^\top g = 0$, that is, $f = g$, so that x and y are positive multiples of $f = g$. ■

Exercises from Appendix B

Exercise B.1. Mark in the list below those sets which are closed and those which are open (the sets are in $\mathbf{R}^n$, $\|\cdot\|$ is a norm on $\mathbf{R}^n$, $n > 0$):

1. All vectors with integer coordinates.

   Solution: closed

2. All vectors with rational coordinates.

   Solution: neither closed, nor open

3. All vectors with positive coordinates.

   Solution: open

4. All vectors with nonnegative coordinates.

   Solution: closed

5. $\{x \in \mathbf{R}^n : \|x\| < 1\}$.

   Solution: open

6. $\{x \in \mathbf{R}^n : \|x\| = 1\}$.

   Solution: closed

7. $\{x \in \mathbf{R}^n : \|x\| \leq 1\}$.

   Solution: closed

8. $\{x \in \mathbf{R}^n : \|x\| \geq 1\}$.

   Solution: closed

9. $\{x \in \mathbf{R}^n : \|x\| > 1\}$.

   Solution: open

10. $\{x \in \mathbf{R}^n : 1 < \|x\| \leq 2\}$.

    Solution: neither closed, nor open

Exercise B.2. Consider the function $f(x_1, x_2) : \mathbf{R}^2 \to \mathbf{R}$ defined as

$$f(x_1, x_2) = \begin{cases} \frac{x_1^2 - x_2^2}{x_1^2 + x_2^2}, & \text{if } (x_1, x_2) \neq 0, \\ 0, & \text{if } x_1 = x_2 = 0. \end{cases}$$

Check whether this function is continuous on the following sets:

1. $\mathbf{R}^2$

   Solution: f is not continuous on the set

2. $\mathbf{R}^2 \setminus \{0\}$

   Solution: f is continuous on the set

3. $\{x \in \mathbf{R}^2 : x_1 = 0\}$

   Solution: f is not continuous on the set (note that in this domain we have $f(x) = -1$ whenever $x_2 \neq 0$ and $f(x) = 0$ whenever $x_2 = 0$)

4. $\{x \in \mathbf{R}^2 : x_2 = 0\}$

   Solution: f is not continuous on the set (note that in this domain we have $f(x) = 1$ whenever $x_1 \neq 0$ and $f(x) = 0$ whenever $x_1 = 0$)

5. $\{x \in \mathbf{R}^2 : x_1 + x_2 = 0\}$

   Solution: f is continuous on the set

6. $\{x \in \mathbf{R}^2 : x_1 - x_2 = 0\}$

   Solution: f is continuous on the set

7. $\{x \in \mathbf{R}^2 : |x_1 - x_2| \leq x_1^4 + x_2^4\}$

   Solution: f is continuous on the set.

Exercise B.3. Let $f : \mathbf{R}^n \to \mathbf{R}^m$ be a continuous mapping. Among the following statements, mark those which are always true:

1. If U is an open set in $\mathbf{R}^m$, then so is the set $f^{-1}(U) := \{x \in \mathbf{R}^n : \ f(x) \in U\}$.

   Solution: true

2. If U is an open set in $\mathbf{R}^n$, then so is the set $f(U) = \{f(x) : \ x \in U\}$.

   Solution: not always true (take $f \equiv 0$)

3. If F is a closed set in $\mathbf{R}^m$, then so is the set $f^{-1}(F) = \{x \in \mathbf{R}^n : \ f(x) \in F\}$.

   Solution: true

4. If F is a closed set in $\mathbf{R}^n$, then so is the set $f(F) = \{f(x) : \ x \in F\}$.

   Solution: not always true (take $f(x) = \exp\{x\} : \mathbf{R} \to \mathbf{R}$ and look at $f(\mathbf{R})$).

Exercise B.4. Prove that in general *none* of Theorems B.25, B.29, and B.31 remains valid when

1. X is closed, but not bounded;

   Solution: Take the mapping $x \mapsto \exp\{x\} : X := \mathbf{R} \to \mathbf{R}$, so that X is closed and f is continuous on X. Here:

   - f is unbounded on X, and $f(X)$ is not closed, in contrast to the conclusion of Theorem B.25
   - f is not uniformly continuous on X, in contrast to the conclusion of Theorem B.29
   - f does not achieve its minimum on X, in contrast to the conclusion of Theorem B.31

2. X is bounded, but not closed.

   Solution: Take the mapping $x \mapsto \frac{1}{x} : X := (0, 1) \to \mathbf{R}$, so that X is bounded and f is continuous on X. Here:

   - f is unbounded on X, and $f(X)$ is not closed, in contrast to the conclusion of Theorem B.25
   - f is not uniformly continuous on X, in contrast to the conclusion of Theorem B.29
   - f does not achieve its minimum on X, in contrast to the conclusion of Theorem B.31

Exercises from Appendix D

Exercise D.1.

1. Find the dimension of $\mathbf{R}^{m\times n}$ and point out a basis in this space.

   Solution: The dimension is mn, and a basis is, e.g., the basis $\left\{e_i f_j^\top : i \le m,\, j \le n\right\}$, where $e_1, ..., e_m$ and $f_1, ..., f_n$ are the standard basic orths in $\mathbf{R}^m$, resp., $\mathbf{R}^n$

2. Build an orthonormal basis in $\mathbf{S}^m$.

   Solution: An orthonormal basis in $\mathbf{S}^m$ is composed of m matrices $e_i e_i^\top$ and $\frac{m(m-1)}{2}$ matrices $\frac{1}{\sqrt{2}}[e_i e_j^\top + e_j e_i^\top]$, $1 \le i < j \le m$, where e_i are the standard basic orths in $\mathbf{R}^m$.

Exercise D.2. In the space $\mathbf{R}^{m\times m}$ of square $m \times m$ matrices, there are two interesting subsets: the set $\mathbf{S}^m$ of *symmetric* matrices $\left\{A : A = A^\top\right\}$ and the set $\mathbf{J}^m$ of *skew-symmetric* matrices $\{A = [A_{ij}] : A_{ij} = -A_{ji},\ \forall i, j\}$.

1. Verify that both $\mathbf{S}^m$ and $\mathbf{J}^m$ are linear subspaces of $\mathbf{R}^{m\times m}$.

   Solution: This is evident.

2. Find the dimension of $\mathbf{S}^m$ and point out a basis in $\mathbf{S}^m$.

   Solution: The dimension of $\mathbf{S}^m$ is $\frac{m(m+1)}{2}$, the basis for $\mathbf{S}^m$ was built in Exercise D.1.2.

3. Find the dimension of $\mathbf{J}^m$ and point out a basis in $\mathbf{J}^m$.

   Solution: The dimension of $\mathbf{J}^m$ is $\frac{m(m-1)}{2}$ (note that all of the diagonal entries of the matrices in $\mathbf{J}^m$ must be zero), an orthonormal basis for $\mathbf{J}^m$ is, e.g., $\frac{1}{\sqrt{2}}[e_i e_j^\top - e_j e_i^\top]$, $1 \le i < j \le m$.

4. What is the sum of $\mathbf{S}^m$ and $\mathbf{J}^m$? What is the intersection of $\mathbf{S}^m$ and $\mathbf{J}^m$?

   Solution: Their sum is the entire $\mathbf{R}^{m\times m}$, and their intersection is $\{0\}$.

Exercise D.3. Is the "3-factor" extension of Fact D.1 valid, at least in the case of square matrices X, Y, Z of the same size? That is, for square matrices X, Y, Z of the same size, is it always true that $\mathrm{Tr}(XYZ) = \mathrm{Tr}(YXZ)$?

Solution: Beyond the trivial case of 1×1 matrices, this is wrong, as is immediately shown by numerical experimentation.

Exercise D.4. Given $P \in \mathbf{S}^p$, $Q \in \mathbf{R}^{r\times p}$, and $R \in \mathbf{S}^r$, consider the matrices

$$A = \begin{bmatrix} P & Q^\top \\ Q & R \end{bmatrix},\quad B = \begin{bmatrix} P & -Q^\top \\ -Q & R \end{bmatrix},\quad C = \begin{bmatrix} R & Q \\ Q^\top & P \end{bmatrix},\quad D = \begin{bmatrix} R & -Q \\ -Q^\top & P \end{bmatrix}.$$

Prove that $\lambda(A) = \lambda(B) = \lambda(C) = \lambda(D)$. Thus, the matrices A, B, C, D simultaneously are/are not positive semidefinite. As a consequence, the Schur Complement Lemma says that when $R \succ 0$, we have $A \succeq 0$ if and only if $P - Q^\top R^{-1} Q \succeq 0$; since $A \succeq 0$ if and only if $C \succeq 0$, we see that the same lemma says that when $P \succ 0$, we have $A \succeq 0$ if and only if $R - QP^{-1}Q^\top \succeq 0$.

Solution: Indeed, the matrices are rotations of each other:

$$B = UAU^\top, \quad C = VAV^\top, \quad D = WAW^\top$$

where

$$U = \begin{bmatrix} -I_p & \\ & I_r \end{bmatrix}, \quad V = \begin{bmatrix} & I_r \\ I_p & \end{bmatrix}, \quad W = \begin{bmatrix} & -I_r \\ I_p & \end{bmatrix},$$

and clearly the matrices U, V, W are orthogonal.

Exercise D.5. Let $\mathbf{S}^n_{++} := \operatorname{int} \mathbf{S}^n_+ = \{X \in \mathbf{S}^n : X \succ 0\}$, and consider $X, Y \in \mathbf{S}^n_{++}$. Then, $X \preceq Y$ holds if and only if $X^{-1} \succeq Y^{-1}$ (the $\succeq$-antimonotonicity of X^{-1}, $X \in \mathbf{S}^n_{++}$). Is it true that from $0 \prec X \preceq Y$ it always follows that $X^{-2} \succeq Y^{-2}$?

Solution: For $Z \succ 0$, we clearly have $Z \preceq I_n$ if and only if $Z^{-1} \succeq I_n$, and therefore for $X \succ 0$, $Y \succ 0$ we have

$$X \preceq Y \iff Y^{-1/2} X Y^{-1/2} \preceq I_n \iff Y^{1/2} X^{-1} Y^{1/2} \succeq I_n \iff X^{-1} \succeq Y^{-1}.$$

Numerical experimentation shows that $0 \prec X \preceq Y$ not always implies that $X^{-2} \succeq Y^{-2}$.

Exercise D.6. Let $A, B \in \mathbf{S}^n$ be such that $0 \preceq A \preceq B$. For each one of the following, either prove the statement or produce a counter example:

1. $A^2 \preceq B^2$;

   Solution: We can verify (with Mathematica) that for $n = 2$, taking

   $$A = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}, \quad B = \begin{bmatrix} 2 & 0.5 \\ 0.5 & 1 \end{bmatrix}$$

   gives a counterexample to the claim.

2. $0 \preceq A^{1/2} \preceq B^{1/2}$.

   Solution: This is always true.
   Note that

   $$B - A = \frac{1}{2}\left[(B^{1/2} + A^{1/2})(B^{1/2} - A^{1/2}) + (B^{1/2} - A^{1/2})(B^{1/2} + A^{1/2})\right].$$

   Hence, $B - A \in \mathbf{S}^n_+$ implies $(B^{1/2} + A^{1/2})(B^{1/2} - A^{1/2}) + (B^{1/2} - A^{1/2})(B^{1/2} + A^{1/2}) \in \mathbf{S}^n_+$. Because $B^{1/2} - A^{1/2}$ is a symmetric matrix, we can rewrite it in terms of its eigenvector decomposition as

   $$B^{1/2} - A^{1/2} = UDU^\top,$$

   where U is an orthogonal matrix and D is a diagonal matrix. Then, by defining $X := 2U^\top (B - A)U$ and $Y := U^\top (B^{1/2} + A^{1/2})U$, we observe that

   $$X = YD + DY \tag{$*$}$$

   holds. Because $B - A \in \mathbf{S}^n_+$, we have $X \in \mathbf{S}^n_+$ (see Fact D.31). Likewise $Y \in \mathbf{S}^n_+$ because $B^{1/2} + A^{1/2} \in \mathbf{S}^n_+$ (since both A and B are positive semidefinite). In addition, observe that

   $$\begin{array}{rcl} A' & := & Y - D = 2U^\top A^{1/2} U \\ B' & := & Y + D = 2U^\top B^{1/2} U. \end{array} \tag{$**$}$$

   Therefore, both A' and B' are in $\mathbf{S}^n_+$. Finally, let us consider the diagonal elements of the matrices X, Y, A' and B'. From $(*)$, $(**)$ we see that

   $$\begin{aligned} X_{ii} &= 2Y_{ii} D_{ii} \\ A'_{ii} &= Y_{ii} - D_{ii} \\ B'_{ii} &= Y_{ii} + D_{ii} \end{aligned}$$

Because all of X, Y, A' and B' are in $\mathbf{S}^n_+$, we have all these diagonal elements are nonnegative and $Y_{ii} \geq 0$ for all $i \in [m]$. In particular, we have $Y_{ii} \geq |D_{ii}|$ for all i. Then for any $j \in [m]$ such that $D_{jj} \neq 0$, we have $Y_{jj} > 0$. Moreover, from $X_{jj} = 2Y_{jj}D_{jj}$ and $Y_{jj} > 0$, we deduce that $D_{jj} \geq 0$ as well. This then implies that D has a nonnegative diagonal, and hence $B^{1/2} - A^{1/2} \in \mathbf{S}^n_+$ as desired.

An alternative proof of "$\succeq$-monotonicity" of the square root of a positive semidefinite matrix is given in Example IV.20.5 in section 20.2.

Exercise D.7. A matrix $A \in \mathbf{S}^n$ is called *diagonally dominant* if it satisfies the relation

$$a_{ii} \geq \sum\nolimits_{j \neq i} |a_{ij}|, \quad i = 1, \ldots, n.$$

Prove that every diagonally dominant matrix A is positive semidefinite.

Solution: Let x be an eigenvector of A with eigenvalue λ, and let x_i be the entry of x with the maximum absolute value. As x is an eigenvector, $x \neq 0$ and so $x_i \neq 0$. Replacing, if necessary, x with $-x$, we can assume that $x_i > 0$. Then, as x is an eigenvector of A with eigenvalue λ, we deduce from $Ax = \lambda x$ that

$$a_{ii}x_i + \sum\nolimits_{j \neq i} a_{ij}x_j = \lambda x_i.$$

Moreover, using the fact that $x_i > 0$ is the largest magnitude coordinate in x, we get

$$\sum\nolimits_{j \neq i} a_{ij}x_j \leq \left|\sum\nolimits_{j \neq i} a_{ij}x_j\right| \leq \sum\nolimits_{j \neq i} |a_{ij}x_j| \leq x_i \sum\nolimits_{j \neq i} |a_{ij}| \leq a_{ii}x_i.$$

Combining these two relations, we arrive at

$$\lambda x_i = a_{ii}x_i + \sum\nolimits_{j \neq i} a_{ij}x_j \geq a_{ii}x_i - \left|\sum\nolimits_{j \neq i} a_{ij}x_j\right| \geq 0.$$

This means that $\lambda \geq 0$, and since the eigenvalue λ was arbitrary, all eigenvalues of A are non-negative, and hence $A \succeq 0$.

Exercise D.8. Prove the following matrix analogy of the scalar inequality $ab \leq \frac{a^2+b^2}{2}$ for $a, b \in \mathbf{R}$:

$$AB^\top + BA^\top \preceq AA^\top + BB^\top, \qquad \forall A, B \in \mathbf{R}^{m \times n}.$$

Solution: Note that we can rewrite this expression as

$$AA^\top - AB^\top - BA^\top + BB^\top = (A - B)(A - B)^\top.$$

Then, the positive semidefiniteness of this matrix is immediate.

Exercise D.9.

1. Let I_k denote the $k \times k$ identity matrix, and let A be an $m \times n$ matrix. Prove that the following three properties are equivalent to each other:
   - $A^\top A \preceq I_n$;
   - $AA^\top \preceq I_m$;
   - $\begin{bmatrix} I_m & A \\ A^\top & I_n \end{bmatrix} \succeq 0$.

   Solution: By the Schur Complement Lemma,

   $$X = \begin{bmatrix} I_m & A \\ A^\top & I_n \end{bmatrix} \succeq 0 \iff I_m - AA^\top \succeq 0.$$

   Invoking the concluding comment in Exercise D.4, $X \succeq 0 \iff I_n - A^\top A \succeq 0$.
2. Let $A_1, \ldots, A_k$ be $n \times n$ matrices such that

   $$A_1^\top A_1 + \ldots + A_k^\top A_k \preceq I_n.$$

   For each of the following, either prove the statement or produce a counter example:
   - $A_1 A_1^\top + \ldots + A_k A_k^\top \preceq I_n$;

Solution: When $n = 1$, the claim clearly is true; when $k = 1$, it is true due to item 1 of Exercise. When $k > 1$ and $n > 1$, the claim is wrong in general: set $\kappa = \min[k, n]$, $A_i = ee_i^\top$ with unit e, the first κ basic orths of $\mathbf{R}^n$ in the role of e_i when $i \le \kappa$, and $e_i = 0$ for $\kappa < i \le k$. With this setup, $\sum_i A_i^\top A_i = \sum_{i=1}^{\kappa} e_i e_i^\top \preceq I_n$, while $\sum_i A_i A_i^\top = \kappa e e^\top \not\preceq I_n$.

- $\begin{bmatrix} A_1A_1^\top & A_1A_2^\top & \cdots & A_1A_k^\top \\ A_2A_1^\top & A_2A_2^\top & \cdots & A_2A_k^\top \\ \vdots & \vdots & \ddots & \vdots \\ A_kA_1^\top & A_kA_2^\top & \cdots & A_kA_k^\top \end{bmatrix} \preceq I_{kn}.$

Solution: Observe that by the Schur Complement Lemma and the concluding comment in Exercise D.4 we have

$$\begin{aligned} & I_n - (A_1^\top A_1 + \ldots + A_k^\top A_k) \succeq 0 \\ & \iff \begin{bmatrix} I_n & A_1^\top & \ldots & A_k^\top \\ A_1 & & & \\ \vdots & & I_{kn} & \\ A_k & & & \end{bmatrix} \succeq 0 \\ & \iff I_{kn} - \begin{bmatrix} A_1 \\ \vdots \\ A_k \end{bmatrix} \begin{bmatrix} A_1^\top & \ldots & A_k^\top \end{bmatrix} = I_{kn} - \begin{bmatrix} A_1A_1^\top & A_1A_2^\top & \cdots & A_1A_k^\top \\ A_2A_1^\top & A_2A_2^\top & \cdots & A_2A_k^\top \\ \vdots & \vdots & \ddots & \vdots \\ A_kA_1^\top & A_kA_2^\top & \cdots & A_kA_k^\top \end{bmatrix} \succeq 0, \end{aligned}$$

which is exactly what is required.

References

[Axl15] S. Axler, *Linear algebra done right, 3rd edition*, Undergraduate Texts in Mathematics, Springer, 2015.

[BNO03] D. Bertsekas, A. Nedic, and A. Özdağlar, *Convex analysis and optimization*, Athena Scientific, 2003.

[BTN] A. Ben-Tal and A. Nemirovski, *Lectures on modern convex optimization: analysis, algorithms, and engineering applications*, SIAM 2001 and `https://www.isye.gatech.edu/~nemirovs/LMCOLN.pdf` 2023.

[BV04] S. Boyd and L. Vandenberghe, *Convex optimization*, Cambridge University Press, 2004.

[Edw12] C. H. Edwards, *Advanced calculus of several variables*, Courier Corporation, 2012.

[Gel89] I. M. Gel'fand, *Lectures on linear algebra*, Dover Books on Mathematics, Dover Publications, 1989.

[HUL93] J.-B. Hiriart-Urruty and C. Lemarechal, *Convex analysis and minimization algorithms, I: Fundamentals, II: Advanced theory and bundle methods*, Springer, 1993.

[IT79] A. D. Ioffe and V. M. Tikhomirov, *Theory of extremal problems*, Nauka, 1974 (in Russian). English translation: Studies in Mathematics and its Applications, v. 6, North-Holland, 1979.

[Nem24] A. Nemirovski, *Introduction to linear optimization*, World Scientific, 2024 and `https://www.isye.gatech.edu/~nemirovs/WSbook.pdf`, 2024.

[Nes18] Yu. Nesterov, *Lectures on convex optimization, 2nd edition*, Springer, 2018.

[Pas22] D. S. Passman, *Lectures on linear algebra*, World Scientific, 2022.

[Roc70] R. T. Rockafellar, *Convex analysis*, Princeton University Press, 1970.

[Rud13] W. Rudin, *Principles of mathematical analysis, 3rd edition*, McGraw Hill India, 2013.

[Str06] G. Strang, *Linear algebra and its applications*, Belmont, CA: Thomson, Brooks/Cole, 2006.