Tutorial: *Mirror Descent Algorithms for Large-Scale Deterministic and Stochastic Convex Optimization*

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

COLT 2012
Edinburgh, June 24-27, 2012

♣ **Problem of Primary Interest:** *Convex Minimization*

$$\mathrm{Opt} = \min_{x \in X} f(x) \qquad (P)$$

- X: convex compact subset of Euclidean space E
- $f : X \to \mathbf{R}$: convex Lipschitz continuous

♠ f is represented by a *First Order oracle:*

- given on input $x \in X$, FO returns the value $f(x)$ and a subgradient $f'(x)$ of f at x
- the vector field $x \mapsto f'(x)$ is assumed to be bounded on X

♣ **Mirror Descent for** (P), *milestones:*

- *Subgradient Descent* ("Euclidean prototype"): N. Shor, 1967:

$$X \ni x_\tau \mapsto x_{\tau+1} = \mathrm{Proj}_X(x_\tau - \gamma_\tau f'(x_\tau))$$

- $\gamma_\tau > 0$: stepsizes • $\mathrm{Proj}_X(y) = \mathrm{argmin}_{z \in X} \|y - z\|_2$

- *General Mirror Descent scheme*: Nem., 1979
- *Modern Proximal Point form*: A. Beck & M. Teboulle, 2003

Proximal Setup

$$\mathrm{Opt} = \min_{x \in X} f(x) \qquad (P)$$

- X: convex compact subset of Euclidean space E

♣ **Setup for MD** ("proximal setup") is given by

- a *norm* $\|\cdot\|$ on E
- a *distance-generating function* $\omega(x) : X \to \mathbf{R}$ which should be
  - *convex and continuous on* X
  - *strongly convex, modulus 1, w.r.t.* $\|\cdot\|$:

$$\langle \omega'(x) - \omega'(x'), x - x' \rangle \geq \|x - x'\|^2$$
$$\text{for all } x, x' \in X^o = \{x \in X : \partial\omega(x) \neq \emptyset\}$$

  - admitting a continuous on X^o selection $\omega'(x)$ of subgradients

♠ **Example:** *Euclidean setup:*

$$E = \mathbf{R}^n,\ \|x\| = \|x\|_2,\ \omega(x) = \tfrac{1}{2} x^T x$$

Proximal Setup (continued)

♣ **Proximal setup** $\|\cdot\|, \omega(\cdot)$ for $X \subset E$ induces:

- *ω-center of X* $x_\omega = \mathrm{argmin}_{x \in X} \omega(x)$
- *Bregman distance* $V_x(y) = \omega(y) - \omega(x) - \langle \omega'(x), y - x \rangle$, $x \in X^o, y \in X$. By strong convexity of $\omega(\cdot)$,
$$V_x(y) \geq \tfrac{1}{2}\|y - x\|^2$$
- *ω-radius of X* $\Omega = \Omega[X, \omega(\cdot)] = \sqrt{2[\max_{x \in X} \omega(x) - \min_{x \in X} \omega(x)]}$

  For $x \in X$ one has
$$\tfrac{1}{2}\|x - x_\omega\|^2 \leq V_{x_\omega}(x) \leq \omega(x) - \omega(x_\omega) \leq \tfrac{1}{2}\Omega^2$$
$$\Rightarrow \|x - x_\omega\| \leq \Omega \ \forall x \in X$$
- *prox-mapping*

$[x \in X^o, \xi \in E] \mapsto \mathrm{Prox}_x(\xi) := \mathrm{argmin}_{z \in X} [\langle \xi, z \rangle + V_x(z)] \in X^o$

♠ With Euclidean setup,
$$V_x(y) = \tfrac{1}{2}\|x - y\|_2^2, \ \mathrm{Prox}_x(\xi) = \mathrm{Proj}_X(x - \xi)$$
$\Rightarrow$ Subgradient Descent is the recurrence
$$x_{\tau+1} = \mathrm{Prox}_{x_\tau}(\gamma_\tau f'(x_\tau))$$

Basic Mirror Descent

- X: convex compact subset of Euclidean space E
- $\|\cdot\|, \omega(\cdot)$: proximal setup for (E, X)

♣ **MD** works with a *sequence of vector fields* $\{g_\tau(\cdot) : X \to E\}_\tau$ represented by *an oracle*. At call $\tau = 1, 2, ...$, the query point being x_τ, the oracle returns the vector $g_\tau(x_\tau) \in E$.

- In most of applications, the sequence $\{g_\tau(\cdot)\}_\tau$ is just stationary: $g_\tau(\cdot) \equiv g(\cdot)$.

♠ MD is the recurrence

$$x_1 = x_\omega := \mathrm{argmin}_X \omega(\cdot);\ x_{\tau+1} = \mathrm{Prox}_{x_\tau}(\gamma_\tau g_\tau(x_\tau))$$

- $x_\tau \in X^o$: seach points
- $\gamma_\tau > 0$: stepsizes

Basic Mirror Descent (continued)

$$x_1 = x_\omega := \mathrm{argmin}_X\, \omega;\ x_{\tau+1} = \mathrm{Prox}_{x_\tau}(\gamma_\tau g_\tau(x_\tau))$$

♣ **Main Property of MD:** *Under Boundedness Assumption*

$$\sup_{x\in X,\tau} \|g_\tau(x)\|_* \le L < \infty$$

- $\|\xi\|_* = \max\{\langle \xi, x\rangle : \|x\| \le 1\}$ is the conjugate of $\|\cdot\|$

the residual

$$\varepsilon_t := \max_{z\in X} \sum_{\tau\le t} \lambda_\tau^t \langle g_\tau(x_\tau), x_\tau - z\rangle,\ \ \lambda_\tau^t = \gamma_\tau / \sum_{s\le t}\gamma_s$$

obeys the bound

$$\varepsilon_t \le \frac{\Omega^2 + \sum_{\tau\le t}\gamma_\tau^2 \|g_\tau(x_\tau)\|_*^2}{2\sum_{\tau\le t}\gamma_\tau},\ t = 1, 2, ...$$

- *In particular, when* $\frac{\Omega}{L\sqrt{t}} \le \gamma_\tau \le \frac{\Omega}{\|g_\tau(x_\tau)\|_*\sqrt{t}}$ *for* $1 \le \tau \le t$ *(e.g.,* $\gamma_\tau = \frac{\Omega}{L\sqrt{t}}$, *or* $\gamma_\tau = \frac{\Omega}{\|g_\tau(x_\tau)\|_*\sqrt{t}}$, $1 \le \tau \le t$*), one has*

$$\varepsilon_t \le \Omega L/\sqrt{t}.$$

♠ **Fact:** *When* $g_\tau(\cdot)$ *come from problem "with convex structure," the residual* ε_t *upper-bounds inaccuracy of the approximate solution* $x^t := \sum_{\tau\le t}\lambda_\tau^t x_\tau$ *to the problem.*

Basic Mirror Descent (continued)

Example 1: Convex Minimization $\mathrm{Opt} = \min_X f$**.** Applying MD to $\{g_\tau(\cdot) \equiv f'(\cdot)\}_\tau$ and assuming w.l.o.g. the Lipschitz constant $L_{\|\cdot\|}(f)$ of f taken w.r.t. $\|\cdot\|$ to upper-bound $\|f'(\cdot)\|_*$, one has $f(x^t) - \mathrm{Opt} \leq \mathcal{E}_t$:

$$\begin{aligned}\mathcal{E}_t &= \max_{z\in X} \textstyle\sum_{\tau\leq t} \lambda_\tau^t \langle f'(x_\tau), x_\tau - z\rangle \geq \max_{z\in X} \sum_{\tau\leq t} \lambda_\tau^t [f(x_\tau) - f(z)] \\ &\geq \max_{z\in X} [f(\textstyle\sum_{\tau\leq t} \lambda_\tau^t x_\tau) - f(z)] = f(x^t) - \mathrm{Opt}\end{aligned}$$

$\Rightarrow$ *For every t, t-step MD with appropriate stepsizes ensures*

$$f(x^t) - \mathrm{Opt} \leq \Omega L_{\|\cdot\|}(f)/\sqrt{t}$$

Example 1.A: Convex Online Minimization. *When $g_\tau(x) = f'_\tau(x)$, with convex functions $f_\tau(\cdot) : X \to \mathbf{R}$ satisfying $\|f'_\tau(x)\|_* \leq L < \infty$ for all $x \in X, \tau$, t-step MD with stepsizes $\gamma_\tau = \frac{\Omega}{L\sqrt{t}}$, $1 \leq \tau \leq t$, ensures that*

$$\tfrac{1}{t}\textstyle\sum_{\tau\leq t} f_\tau(x_\tau) \leq \frac{\Omega L}{\sqrt{t}} + \min_{x\in X} \frac{1}{t}\sum_{\tau\leq t} f_\tau(x)$$

Basic Mirror Descent (continued)

Example 2: Convex-Concave Saddle Point problem

$$\mathrm{SadVal} = \min_{u\in U}\max_{v\in V} f(u,v).$$

♣ **Situation:**

- $X = U \times V \subset E_u \times E_v =: E$ with compact convex U, V
- $f(u,v) : X \to \mathbf{R}$: convex in $x \in U$, concave in $v \in V$, Lipschitz continuous

♠ f, U, V give rise to two convex optimization problems:

$$\begin{array}{rcll} \mathrm{Opt}(P) & = & \min_{u\in U}\left[\overline{f}(u) := \max_{v\in V} f(u,v)\right] & (P) \\ \mathrm{Opt}(D) & = & \max_{v\in V}\left[\underline{f}(v) := \min_{u\in U} f(u,v)\right] & (D) \end{array}$$

with equal optimal values: $\mathrm{Opt}(P) = \mathrm{Opt}(D)$, and to *vector field*

$$g(x=[u;v]) = \left[\begin{array}{c} g_u(u,v) \in \partial_u f(u,v) \\ g_v(u,v) \in \partial_v(-f(u,v)) \end{array}\right] : X := U \times V \to E$$

♠ Optimal solutions u_*, v_* to (P), (D) are exactly the *saddle points* of f on $U \times V$:

$$f(u,v_*) \geq f(u_*,v_*) \geq f(u_*,v)\ \forall (u \in U, v \in V):$$

MD for Saddle Point problems

$$\begin{array}{rcll}\mathrm{Opt}(P) & = & \min_{u\in U}\left[\overline{f}(u) := \max_{v\in V} f(u,v)\right] & (P)\\ \mathrm{Opt}(D) & = & \max_{v\in V}\left[\underline{f}(v) := \min_{u\in U} f(u,v)\right] & (D)\end{array}$$
$$\Rightarrow g(u;v) = [f'_u(u,v); -f'_v(u,v)] : U\times V \to E$$

♣ **Fact:** *Applying MD to $g_\tau(\cdot)\equiv g(\cdot)$, the residual*

$$\varepsilon_t = \max_{z\in X}\sum_{\tau\le t}\lambda_\tau^t\langle g(x_\tau), x_\tau - z\rangle,\ \ \lambda_\tau^t = \gamma_\tau/\sum_{s\le t}\gamma_s$$

upper-bounds the saddle point inaccuracy ("duality gap") of the approximate solution $x^t = [u^t; v^t] := \sum_{\tau\le t}\lambda_\tau^t x_\tau$ to (P,D):

$$[\overline{f}(u^t) - \mathrm{Opt}(P)] + [\mathrm{Opt}(D) - \underline{f}(v^t)] = \overline{f}(u^t) - \underline{f}(v_t) \le \varepsilon_t$$

$$\begin{array}{l}\forall [u;v]\in U\times V : \varepsilon_t \ge \sum_{\tau\le t}\lambda_\tau^t\langle g(x_\tau), x_\tau - [u;v]\rangle\\ = \sum_{\tau\le t}\lambda_\tau^t[\langle f'_u(u_\tau,v_\tau), u_\tau - u\rangle + \langle -f'_v(u_\tau,v_\tau), v_\tau - v\rangle]\\ \ge \sum_{\tau\le t}\lambda_\tau^t[f(u_\tau,v_\tau) - f(u,v_\tau) - f(u_\tau,v_\tau) + f(u_\tau,v)]\\ = \sum_{\tau\le t}\lambda_\tau^t[f(u_\tau,v) - f(u,v_\tau)] \ge f(u^t,v) - f(u,v^t)\\ \Rightarrow \varepsilon_t \ge \max_{u\in U, v\in V}[f(u^t,v) - f(u,v^t)] = \overline{f}(u^t) - \underline{f}(v^t).\end{array}$$

MD for Saddle Point problems (continued)

$$\begin{array}{rcll} \mathrm{Opt}(P) & = & \min_{u\in U}\left[\bar{f}(u) := \max_{v\in V} f(u,v)\right] & (P) \\ \mathrm{Opt}(D) & = & \max_{v\in V}\left[\underline{f}(v) := \min_{u\in U} f(u,v)\right] & (D) \end{array}$$
$$\Rightarrow g(u;v) = [f'_u(u,v); -f'_v(u,v)] : U \times V \to E$$

♠ Assuming that $\|\cdot\|$ respects representation $E = E_u \times E_v$: $\|[u;v]\| \equiv \|[u;-v]\|$, we can ensure that $\|g(\cdot)\|_* \le L_{\|\cdot\|}(f)$.

$\Rightarrow$ *t-step MD with properly chosen stepsizes ensures that*

$$[\bar{f}(u^t) - \mathrm{Opt}(P)] + [\mathrm{Opt}(D) - \underline{f}(v^t)] \le \Omega L_{\|\cdot\|}(f)/\sqrt{t}.$$

♠ Similar results for other "problems with convex structure:"

- variational inequalities with monotone operators
- convex Nash equilibrium problems

Reason for Main Property

♣ **Fact:** With $V_x(z) = \omega(z) - \omega(x) - \langle \omega'(x), z - x \rangle$ one has

$$x_+ = \mathrm{Prox}_x(\xi) := \mathrm{argmin}_{z \in X} [\langle \xi, z \rangle + V_x(z)] \quad (1)$$

$$\Rightarrow \forall (z \in X) : \langle \xi, x_+ - z \rangle \leq V_x(z) - V_{x_+}(z) - V_x(x_+) \quad (2)$$

Proof: rearrange terms in the optimality conditions for (1):

$$\langle \xi + \omega'(x_+) - \omega'(x), z - x_+ \rangle \geq 0 \, \forall z \in X$$

♠ **Fact:** (2) implies that

$$\forall (z \in X) : \langle \xi, x - z \rangle \leq V_x(z) - V_{x_+}(z) + \tfrac{1}{2}\|\xi\|_*^2 \quad (3)$$

Proof: by (2),

$$\langle \xi, x - z \rangle \leq V_x(z) - V_{x_+}(z) + [\langle \xi, x - x_+ \rangle - V_x(x_+)],$$

and $\langle \xi, x - x_+ \rangle - V_x(x_+) \leq \|\xi\|_* \|x - x_+\| - \frac{1}{2}\|x - x_+\|^2 \leq \frac{1}{2}\|\xi\|_*^2$.

- By (3), $x_1 = \mathrm{argmin}_X \omega$; $x_{\tau+1} = \mathrm{Prox}_{x_\tau}(\gamma_\tau g_\tau)$ implies

$$\gamma_\tau \langle g_\tau, x_\tau - x \rangle \leq V_{x_\tau}(z) - V_{x_{\tau+1}}(z) + \tfrac{1}{2}\gamma_\tau^2 \|g_\tau\|^2 \; \forall (z \in X, \tau)$$

$$\Rightarrow \quad \textstyle\sum_{\tau \leq t} \gamma_\tau \langle g_\tau, x_\tau - z \rangle \leq \frac{1}{2}\Omega^2 + \frac{1}{2} \sum_{\tau \leq t} \gamma_\tau^2 \|g_\tau\|_*^2 \; \forall z \in X$$

- Dividing by $\sum_{\tau \leq t} \gamma_\tau$ and maximizing in $z \in X$, we get

$$\varepsilon_t := \max_{z \in X} \left[\textstyle\sum_{\tau \leq t} \lambda_\tau^t \langle g_\tau, x_\tau - z \rangle \right] \leq \frac{\Omega^2 + \sum_{\tau \leq t} \gamma_\tau^2 \|g_\tau\|_*^2}{2 \sum_{\tau \leq t} \gamma_\tau}$$

Role of Symmetry

$$\mathcal{E}_t \le \Omega[\sup_{x\in X,\tau} \|g_\tau(x)\|_*]/\sqrt{t} \qquad (*)$$

♣ When X is "nearly symmetric," the MD efficiency estimate can be improved. Assume that

- *X contains $\|\cdot\|$-ball of radius $\theta\Omega$*
- *The vector fields $\{g_\tau(\cdot)\}_\tau$ are uniformly semi-bounded:*

$$M := \sup_{x,x'\in X,\tau} \langle g_\tau(x), x' - x\rangle < \infty$$

Then *for every $t \ge 4/\theta^2$, the t-step MD with the stepsizes*

$$\gamma_\tau = \frac{\Omega}{\|g_\tau(x_\tau)\|_*\sqrt{t}}\ 1 \le \tau \le t$$

ensures that

$$\mathcal{E}_t \le 2\theta^{-1} M/\sqrt{t} \qquad (!)$$

♠ **Note:** *When $\theta = O(1)$,*

- $(!)$ *can only be better than* $(*)$
- When $g_\tau(\cdot) \equiv g(\cdot)$ comes from $\min_{u\in U}\max_{v\in V} f(u,v)$, we have $M \le \max_{U\times V} f - \min_{U\times V} f$ $\Rightarrow$ $(!)$ *becomes*

$$\mathcal{E}_t \le O(1)\left[\max_{U\times V} f - \min_{U\times V} f\right]/\sqrt{t}$$

$O(1/\sqrt{t})$ – good or bad?

♣ The MD convergence rate $O(1/\sqrt{t})$ is slow. However, *this is the best possible rate one can expect when solving nonsmooth large-scale convex problems represented by FO oracles*, or any other oracles providing *local* information.

♠ **Bad news:** *Consider Convex Minimization problem*

$$\mathrm{Opt}(f) = \min_x\{f(x) : \|x\| \le R\} \qquad (P_f)$$

where $\|\cdot\|$ *is either the norm* $\|\cdot\|_p$ *on* $E = \mathbf{R}^n$ *(*$p = 1, 2$*), or the nuclear norm on* $\mathbf{R}^{n\times n}$*. Let*

$$\mathcal{F}_{\|\cdot\|}(L) = \{f : E \to \mathbf{R} : \ f \text{ is convex, } L_{\|\cdot\|}(f) \le L\},$$

and assume that when solving (P_f)*,* $f \in \mathcal{F}_{\|\cdot\|}(L)$ *is learned via calls, one per step, to a FO (or any local) oracle.*

Then for every $t \le n$ *and any* t*-step algorithm* $\mathcal{B}$ *one has*

$$\sup_{f\in\mathcal{F}_{\|\cdot\|}(L)} [f(x_{\mathcal{B}}(f)) - \mathrm{Opt}(f)] \ge 0.01LR/\sqrt{t}$$

- $x_{\mathcal{B}}(f)$: solution generated in t steps by $\mathcal{B}$ as applied to (P_f)

$O(1/\sqrt{t})$ – good or bad? (continued)

$$\mathrm{Opt}(f) = \min_{x \in X} f(x), \ X \subset X_R := \{x \in E : \|x\| \le R\} \quad (P_f)$$

$\|\cdot\|$: $\|\cdot\|_p$ norm on $E = \mathbf{R}^n$ ($p = 1, 2$), or nuclear norm on $\mathbf{R}^{n \times n}$.

♠ **Relatively good news:** With appropriate proximal setup, t-step MD as applied to (P_f) ensures

$$f(x^t) - \mathrm{Opt}(f) \le O\left(L_{\|\cdot\|}(f) R / \sqrt{t}\right)$$

- hidden factor: $O(1)$ when $\|\cdot\| = \|\cdot\|_2$, otherwise $O(1)\sqrt{\ln(n+1)}$

Note:

- *Rate of convergence is (nearly) dimension-independent*
- *When X is simple, computational effort per MD step in the large scale case is by order of magnitudes smaller than in all known polynomial time Convex Optimization techniques, like Interior Point methods*

$\Rightarrow$ *When solving problems with convex structure to low or medium accuracy, MD could be the method of choice...*

Favorable Geometry case

$$\varepsilon_t \le \Omega[X,\omega] \sup_{x\in X,\tau} \|g_\tau(x)\|_*/\sqrt{t}$$

♣ **Question:** *How to choose a good proximal setup?*

- In general, the answer depends on the geometry of X and on a priori information on $\{g_\tau(\cdot)\}_\tau$
- There is, however, a *favorable geometry* case when the answer is clear:
  - Assuming w.l.o.g. that $X^+ = \frac{1}{2}[X - X]$ linearly spans E, X^+ is the unit ball of norm $\|\cdot\|_X$ given solely by X.
  - A *Favorable Geometry case* is the one where X admits a d.-g.f. $\omega_X(\cdot)$ such that $\|\cdot\|_X, \omega_X(\cdot)$ is a valid proximal setup *with "moderate"* $\Omega_X := \Omega[X,\omega_X]$ ($O(1)$, or $O(1)\ln^{O(1)}(\dim X)$).

Favorable Geometry case (continued)

$$\mathcal{E}_t \le \Omega[X,\omega] \sup_{x\in X,\tau} \|g_\tau(x)\|_*/\sqrt{t}$$

♠ **Observation:** Let $\omega_X(\cdot)$ complement $\|\cdot\|_X$ to a proximal setup. Then for every proximal setup $\|\cdot\|, \omega(\cdot)$ for X and every $\{g_\tau(\cdot)\}_\tau$ one has

$$\sup_{x\in X,\tau} \|g_\tau(x)\|_{X,*} \le \Omega[X,\omega] \sup_{x\in X,\tau} \|g_\tau(x)\|_* \qquad (!)$$
$$\Rightarrow \Omega_X \sup_{x\in X,\tau} \|g_\tau(x)\|_{X,*} \le \Omega_X\Omega[X,\omega] \sup_{x\in X,\tau} \|g_\tau(x)\|_*$$

$\Rightarrow$ *Passing from $\|\cdot\|, \omega(\cdot)$ to $\|\cdot\|_X, \omega_X(\cdot)$ spoils MD efficiency at worst by factor $\Omega_X = \Omega[X,\omega_X]$. Thus, with moderate Ω_X, the proximal setup $\|\cdot\|_X, \omega_X(\cdot)$ is nearly optimal.*

♠ **Reason for** (!): For every $g \in E$ and every x with $\|x\|_X \le 1$, so that $x = [u-v]/2$ with $u, v \in X$:

$$\langle g, x\rangle = \tfrac{1}{2}[\langle g, u - x_\omega\rangle + \langle g, x_\omega - v\rangle] \le \tfrac{1}{2}\|g\|_*[\|u - x_\omega\| + \|v - x_\omega\|]$$
$$\le \Omega[X,\omega]\|g\|_* \quad \Rightarrow \quad \|g\|_{X,*} \le \Omega[X,\omega]\|g\|_*$$

Favorable Geometry: Examples

♠ **Examples of Favorable Geometry domains X:**

$$X = B^1 \times ... \times B^K$$

where K is moderate and B^k are *favorable geometry atoms*:

- ℓ_1/ℓ_2 *balls* $B = \{y = [y^1; ...; y^n] : \sum_{j=1}^n \|y^j\|_2 \le 1\}$:

$$\|y\|_B = \sum_{j=1}^n \|y^j\|_2,\ \omega_B(y) = O(1)\sqrt{\ln(n+1)}\sum_{j=1}^n \|y^j\|_2^{\vartheta_n}$$
$$\vartheta_n = \min[2, 1 + 1/\ln(n)] \Rightarrow \Omega_B \le O(1)\sqrt{\ln(n+1)}$$

Note: $n = 1$ *gives rise to Euclidean setup for* $\|\cdot\|_2$*-ball.*

- *Nuclear norm balls* $B = \{y \in \mathbf{R}^{m\times n} : \sum_{i=1}^m \sigma_i(y) \le 1\}$, $m \le n$:

$$\|y\|_B = \sum_{j=1}^m \sigma_j(y),\ \omega_B(y) = O(1)\sqrt{\ln(m+1)}\sum_{j=1}^m \sigma_j^{\theta_m}(y)$$
$$\theta_m = \min[2, 1 + 1/(2\ln(m))] \Rightarrow \Omega_B \le O(1)\sqrt{\ln(m+1)}$$

♠ An induced proximal setup for X is, e.g.,

$$\|(x_1, ..., x_K)\| = \max_k \|x_k\|_{B^k},\ \omega(x_1..., x_k) = \sum_k \omega_{B^k}(x_k)$$
$$\Rightarrow \Omega_X = \sqrt{\sum_k \Omega_{B^k}^2} \le O(1)\sqrt{K\ln(\dim X)}$$

- $K = O(1) \Rightarrow$ *Favorable Geometry case*. This remains true if $X \subset B^1 \times ... \times B^K$ and $\|\cdot\|_X$ is within $O(1)$ factor of $\|\cdot\|$.

Favorable Geometry: Counter-Examples

♠ A domain with *intrinsically bad* geometry is the usual box $X = \{x \in \mathbf{R}^n : \|x\|_\infty \leq 1\}$. *For every proximal setup with* $\|\cdot\| = \|\cdot\|_x = \|\cdot\|_\infty$ *one has* $\Omega[X, \omega] \geq \sqrt{n}$.

♠ In fact, *large-scale* $\|\cdot\|_p$*-balls with all* $p > 2$ *"are bad:"*

Let $p \geq 2$. *Consider Convex Minimization problem*

$$\mathrm{Opt}(f) = \min_x\{f(x) : x \in \mathbf{R}^n, \|x\|_p \leq R\}, \qquad (P_f)$$

$$f \in \mathcal{F}_{n,p}(L) = \{f : \mathbf{R}^n \to \mathbf{R} : \ f \text{ is convex, } L_{\|\cdot\|_p}(f) \leq L\}$$

Assume that when solving (P_f), $f \in \mathcal{F}_{n,p}(L)$ *is learned via calls, one per step, to a FO (or any local) oracle. Then for every* $t \leq n$ *and any* t*-step algorithm* $\mathcal{B}$ *one has*

$$\sup_{f \in \mathcal{F}_{n,p}(L)} [f(x_{\mathcal{B}}(f)) - \mathrm{Opt}(f)] \geq 0.01 LR/t^{1/p}$$

- $x_{\mathcal{B}}(f)$: solution generated in t steps by $\mathcal{B}$ as applied to (P_f)

$\Rightarrow$ *As* $p > 2$ *grows, our abilities to minimize oracle-represented nonsmooth convex functions over* $\|\cdot\|_p$*-balls at a dimension independent rate deteriorate and disappear at* $p = \infty$.

Favorable Geometry: Illustration

♣ The most attractive feature of MD is *ability to adjust itself, to some extent, to problem's geometry and to ensure, under favorable circumstances, (nearly) dimension independent rate of convergence*. For example:

- When minimizing convex f over ℓ_2-ball $\{x \in \mathbf{R}^n : \|x\|_2 \leq 1\}$, MD with *Euclidean setup* ensures

$$f(x^t) - \min_{x \in X} f(x) \leq O(1)[\max_X f - \min_X f]/\sqrt{t}$$

- When minimizing convex f over ℓ_1-ball $\{x \in \mathbf{R}^n : \|x\|_1 \leq 1\}$, MD with appropriate *Non-Euclidean setup* ensures

$$f(x^t) - \min_{x \in X} f(x) \leq O(1)\sqrt{\ln(n+1)}[\max_X f - \min_X f]/\sqrt{t},$$

and similarly for minimizing over nuclear norm ball in $\mathbf{R}^{n \times n}$.

- "Wrong setup" (Euclidean when minimizing over ℓ_1/nuclear norm ball, or ℓ_1/nuclear norm when minimizing over ℓ_2-ball) can spoil the efficiency by factor *as large as* $O(\sqrt{n/\ln(n)})$.

Stochastic case

♣ **Situation:** Given $X \subset E$ and proximal setup $\|\cdot\|, \omega(\cdot)$, we want to process vector fields $g_\tau(x) : X \to E$ represented by *Stochastic Oracle*. At τ-th call to SO, the query point being $x_\tau \in X$, the oracle returns *an estimate* $h_\tau(x_\tau; \xi_\tau) \in E$ of $g_\tau(x_\tau)$. Here $h_\tau(\cdot;\cdot)$ are deterministic functions, and $\xi_1, \xi_2, ...$ are *i.i.d.* disturbances.

♠ **Example:** Problem $\min_{x \in X} \left[f(x) = \mathbf{E}_{\xi \sim P} F(x, \xi) \right]$ with convex in $x \in X$ integrant F.

The associated vector field $g(x) = f'(x)$ is usually difficult to compute. However, assuming one can sample from P and F is easy to compute, we can set

$$h_\tau(x; \xi_\tau) = F'_x(x, \xi_\tau) \text{ with } \xi_1, \xi_2, ... \text{ drawn from } P$$

♠ **Standing Assumption:** *When processing* $\{g_\tau(\cdot)\}_\tau$, *for some* L, σ, μ *and all* $x \in X$, τ *it holds:*

$$\|g_\tau(x)\|_* \le L,\ \|\mathbf{E}_\xi\{\Delta_\tau(x;\xi)\}\|_* \le \mu,\ \mathbf{E}_\xi\{\|\Delta_\tau(x;\xi)\|_*^2\} \le \sigma^2$$

- $\Delta_\tau(x;\xi) := h_\tau(x;\xi) - g_\tau(x)$: oracle's error

Stochastic Mirror Descent

- X: convex compact subset of Euclidean space E
- $\|\cdot\|, \omega(\cdot)$: proximal setup for $(E, X) \Rightarrow \Omega = \sqrt{2[\max_X \omega - \min_x \omega]}$
- $\{g_\tau(x) : X \to E\}_\tau$: vector fields of interest, $\|g_\tau(x)\|_* \le L < \infty$
- $\{h_\tau(x;\xi) = g_\tau(x) + \Delta_\tau(x;\xi) : X \times \Xi \to E\}_\tau$: Stochastic oracle
  $\|\mathbf{E}_{\xi\sim P}\Delta_\tau(x;\xi)\|_* \le \mu$, $\mathbf{E}_{\xi\sim P}\{\|\Delta_\tau(x;\xi)\|_*^2\} \le \sigma^2$

♣ **Stochastic Mirror Descent** is the recurrence

$$x_1 = x_\omega := \operatorname{argmin}_X \omega;\ x_{\tau+1} = \operatorname{Prox}_{x_\tau}(\gamma_\tau h_\tau(x_\tau;\xi_\tau))$$
$$x^t = \textstyle\sum_{\tau\le t} \lambda_\tau^t x_\tau,\ \lambda_\tau^t = \gamma_\tau / \sum_{s\le t}\gamma_s$$

- $\xi_\tau \sim P$: independent
- $\gamma_\tau > 0$: deterministic stepsizes

Stochastic Mirror Descent (continued)

$$x_1 = x_\omega := \mathrm{argmin}_X\, \omega;\ x_{\tau+1} = \mathrm{Prox}_{x_\tau}(\gamma_\tau[g_\tau(x_\tau) + \Delta_\tau(x_\tau;\xi_\tau)])$$
$$x^t = \textstyle\sum_{\tau\le t} \lambda_\tau^t x_\tau,\ \ \lambda_\tau^t = \gamma_\tau / \sum_{s\le t}\gamma_s$$
$$\|g_\tau(x)\|_* \le L,\ \|\mathbf{E}_{\xi\sim P}\Delta_\tau(x;\xi)\|_* \le \mu,\ \ \mathbf{E}_{\xi\sim P}\{\|\Delta_\tau(x;\xi)\|_*^2\} \le \sigma^2$$

♣ **Main Property of SMD:** *One has*

$$\mathbf{E}\left\{\mathcal{E}_t := \max_{z\in X} \textstyle\sum_{\tau\le t} \lambda_\tau^t \langle g(x_\tau), x_\tau - z\rangle\right\} \le \frac{\Omega^2 + [L^2 + 2\sigma^2]\sum_{\tau\le t}\gamma_\tau^2}{\sum_{\tau\le t}\gamma_\tau} + 2\mu\Omega$$

- *In particular,* $\gamma_\tau = \Omega/\sqrt{[L^2+2\sigma^2]t}$, $1 \le \tau \le t$, *yields*

$$\mathbf{E}\{\mathcal{E}_t\} \le \Xi/\sqrt{t} + 2\mu\Omega,\ \ \Xi = 2\Omega\sqrt{L^2+2\sigma^2}.$$

- *Strengthening the bound on the second moment of* $\|\Delta_\tau\|_*$ *to* $\mathbf{E}\{\exp\{\|\Delta_\tau\|_*^2/\sigma^2\}\} \le \exp\{1\}$, *large deviation probabilities obey an exponential bound:*

$$\forall\theta > 0 : \mathrm{Prob}\left\{\mathcal{E}_t > [\Xi + \theta\Sigma]/\sqrt{t} + 2\mu\Omega\right\} \le O(1)\mathrm{e}^{-\theta}$$
$$[\Sigma = 4\Omega\sigma]$$

Stochastic Mirror Descent (continued)

♣ When $g_\tau(\cdot) \equiv g(\cdot)$ is associated with a problem with convex structure, e.g.,

A. $\min_{x\in X} f(x) \Rightarrow g(x) = f'(x)$, or

B. $\min_{u\in U}\max_{v\in V} f(u,v) \Rightarrow g(u,v) = [f'_u(u,v); -f'_v(u,v)]$,

the residual $\mathcal{E}_t$ upper-bounds inaccuracy of the approximate solution x^t to the problem of interest.

$\Rightarrow$ *t-step SMD allows to solve stochastic convex problems with expected inaccuracy $O(1/\sqrt{t})$.* For example,

- in the case of A, we get
$$\mathbf{E}\{f(x^t) - \min_X f\} \le 2\Omega\sqrt{L^2+2\sigma^2}/\sqrt{t} + 2\mu\Omega$$
- in the case of B, we get

$$\mathbf{E}\{[\bar{f}(u^t) - \min_U \bar{f}] + [\max_V \underline{f} - \underline{f}(v^t)]\} \le 2\Omega\sqrt{L^2+2\sigma^2}/\sqrt{t} + 2\mu\Omega.$$

♠ **Note:** In typical stochastic problems, in *every* dimension, not only a large one, $O(1/\sqrt{t})$ is the *best rate allowed by Statistics*.

Stochastic Mirror Descent: Illustration

♣ Consider *Binary Classification problem* where we draw from a distribution P *examples* $\xi_\tau = (\eta_\tau, y_\tau) \in \mathbf{R}^N \times \{\pm 1\}$ and want to build a *linear classifier* $y \sim \mathrm{sign}(\langle x, \eta \rangle)$.

♠ The problem can be modeled as

$$\mathrm{Opt}(\rho) = \min_{\|x\| \le 1} [p_\rho(x) = p(\rho x) := \mathbf{E}\{\max[1 - y\langle \rho x, \eta \rangle, 0]\}]$$

[$p(x)$: convex upper bound on the probability for x to mis-classify]

- Let $\|\cdot\|$ be (a) $\|\cdot\|_2$, or (b) $\|\cdot\|_1$, or (c) nuclear norm on $\mathbf{R}^N = \mathbf{R}^{m \times n}$

♠ Assuming $\mathbf{E}\{\|\eta\|_*^2\} \le R^2 < \infty$ and setting

$$h(x; \eta, y) \equiv -\rho y \chi(1 - y\langle \rho x, \eta \rangle > 0)\eta,$$
$$g(x) := \mathbf{E}_{\eta, y}\{h(x; \eta, y)\} \in p'_\rho(x)$$

we satisfy Standing Assumption with

$$X = \{\|x\| \le 1\}, \ L = \rho R, \ \sigma = 2\rho R, \ \mu = 0.$$

$\Rightarrow$ *For every $t \ge 1$, drawing a t-element sample from P and applying t-step SMD with appropriate proximal setup, we get a linear classifier ρx^t, $\|x^t\| \le 1$, such that*

$$\mathbf{E}\{p(x^t)\} \le \mathrm{Opt}(\rho) + \rho R t^{-1/2} \times \begin{cases} O(1), & \text{case (a)} \\ O(1)\sqrt{\ln(N)}, & \text{cases (b), (c)} \end{cases}$$

Utilizing Problem's Structure: Mirror Prox

$$\mathrm{Opt} = \min_{x \in X} f(x) \qquad (P)$$

♣ Unimprovable or not, convergence rate $O(1/\sqrt{t})$ is slow. When we can do better?

- One can use *bundle* versions of MD re-utilizing past information. In practice, this improves the convergence pattern at the price of *controlled* increase in the computational cost of a step. *Theoretical complexity bounds, however, remain intact.*
- When f is smooth: $\|f'(x) - f'(x')\|_* \leq \mathcal{M}\|x - x'\|$, the MD efficiency improves to $f(x^t) \leq \Omega^2 \mathcal{M}/t$. *This is of no actual interest: with Nesterov's optimal method for smooth convex minimization one achieves unimprovable in the large-scale case efficiency* $O(1)\Omega^2 \mathcal{M}/t^2$.
- When f is *strongly convex*, properly modified MD converges at the rate $O(1/t)$.
- *For a wide spectrum of "well-structured" f, rate $O(1/t)$ can be achieved by smooth saddle point reformulation of (P).*

Extra-Gradient MD – Mirror Prox

♣ **Situation:** X is a convex compact subset of Euclidean space E, $\|\cdot\|, \omega(\cdot)$ is a proximal setup, $g(\cdot): X \to E$ is a vector field represented by an oracle.

- At τ-th call, $x_\tau \in X$ be the query point, the oracle returns an estimate $h(x_\tau;\xi_\tau) = g(x_\tau) + \Delta(x_\tau;\xi_\tau)$ of $g(x_\tau)$, ξ_τ are i.i.d., and

$$\|\mathbf{E}_\xi\{\Delta(x;\xi)\}\|_* \le \mu, \quad \mathbf{E}_\xi\{\|\Delta(x;\xi)\|_*^2\} \le \sigma^2, \ \forall x \in X$$

- $g(\cdot)$ satisfies

$$\|g(x) - g(x')\|_* \le \mathcal{M}\|x - x'\| + L \ \forall (x, x' \in X)$$

- **Note:** $L = \sigma = \mu = 0 \Leftrightarrow g(\cdot)$ is *Lipschitz* & *precisely* observed.

♣ **Mirror Prox** is the recurrence

$$\begin{array}{l} x_1 = x_\omega; \\ x_\tau \mapsto w_\tau = \mathrm{Prox}_{x_\tau}(\gamma_\tau h(x_\tau;\xi_{2\tau-1})) \\ \qquad\qquad \mapsto x_{\tau+1} = \mathrm{Prox}_{x_\tau}(\gamma_\tau h(w_\tau;\xi_{2\tau})) \\ x^t = \sum_{\tau\le t} \lambda_\tau^t w_\tau, \ \lambda_\tau^t = \gamma_\tau / \sum_{s\le t}\gamma_s \end{array}$$

with deterministic stepsizes $\gamma_\tau > 0$.

Mirror Prox (continued)

- $X \subset E, \|\cdot\|, \omega \Rightarrow \Omega$
- $g(\cdot) : X \to E$: $\|g(x) - g(x')\|_* \le \mathcal{M}\|x - x'\| + L$
- oracle $x \mapsto h(x;\xi) = g(x) + \Delta(x;\xi)$:
$$\|\mathbf{E}_\xi\{\Delta(x;\xi)\}\|_* \le \mu, \ \ \mathbf{E}_\xi\{\|\Delta(x;\xi)\|_*^2\} \le \sigma^2$$
- $x_\tau \mapsto w_\tau = \mathrm{Prox}_{x_\tau}(\gamma_\tau h(x_\tau;\xi_{2\tau-1})) \mapsto x_{\tau+1} = \mathrm{Prox}_{x_\tau}(\gamma_\tau h(w_\tau;\xi_{2\tau}))$
$$x^t = \textstyle\sum_{\tau\le t} \lambda_\tau^t w_\tau, \ \ \lambda_\tau^t = \gamma_\tau / \sum_{s\le t}\gamma_s$$

♣ **Main Property of MP:** *Let* $0 < \gamma_\tau \le \frac{1}{2\mathcal{M}}$. *Then*

$$\mathbf{E}\left\{\mathcal{E}_t := \max_{z\in X} \textstyle\sum_{\tau\le t} \lambda_\tau^t \langle g(x_\tau), x_\tau - z\rangle\right\} \le \frac{\Omega^2 + [3L^2 + 7\sigma^2]\sum_{\tau\le t}\gamma_\tau^2}{\sum_{\tau\le t}\gamma_\tau} + 2\mu\Omega$$

- In particular, $\gamma_\tau = \min\left[(2\mathcal{M})^{-1}, \Omega/\sqrt{[3L^2+7\sigma^2]t}\right], \tau \le t,$

yields

$$\mathbf{E}\{\mathcal{E}_t\} \le 2\Omega^2\mathcal{M}/t + \Xi/\sqrt{t} + 2\mu\Omega, \ \ \Xi = 2\Omega\sqrt{3L^2+7\sigma^2}.$$

♠ **Note:** In the smooth deterministic case $L = \sigma = \mu = 0$, we get $O(1/t)$ convergence!

Mirror Prox (continued)

- $X \subset E$, $\|\cdot\|, \omega \Rightarrow \Omega$
- $g(\cdot) : X \to E$: $\|g(x) - g(x')\|_* \le \mathcal{M}\|x - x'\| + L$
- oracle $x \mapsto h(x;\xi) = g(x) + \Delta(x;\xi)$:
$$\|\mathbf{E}_\xi\{\Delta(x;\xi)\}\|_* \le \mu, \quad \mathbf{E}_\xi\{\|\Delta(x;\xi)\|_*^2\} \le \sigma^2$$
- $x_\tau \mapsto w_\tau = \mathrm{Prox}_{x_\tau}(\gamma_\tau h(x_\tau;\xi_{2\tau-1})) \mapsto x_{\tau+1} = \mathrm{Prox}_{x_\tau}(\gamma_\tau h(w_\tau;\xi_{2\tau}))$
$$x^t = \textstyle\sum_{\tau\le t} \lambda_\tau^t w_\tau, \ \lambda_\tau^t = \gamma_\tau / \sum_{s\le t}\gamma_s$$

♠ $\gamma_\tau = \min\left[(2\mathcal{M})^{-1}, \Omega/\sqrt{[3L^2+7\sigma^2]t}\right]$, $1 \le \tau \le t$, yields

$$\mathbf{E}\{\mathcal{E}_t\} \le \frac{2\Omega^2\mathcal{M}}{t} + \frac{\Xi}{\sqrt{t}} + 2\mu\Omega, \ \Xi = 2\Omega\sqrt{3L^2+7\sigma^2}.$$

- Strengthening the bound on the second moment of $\|\Delta\|_*$ to $\mathbf{E}\{\exp\{\|\Delta\|_*^2/\sigma^2\}\} \le \exp\{1\}$, large deviation probabilities obey an exponential bound:

$$\forall\theta > 0 : \mathrm{Prob}\left\{\mathcal{E}_t > 2\Omega^2\mathcal{M}/t + [\Xi + \theta\Sigma]/\sqrt{t} + 2\mu\Omega\right\} \le O(1)\mathrm{e}^{-\theta}$$
$$[\Sigma = 9\Omega\sigma]$$

Application: $O(1/t)$ Nonsmooth Convex Minimization

$$\mathrm{Opt}(P) = \min_{u \in U} f(u) \qquad (P)$$

♣ **Corollary:** *Let (P) be a convex program with compact $U \subset E_U$ and with f such that*

$$f(u) = \max_{v \in V} \phi(u, v)$$

- V: compact convex subset of Euclidean space E_v
- $\phi(u, v)$: convex-concave with *Lipschitz continuous* gradient

so that (P) is the primal form of the saddle point problem

$$\min_{u \in U} \max_{v \in V} \phi(u, v) \qquad (SP)$$

The vector field $g(u, v) = [\phi'_u(u, v); -\phi'_v(u, v)]$ associated with (SP) is Lipschitz continuous. Equipping

- *$E := E_U \times E_V, X := U \times V$ — with a proximal setup $\|\cdot\|, \omega$,*
- *$g(\cdot)$ — with a precise deterministic oracle,*

t-step MP yields $(u^t, v^t) \in U \times V$ such that

$$f(u^t) - \mathrm{Opt}(P) \le O(1)\Omega\mathcal{M}/t$$

$$\mathcal{M} = \min\{M : \|g(x) - g(x')\|_* \le M\|x - x'\| \ \forall (x, x' \in X)\}$$

$O(1/t)$ Nonsmooth Convex Minimization (continued)

$$\min_{u \in U} [f(u) = \max_{v \in V} \phi(u, v)]$$

♣ **Fact:** *If $\phi(u, v)$ is*

- *convex-concave with Lipschitz continuous gradient,*
- *affine in u,*
- *strongly concave in v,*

then properly modified MP ensures $O(1/t^2)$ convergence rate.

♠ **Note:** The premise does *not* imply smoothness of f.

Smooth and Bilinear Saddle Point Representations

♣ **Fact:** *Representations* $f(u) = \max_{v \in V} \phi(u, v)$ *with smooth convex-concave, and even with bilinear* ϕ *are available for wide spectrum of convex functions* f*. Whenever it is the case,* f *can be minimized via MP at the rate* $O(1/t)$*.*

- $f(u) = \max\limits_{k \le K} f_k(u)$ with smooth convex f_k

$$\Rightarrow f(u) = \max_{v \ge 0, \sum_k v_k = 1} \sum_k v_k f_k(u)$$

- $f(u) = \|Au - b\| \Rightarrow f(u) = \max_{\|v\|_* \le 1} \langle v, Ay - b \rangle$
- $f(u) = \|y\| + \frac{1}{2}\|Au - b\|_2^2$

$$\Rightarrow f(u) = \max_{\|v\|_* \le 1, w} \left[\langle u, v \rangle + \langle w, Au - b \rangle - \frac{1}{2} w^T w \right]$$

- $f(u)$: sum of k largest eigenvalues of $\mathcal{A}(u) = Au - b \in \mathbf{S}^n$

$$\Rightarrow f(u) = \max_v \left[\mathrm{Tr}(v\mathcal{A}(u)) : 0 \preceq v \preceq I_n, \mathrm{Tr}(v) = k \right]$$

- $f(u) = \inf_{b \in \mathbf{R}} \left[\frac{1}{N} \sum_{i=1}^N \max[1 - y_i(\langle u, \eta_i \rangle + b), 0] \right]$

$$\Rightarrow f(u) = \max_{v \in V} \sum_{i=1}^N v_i [1 - y_i \langle u, \eta_i \rangle]$$

$$V = \{v : 0 \le v_i \le 1/N \forall i, \textstyle\sum_i y_i v_i = 0\} \subset \{v \in \mathbf{R}^N : \|v\|_1 \le 1\}$$

$O(1/t)$ Nonsmooth Convex Minimization: Comments

$$\mathrm{Opt}(P) = \min_{u \in U} f(u) \qquad (P)$$

• Convex programs always have a lot of structure (otherwise, how could we know that the problem is convex?)
Accelerating algorithms by utilizing problem's structure is an old and still challenging goal.

• A common way to utilize structure is via "structure-revealing" *conic* formulations (Linear/Conic Quadratic/Semidefinite) and Interior Point Methods. However, *in the large scale case IPM iteration may become prohibitively costly*.

• Utilizing structure *within the realm of oracle-oriented methods with computationally cheap iterations* is due to Nesterov (2003).
Nesterov's Smoothing (2003) uses saddle point representation of a nonsmooth f to approximate f by a *smooth* function which is further minimized by Nesterov's algorithm for smooth convex minimization. The resulting convergence rate is $O(1/t)$.

• MP offers another way to utilize saddle point representation to achieve the same $O(1/t)$ rate.

"Practical scopes" of these two approaches are nearly identical.

$O(1/t)$ Nonsmooth Convex Minimization: Examples

♣ **Problem of interest:**

$$\mathrm{Opt}(P) = \min_{\|u\|\le 1} \|Au - b\|_p, \;\; A : M \times N \qquad (P)$$

where $p = 2$ or $p = \infty$, and $\|\cdot\|$ is

(a) $\|\cdot\|_2$ on $\mathbf{R}^N$, or (b) $\|\cdot\|_1$ on $\mathbf{R}^N$, or (c) nuclear norm on $\mathbf{R}^N = \mathbf{R}^{m\times n}$

♠ Bilinear saddle point reformulation is

$$\mathrm{SadVal} = \min_{u\in U} \max_{v\in V} \langle v, Au - b\rangle$$

$$U = \{\|u\| \le 1\},\; V = \{\|v\|_q \le 1\}, q = \tfrac{p}{p-1} \in \{1, 2\}$$

and its domain is the product of two favorable geometry atoms.

♠ Applying t-step MP with appropriate setup, we get u^t with $\|u^t\| \le 1$ and

$$f(u^t) - \mathrm{Opt}(P) \le \kappa \|A\|_{\|\cdot\|,p}/t$$

$$\|A\|_{\|\cdot\|,p} = \max\{\|Au\|_p : \|u\| \le 1\}$$

$$\kappa = O(1)\ln^{1/2-1/p}(M+1) \times \begin{cases} 1, & \text{case (a)} \\ \sqrt{\ln(N+1)}, & \text{case (b)} \\ \sqrt{\ln(m+1)}, & \text{case (c)} \end{cases}$$

$O(1/t)$ Nonsmooth Convex Minimization: Examples

$$\mathrm{Opt}(P) = \min_{\|u\|\le 1} \|Au - b\|_p, \;\; A : M \times N, \; p \in \{2, \infty\} \qquad (P)$$

$\|\cdot\|$: (a) $\|\cdot\|_2$ on $\mathbf{R}^N$ | (b) $\|\cdot\|_1$ on $\mathbf{R}^N$ | (c) nuclear norm on $\mathbf{R}^N = \mathbf{R}^{m\times n}$

$$\Rightarrow \boxed{f(u^t) - \mathrm{Opt}(P) \le O(1)\ln(MN)\|A\|_{\|\cdot\|,p}/t}$$

♠ MP step reduces to computing $O(1)$ matrix-vector products involving A and A^*, plus

— $O(M+N)$ a.o. in cases (a), (b)

— computing svd's of two $m \times n$ matrices in case (c).

$\Rightarrow$ Except for case (c), *MP is computationally cheap...*

♠ **Note:** When solving a *Least Squares* problem

$$\text{(LS)} \quad \mathrm{Opt}(A, b) = \min_{\|u\|_2 \le 1} \|Au - b\|_2 \qquad [A : n \times n]$$

with A represented by *multiplication oracle* $u, u' \mapsto Au, A^T u'$, the rate $O(1/t)$ is unimprovable in the large-scale case:

• *The worst-case, over (A, b) with $\|A\|_{2,2} \le 1$ and $\mathrm{Opt}(A, b) = 0$, inaccuracy in terms of the objective of (LS) is, for every t-step algorithm, at least $O(1)/t$, provided $t \le n/4$.*

Acceleration by Randomization

♣ **Problem of interest:**

$$\begin{array}{rl} & \mathrm{Opt} = \min_{\|u\|_1\le 1} \|Au-b\|_p \qquad [A: m\times n,\ p\in\{2,\infty\}] \\ \Leftrightarrow\ (\ell_1): & \min_{\|u\|_1\le 1} \max_{\|v\|_{p/(p-1)}\le 1} \langle v, Au-b\rangle \\ \Rightarrow & g(u,v) = [A^T v; b-Au] : X := U\times V \to \mathbf{R}^{m+n} \\ & U=\{u: \|u\|_1\le 1\},\ V=\{v: \|v\|_{p/(p-1)}\le 1\}. \end{array}$$

♠ *Omitting from now on logarithmic in m, n factors,* MP solves (ℓ_1) within accuracy ε in

$$N(\varepsilon) = \|A\|_{1,p}/\varepsilon, \quad \|A\|_{1,p} = \max_{j\le n} \|\mathrm{Col}_j[A]\|_p$$

steps, with two multiplications of vectors from U and from V by A, A^T, plus $O(m+n)$ a.o. "overhead," per step.

$\Rightarrow$ *The arithmetic cost of ε-solution for a general-type A is*

$$\mathcal{C}_d(\varepsilon) = mn\|A\|_{1,p}/\varepsilon \text{ a.o.}$$

In fact, this is the best operation count achievable in the large-scale case with known so far *deterministic* algorithms.

- *For large m, n, matrix-vector multiplications may become too time consuming...*

Acceleration by Randomization (continued)

♠ Matrix-vector multiplications are easy to randomize:

In order to compute Bu, $B \in \mathbf{R}^{M\times N}$, we draw an index $\jmath$ at random according to

$$\mathrm{Prob}\{\jmath = j\} = \mathrm{sign}(u_j)/\|u\|_1,\ 1 \le j \le N$$

and return the vector

$$h = \|u\|_1 \mathrm{sign}(u_\jmath)\mathrm{Col}_\jmath[B]$$

Note:

- $\mathbf{E}\{h\} = Bu$, $\|h\|_q \le \|u\|_1 \|B\|_{1,q}$
- Generating h costs $O(1)(M+N)$ a.o. (assuming cost $O(1)$ of computing/extracting individual entry of B).

Acceleration by Randomization (continued)

$$\begin{array}{rl} & \mathrm{Opt} = \min_{\|u\|_1\le 1} \|Au-b\|_p \qquad [A: m\times n,\ p\in\{2,\infty\}] \\ \Leftrightarrow (\ell_1): & \min_{\|u\|_1\le 1} \max_{\|v\|_{p/(p-1)}\le 1} \langle v, Au-b\rangle \\ \Rightarrow & g(u,v) = [A^T v; b-Au] : X := U\times V \to \mathbf{R}^{m+n} \\ & U=\{u: \|u\|_1\le 1\},\ V=\{v: \|v\|_{p/(p-1)}\le 1\}. \end{array}$$

♠ *When solving* (ℓ_1) *with* $p=\infty$ *by MP with the precise values of* $g(\cdot)$ *replaced with their cheap unbiased random estimates, we* $(1-\delta)$*-reliably get* ε*-solution to* (ℓ_1) *in* $\mathrm{Ln}(1/\delta)\left[\|A\|_{1,\infty}/\varepsilon\right]^2$ *steps, the total computational effort being*

$$\mathcal{C}_r = (m+n)\,\mathrm{Ln}(1/\delta)\left[\|A\|_{1,\infty}/\varepsilon\right]^2 \text{ a.o.}$$

♠ The "deterministic" operations count is $\mathcal{C}_b = mn\|A\|_{1,\infty}/\varepsilon$.

$\Rightarrow$ *With the relative accuracy* $\varepsilon/\|A\|_{1,\infty}$ *and* δ *fixed and* m, n *large, randomized algorithm by far outperforms its deterministic competitors.*

- In addition, *Randomized MP exhibits sublinear time behavior: when* m, n *are large,* ε*-solution is obtained, in a* $(1-\delta)$*-reliable fashion, by inspecting negligibly small fraction of the* mn *data entries.*

Acceleration by Randomization (continued)

♠ In the case of $p = \infty$, our construction basically recovers the *ad hoc* sublinear time algorithm for matrix games (Grigoriadis & Khachiyan, 1994).

♠ *In the case of* $p = 2$, randomization leads to iteration count $\ln(1/\delta)[\|A\|_{1,2}/\varepsilon]^2\Gamma^2[A]$, $\Gamma(A) = \sqrt{m}\|A\|_{1,\infty}/\|A\|_{1,2} \in [1, \sqrt{m}]$ and operation count $\mathcal{C}_r = (m+n)\ln(1/\delta)[\|A\|_{1,2}/\varepsilon]^2\Gamma^2[A]$ a.o. vs. the "deterministic" operation count $\mathcal{C}_d = mn[\|A\|_{1,2}/\varepsilon]$ a.o.

- with $\Gamma[A]$ like $O(1)\ln(mn)$, everything is as when $p = \infty$
- with $\Gamma[A]$ as large as $O(\sqrt{m})$, randomization is really bad.

♠ **However:** Preprocessing $[A, b] \Rightarrow [\bar{A}, \bar{b}] = \mathbf{F}\mathrm{Diag}\{\chi\}[A, b]$ with $m \times m$ DFT matrix $\mathbf{F}$ and $\chi \sim \mathrm{Uniform}(\{-1; 1\}^m)$ yields *equivalent* problem (P) and ensures $(1-\delta)$-reliably $\Gamma[\bar{A}] \leq \sqrt{\ln(mn/\delta)}$.

$\Rightarrow$ *With randomization and preprocessing, the operation count is*

$$\mathcal{C}_r = mn + (m+n)\ln^2(1/\delta)[\|A\|_{1,2}/\varepsilon]^2$$

which for small and fixed $\varepsilon/\|A\|_{1,2}$ *and large* m, n *is negligibly small as compared to* $\mathcal{C}_d = mn[\|A\|_{1,2}/\varepsilon]$ *a.o.*

How it Works: Policeman vs. Burglar

♣ **Problem:** There are n houses in a city, i-th with wealth w_i. Every evening, Burglar chooses a house i to be attacked, and Policeman chooses his post near a house j. The probability for Policeman to catch Burglar is

$\exp\{-\theta \mathrm{dist}(i,j)\}$, $\mathrm{dist}(i,j)$: distance between houses i and j.

Burglar wants to maximize his expected profit

$$w_i(1 - \exp\{-\theta \mathrm{dist}(i,j)\}),$$

the interest of Policeman is completely opposite.

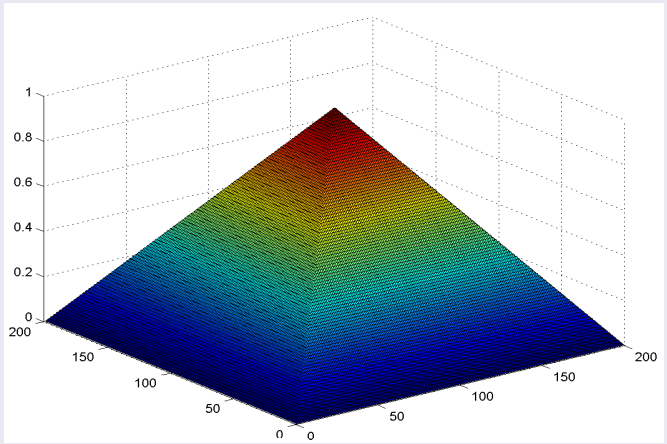
- *What are the optimal mixed strategies of Burglar and Policeman?*

♠ **Equivalently:** *Solve the matrix game*

$$\min_{u \geq 0,\ \sum_{j=1}^n u_j = 1} \ \max_{v \geq 0,\ \sum_{i=1}^n v_i = 1} \ \phi(u,v) := v^T A u$$

$$A_{ij} = w_i(1 - \exp\{-\theta \mathrm{dist}(i,j)\})$$

Policeman vs. Burglar (continued)



Wealth on $n \times n$ square grid of houses

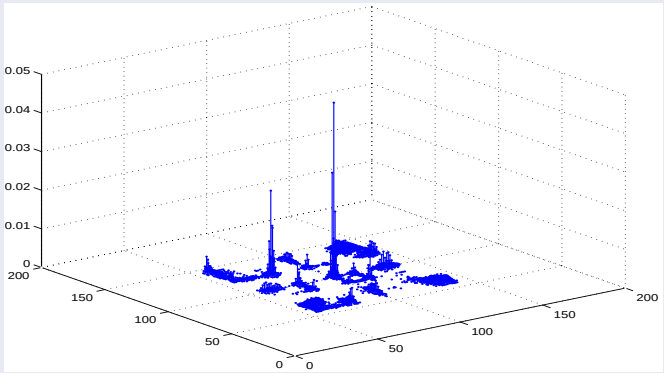
	IPM	MP	Rand MP
N	Steps/CPU, sec/$\mathcal{E}$	Steps/CPU, sec/$\mathcal{E}$	Steps/CPU, sec/$\mathcal{E}$
1600	21/120/6.0e-9	78/6/1.0e-3	10556/264/1.0e-3
6400	21/6930/1.1e-8	80/31/1.0e-3	10408/796/1.0e-3
14400	not tested	95/171/1.0e-3	9422/1584/1.0e-3
40000	out of memory	$15^{\dagger}$/$5533^{\dagger}$/$0.022^{\dagger}$	10216/4931/1.0e-3

Policeman vs. Burglar, N houses

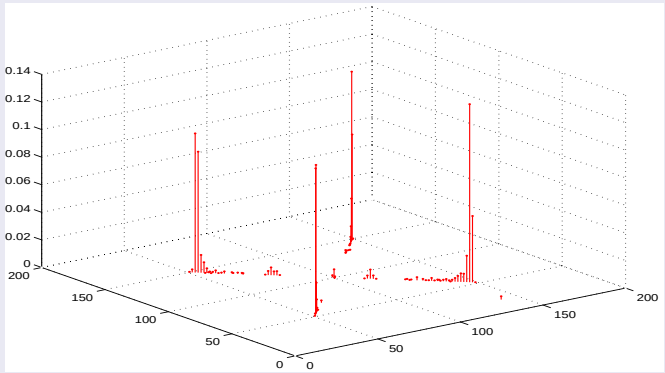
Target residual $\varepsilon_t \leq$ 1.e-3 IPM: `mosekopt`

†: *termination when reaching the CPU limit of 5,400 sec*

Policeman vs. Burglar (continued)

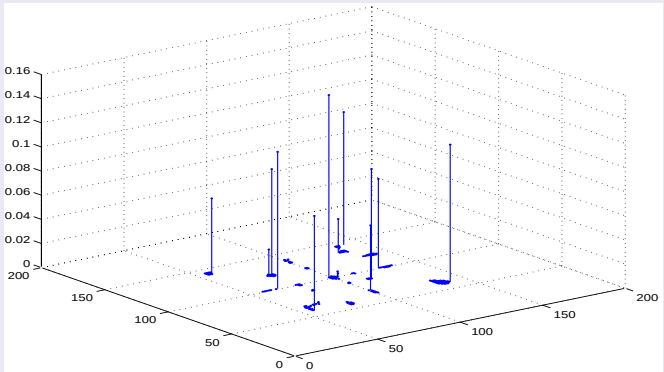


Policeman

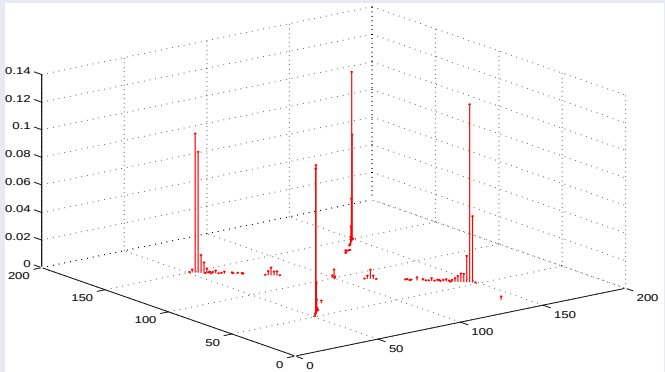


Burglar

♠ The resulting highly sparse near-optimal solution can be refined by further optimizing it on its support by an interior point method. This reduces inaccuracy from 0.0008 to 0.0005 in just 39′.



Policeman, refined



Burglar, refined

200×200 grid of houses

References

- A. Beck, M. Teboulle, Mirror Descent... – *OR Letters* **31** '03
- M. Grigoriadis, L. Khachiyan, A Sublinear Time... – *OR Letters* **18** '95
- A. Juditsky, F. Kilinç Karzan, A. Nemirovski, ℓ_1 Minimization... ('11), http://www.optimization-online.org
- A. Juditsky, A. Nemirovski, First Order... I,II: S. Sra, S. Novozin, S.J. Wright, Eds., *Optimization for Machine Learning*, MIT Press, 2012
- A. Nemirovski, Efficient Methods... – *Ekonomika i Mat. Metody* **15** '79
- A. Nemirovski, D. Yudin, Problem Complexity... – J. Wiley '83
- A. Nemirovski, Information-Based... – *J. of Complexity* **8** '92
- A. Nemirovski, Prox-Method... – *SIAM J. Optim.* **15** '04
- Yu. Nesterov, A Method for Solving... – *Soviet Math. Dokl.* **27:2** '83
- Yu. Nesterov, Smooth Minimization... – *Math. Progr.* **103** '05
- Yu. Nesterov, Excessive Gap Technique... *SIAM J. Optim.* **16:1** '05
- Yu. Nesterov, Gradient Methods for Minimizing... CORE Discussion Paper '07/76

Tutorial: Mirror Descent Algorithms for Large-Scale Deterministic and Stochastic Convex Optimization
Selected Proofs

1 Mirror Descent: Stochastic Case

Situation: At t-th call to Stochastic Oracle, the query point being $x_t \in X$, the SO returns vector $g_t(x_t) + \Delta_t(x_t, \xi_t)$, with independent $\xi_t \sim P$, $t = 1, 2, ...$ Besides this, for all $x \in X$ and all t it holds

$$\|g_t(\cdot)\|_* \le L < \infty, \mathbf{E}_\xi\{\|\Delta_t(x,\xi)\|_*^2\} \le \sigma^2, \|\mathbf{E}_\xi\{\Delta_t(x,\xi)\}\|_* \le \mu. \tag{1}$$

For the MD recurrence

$$x_1 = x_\omega; x_{t+1} = \mathrm{Prox}_{x_t}(\gamma_t[g_t(x_t) + \Delta_t(x_t, \xi_t)]) \tag{2}$$

with deterministic stepsizes $\gamma_t > 0$, we have, setting $g_t = g_t(x_t)$, $\Delta_t = \Delta_t(x_t, \xi_t)$:

$$\begin{array}{ll}
& \gamma_t\langle g_t + \Delta_t, x_{t+1} - x\rangle \le V_{x_t}(x) - V_{x_{t+1}}(x) - V_{x_t}(x_{t+1}) \text{ [see (2) in Transparencies]} \\
\Rightarrow & \gamma_t\langle g_t + \Delta_t, x_t - x\rangle \le V_{x_t}(x) - V_{x_{t+1}}(x) + [\gamma_t\langle g_t + \Delta_t, x_t - x_{t+1}\rangle - V_{x_t}(x_{t+1})] \\
\Rightarrow & \gamma_t\langle g_t + \Delta_t, x_t - x\rangle \le V_{x_t}(x) - V_{x_{t+1}}(x) + \frac{1}{2}\gamma_t^2\|g_t + \Delta_t\|_*^2 \\
\Rightarrow & \sum_{t=1}^T \gamma_t\langle g_t, x_t - x\rangle \le \frac{1}{2}\Omega^2 + \frac{1}{2}\sum_{t=1}^T \|g_t + \Delta_t\|_*^2 + \sum_{t=1}^T \gamma_t\langle \Delta_t, x - x_t\rangle \\
\Rightarrow & \left[\sum_{t=1}^T \gamma_t\right]\epsilon_T \le \frac{1}{2}\Omega^2 + \frac{1}{2}\sum_{t=1}^T \|g_t + \Delta_t\|_*^2 + \max_{x\in X}\sum_{t=1}^T \gamma_t\langle \Delta_t, x - x_t\rangle \\
& \epsilon_T := \max_{x\in X}\sum_{t=1}^T \lambda_t\langle g_t, x_t - x\rangle,\ \lambda_t = \gamma_t/\sum_{s=1}^T \gamma_s.
\end{array}$$

The bottom line is that

$$\left[\sum_{t=1}^T \gamma_t\right]\epsilon_T \le \frac{1}{2}\Omega^2 + \frac{1}{2}\sum_{t=1}^T \|g_t + \Delta_t\|_*^2 + \max_{x\in X}\sum_{t=1}^T \gamma_t\langle \Delta_t, x - x_t\rangle \tag{3}$$

Our goal is to prove the following

Theorem 1.1 (i) *Assuming* (1)*, for the recurrence* (2) *for every* $T = 1, 2, ...$ *one has*

$$\mathbf{E}\left\{\epsilon_T\right\} \leq \Xi := \frac{\Omega^2 + [L^2 + \frac{3}{2}\sigma^2]\sum_{t=1}^T \gamma_t^2}{\sum_{t=1}^T \gamma_t} + 2\mu\Omega. \tag{4}$$

(ii) *Strengthening* (1) *to*

$$\|g_t(\cdot)\|_* \leq L < \infty, \mathbf{E}_\xi\{\exp\{\|\Delta_t(x,\xi)\|_*^2/\sigma^2\}\} \leq \exp\{1\},\ \|\mathbf{E}_\xi\{\Delta_t(x,\xi)\}\|_* \leq \mu \tag{5}$$

for all $x \in X$ *and all* t*, we have for every* $\theta > 0$*:*

$$\begin{array}{c}\mathrm{Prob}\{\epsilon_T > \Xi + \theta\Sigma\} \leq 6\exp\{-\theta\} + \exp\{-\theta^2/4\},\\ \Sigma = 2\frac{\sigma^2\sum_{t=1}^T \gamma_t^2 + \sigma\Omega\sqrt{\sum_{t=1}^T \gamma_t^2}}{\sum_{t=1}^T \gamma_t}.\end{array} \tag{6}$$

Note that with the stepsizes

$$\gamma_t = \frac{\Omega}{\sqrt{L^2 + 3\sigma^2/2}\sqrt{T}},\ 1 \leq t \leq T \tag{7}$$

one has

$$\Xi = \frac{2\Omega\sqrt{L^2 + 3\sigma^2/2}}{\sqrt{T}} + 2\mu\Omega,\ \Sigma \leq 4\frac{\Omega\sigma}{\sqrt{T}}. \tag{8}$$

Proof.

1^0. We need the following

Lemma 1.1 *Given deterministic* γ_t *and (perhaps, stochastic)* g_t*,* Δ_t *such that* $\|g_t\|_* \leq L < \infty$ *and*

$$\mathbf{E}\{\|\Delta_t\|_*^2\} \leq \sigma^2\,\forall t, \tag{9}$$

one has

$$\mathbf{E}\{\sum_{t=1}^T \gamma_t^2\|g_t + \Delta_t\|_*^2\} \leq 2\sum_{t=1}^T \gamma_t^2[L^2 + \sigma^2]. \tag{10}$$

If (9) *is strengthened to*

$$\mathbf{E}\{\exp\{\|\Delta_t\|_*^2/\sigma^2\}\} \le \exp\{1\}, \tag{11}$$

one has

$$\forall\theta>0: \mathrm{Prob}\left\{\sum_{t=1}^T \gamma_t^2\|g_t+\Delta_t\|_*^2 > 2\sum_{t=1}^T \gamma_t^2[L^2+\sigma^2] + 2\theta\sigma^2\sum_{t=1}^T\gamma_t^2\right\} \le \exp\{1-\theta\}. \tag{12}$$

Proof. (10) is evident due to $\|g+\Delta\|_*^2 \le 2[\|g\|_*^2+\|\Delta\|_*^2]$. To prove (12), note that the quantity

$$\chi(G(\cdot)) = \inf\{s>0 : \mathbf{E}\{\exp\{|G|/s\}\} \le \exp\{1\}\}$$

considered as functional on the space of measurable functions $G(\cdot)$ on the probability space associated with $\mathbf{E}\{\cdot\}$ such that $\chi(G(\cdot))$ is finite, clearly is a norm. It follows that in the case of (11), setting $a = 2\sigma^2\sum_{t=1}^T\gamma_t^2$, we have

$$\mathbf{E}\{\exp\{[2\sum_{t=1}^T \gamma_t^2\|\Delta_t\|_*^2]/a\}\} \le \exp\{1\},$$

whence $\mathrm{Prob}\{[2\sum_{t=1}^T\gamma_t^2\|\Delta_t\|_*^2] \ge \theta a\} \le \exp\{1-\theta\}$. Since

$$\sum_{t=1}^T \gamma_t^2\|g_t+\Delta_t\|_*^2 > 2\sum_{t=1}^T\gamma_t^2[L^2+\sigma^2] + 2\theta\sum_{t=1}^T\gamma_t^2\sigma^2$$

clearly implies that $2\sum_{t=1}^T\gamma_t^2\|\Delta_t\|_*^2 > 2\sum_{t=1}^T\gamma_t^2\sigma^2 + 2\theta\sigma^2\sum_{t=1}^T\gamma_t^2$ due to $\|g_t\|_* \le L$, (12) follows. □

2^0. Our next observation is as follows:

Lemma 1.2 *Let $\Delta_t = H_t(\xi^t) \in E$, where $H_t(\cdot)$ are deterministic functions, and $\xi^t = (\xi_1,...,\xi_t)$ with independent $\xi_t \sim P$, $t = 1,2,...$, and let $x_t = X_t(\xi^{t-1}) \in X$, with deterministic $X_t(\cdot)$. Assuming*

$$\begin{array}{ll}(a) & \mathbf{E}_{\xi_t\sim P}\left\{\|H_t(\xi^t)\|_*^2\right\} \le \sigma^2,\ \forall(t,\xi^{t-1}) \\ (b) & \|\mathbf{E}_{\xi_t\sim P}\{H_t(\xi^t)\}\|_* \le \mu\ \forall(t,\xi^{t-1})\end{array} \tag{13}$$

we have for deterministic $\gamma_t \geq 0$*:*

$$\mathbf{E}\{\max_{x\in X}\sum_{t=1}^{T}\gamma_t\langle\Delta_t, x-x_t\rangle\} \leq \frac{1}{2}\Omega^2 + \frac{1}{2}\sigma^2\sum_{t=1}^{T}\gamma_t^2 + 2\mu\Omega\sum_{t=1}^{T}\gamma_t. \tag{14}$$

Strengthening (13.*a*) *to*

$$\mathbf{E}_{\xi_t\sim P}\left\{\exp\{\|H_t(\xi^t)\|_*^2/\sigma^2\}\right\} \leq \exp\{1\}, \tag{15}$$

we get

$$\begin{array}{rl} \forall(\theta>0): & \\ \mathrm{Prob}\Big\{ & \max_{x\in X}\sum_{t=1}^{T}\gamma_t\langle\Delta_t, x-x_t\rangle \geq \frac{1}{2}\Omega^2 + \frac{1}{2}\sigma^2\sum_{t=1}^{T}\gamma_t^2 + 2\mu\Omega\sum_{t=1}^{T}\gamma_t \\ & \qquad\qquad +\theta\left[2\sigma\Omega\sqrt{\sum_{t=1}^{T}\gamma_t^2} + \sigma^2\sum_{t=1}^{T}\gamma_t^2\right]\Big\} \\ \leq & \exp\{1-\theta\} + \exp\{-\theta^2/4\} \end{array} \tag{16}$$

Proof. A. Let $y_t = Y_t(\xi^{t-1})$ be given by the recurrence

$$y_1 = y_\omega; y_{t+1} = \mathrm{Prox}_{y_t}(-\gamma_t\Delta_t).$$

Then, same as in the derivation of (3),

$$\forall x\in X: \sum_{t=1}^{T}\gamma_t\langle -\Delta_t, y_t - x\rangle \leq \frac{1}{2}\Omega^2 + \frac{1}{2}\sum_{t=1}^{T}\gamma_t^2\|\Delta_t\|_*^2,$$

so that

$$\forall x\in X: \sum_{t=1}^{T}\gamma_t\langle\Delta_t, x-x_t\rangle \leq \frac{1}{2}\Omega^2 + \frac{1}{2}\sum_{t=1}^{T}\gamma_t^2\|\Delta_t\|_*^2 + \sum_{t=1}^{T}\gamma_t\langle\Delta_t, y_t-x_t\rangle,$$

whence

$$A := \max_{x\in X}\sum_{t=1}^{T}\gamma_t\langle\Delta_t, x-x_t\rangle \leq \frac{1}{2}\Omega^2 + \frac{1}{2}\sum_{t=1}^{T}\gamma_t^2\|\Delta_t\|_*^2 + \sum_{t=1}^{T}\gamma_t\langle\Delta_t, y_t-x_t\rangle. \tag{17}$$

Since y_t, x_t are deterministic functions of ξ^{t-1} and γ_t are deterministic, we have

$$\mathbf{E}\{\langle \Delta_t, y_t - x_t\rangle\} = \mathbf{E}\{\mathbf{E}_{\xi_t\sim P}\{\langle \Delta_t, y_t - x_t\rangle\}\} = \mathbf{E}\{\langle \mathbf{E}_{\xi_t\sim P}\{\Delta_t\}, y_t - x_t\rangle\} \le \mathbf{E}\{\mu\|y_t - x_t\|\} \le 2\mu\Omega$$

(we have used that $x_t, y_t \in X$). Thus, taking expectation of both sides in (17), we get

$$\mathbf{E}\{A\} \le \frac{1}{2}\Omega^2 + \frac{1}{2}\sigma^2\sum_{t=1}^{T}\gamma_t^2 + 2\mu\Omega\sum_{t=1}^{T}\gamma_t,$$

and (14) follows.

B. Now assume that in addition to (13.a) relation (15) takes place, and let us prove (16). Note that (15) implies (13.b) by Jensen's inequality.

C. We start with the following observation:

Lemma 1.3 *Let η be a scalar random variable such that $|\mathbf{E}\{\eta\}| \le \nu$ and $\mathbf{E}\{\exp\{\eta^2\}\} \le \exp\{1\}$. Then for all $\alpha \in \mathbf{R}$ it holds $\mathbf{E}\{\exp\{\alpha\eta\}\} \le \exp\{\alpha\nu + \alpha^2\}$.*

Proof. We have $\mathrm{e}^s \le s + \mathrm{e}^{2s^2/3}$ for all s, whence $\mathbf{E}\{\mathrm{e}^{\alpha\eta}\} \le \mathbf{E}\{\alpha\eta + \mathrm{e}^{2\alpha^2\eta^2/3}\}$. When $\alpha^2 \le 3/2$, we have $\mathbf{E}\{\mathrm{e}^{2\alpha^2\eta^2/3}\} \le \exp\{2\alpha^2/3\}$ due to $\mathbf{E}\{\mathrm{e}^{\eta^2}\} \le \exp\{1\}$ and Moment inequality, so that

$$\mathbf{E}\{\mathrm{e}^{\alpha\eta}\} \le |\alpha||\mathbf{E}\{\eta\}| + \mathrm{e}^{2\alpha^2/3} \le \mathrm{e}^{|\alpha|\nu + 2\alpha^2/3},\ 0 \le \alpha^2 \le 3/2.$$

Besides this, we have $\alpha s \le \frac{1}{4}\alpha^2 + s^2$, whence $\mathbf{E}\{\mathrm{e}^{\alpha\eta}\} \le \mathrm{e}^{1+\alpha^2/4}$ due to $\mathbf{E}\{\mathrm{e}^{s^2}\} \le \mathrm{e}$. Combining the bounds, we get $\mathbf{E}\{\mathrm{e}^{\alpha\eta}\} \le \mathrm{e}^{|\alpha|\nu+\alpha^2}$ for all α. □

D. Let $s_t = \gamma_t\langle \Delta_t, y_t - x_t\rangle$. Since y_t, x_t depend solely on ξ^{t-1} and $|s_t| \le 2\gamma_t\|\Delta_t\|_*\Omega$ due to $x_t, y_t \in X$, (15) implies that setting $\sigma_t = 2\gamma_t\sigma\Omega$, we have

$$\mathbf{E}_{\xi_t\sim P}\{\mathrm{e}^{s_t^2/\sigma_t^2}\} \le \exp\{1\}\,\forall(t, \xi^{t-1}) \tag{18}$$

Besides this, $|\mathbf{E}_{\xi_t\sim P}\{s_t\}| = |\gamma_t\langle \mathbf{E}_{\xi_t\sim P}\{\Delta_t\}, y_t - x_t\rangle| \le 2\gamma_t\Omega\mu$ for all t, ξ^{t-1}. Applying Lemma 1.3 to the random variable $\eta = s_t/\sigma_t$, which allows to set $\nu = \mu/\sigma$, we get

$$\mathbf{E}_{\xi_t\sim P}\{\mathrm{e}^{\alpha s_t}\} \le \mathrm{e}^{\mu\alpha\sigma_t/\sigma + \alpha^2\sigma_t^2}.$$

Now, for every $r > 0$ we have

$$\mathrm{Prob}\{\sum_{t=1}^{T}\gamma_t\langle\Delta_t, v_t - x_t\rangle > r\} \le \mathbf{E}\{\exp\{\alpha\sum_{t=1}^{T}s_t\}\}\exp\{-\alpha r\}\,\forall\alpha \ge 0.$$

Setting $S_0 = 0$, $S_t = \sum_{\tau=1}^{t} s_\tau$, we have for $\alpha \ge 0$:

$$\begin{array}{l}\mathbf{E}\{\exp\{\alpha S_t\}\} = \mathbf{E}\{\exp\{\alpha S_{t-1} + \alpha s_t\}\} = \mathbf{E}\{\exp\{\alpha S_{t-1}\}\mathbf{E}_{\xi_t\sim P}\{\mathrm{e}^{\alpha s_t}\}\}\\ \le \mathbf{E}\{\exp\{\alpha S_{t-1}\}\exp\{\mu\alpha\sigma_t/\sigma + \alpha^2\sigma_t^2\}\},\end{array}$$

so that

$$\mathbf{E}\{\exp\{\alpha S_T\}\} \le \exp\{\sum_{t=1}^{T}[\mu\alpha\sigma_t/\sigma + \alpha^2\sigma_t^2]\}$$

and thus

$$\mathrm{Prob}\{\sum_{t=1}^{T}\gamma_t\langle\Delta_t, v_t - x_t\rangle > r\} \le \inf_{\alpha>0}\left[\exp\{\alpha[\sum_{t=1}^{T}\mu\sigma_t/\sigma - r] + \alpha^2\sum_{t=1}^{T}\sigma_t^2\right]. \tag{19}$$

Assuming

$$r = \mu\sum_{t=1}^{T}\sigma_t/\sigma + \rho = 2\mu\Omega\sum_{t=1}^{T}\gamma_t + \rho$$

with some positive ρ, and setting $\alpha = \frac{\rho}{2\sum_{t=1}^{T}\sigma_t^2}$, we get from (19) that

$$\forall\rho > 0 : \mathrm{Prob}\{\sum_{t=1}^{T}\gamma_t\langle\Delta_t, v_t - x_t\rangle > 2\mu\Omega\sum_{t=1}^{T}\gamma_t + \rho\} \le \exp\{-\rho^2/\left(4\sum\nolimits_{t=1}^{T}\sigma_t^2\right)\},$$

or, which is the same,

$$\forall\theta > 0 : \mathrm{Prob}\{\sum\nolimits_{t=1}^{T}\gamma_t\langle\Delta_t, v_t - x_t\rangle > 2\mu\Omega\sum\nolimits_{t=1}^{T}\gamma_t + 2\theta\sigma\Omega\sqrt{\sum\nolimits_{t=1}^{T}\gamma_t^2}\} \le \exp\{-\theta^2/4\}. \tag{20}$$

E. Acting exactly as in the proof of Lemma 1.1 with L set to 0, we get

$$\forall\theta > 0 : \mathrm{Prob}\left\{\sum_{t=1}^{T}\gamma_t^2\|\Delta_t\|_*^2 > \sigma^2\sum_{t=1}^{T}\gamma_t^2 + \theta\sigma^2\sum_{t=1}^{T}\gamma_t^2\right\} \le \exp\{1-\theta\}. \tag{21}$$

This combines with (20) and (17) to imply (16). Lemma 1.2 is proved. □

3^0. Now we can prove Theorem 1.1. Combining (14), (10) and (3), we arrive at (4); (i) is proved. In the case of (5), we have at our disposal both (12) and (16), and these two relations clearly imply item (ii) of Theorem. □

2 Mirror Prox: Stochastic Case

Situation: For every $t = 1, 2, ...$, and $(2t-1)$-st call to Stochastic Oracle, the query point being $x_t \in X$, the SO returns vector $g_t(x_t) + \Delta_{2t-1}(x_t, \xi_{2t-1})$; at $(2t)$-th call, the query point being $w_t \in X$, the SO returns $g_t(w_t) + \Delta_{2t}(w_t, \xi_{2t})$, with independent $\xi_s \sim P$, $s = 1, 2, ...$ Besides this, we have

$$\begin{array}{ll} (a) & \|g_t(x) - g_t(x')\|_* \le \mathcal{M}\|x - x'\| + L \ \forall (x, x' \in X, t = 1, 2, ...) \quad [\mathcal{M}, L < \infty], \\ (b) & \mathbf{E}_\xi\{\|\Delta_s(x, \xi)\|_*^2\} \le \sigma^2, \ \forall (x \in X, s = 1, 2, ...) \\ (c) & \|\mathbf{E}_\xi\{\Delta_s(x, \xi)\}\|_* \le \mu \, \forall (x \in X, s = 1, 2, ...) \end{array} \tag{22}$$

For the MP recurrence

$$x_1 = x_\omega; w_t = \mathrm{Prox}_{x_t}(\gamma_t[g_t(x_t) + \Delta_{2t-1}(x_t, \xi_{2t-1})]); x_{t+1} = \mathrm{Prox}_{x_t}(\gamma_t[g_t(w_t) + \Delta_{2t}(w_t, \xi_{2t})])$$

we have, setting $\widehat{g}_t = g_t(x_t)$, $g_t = g_t(w_t)$, $\eta_t = \Delta_{2t-1}(x_t, \xi_{2t-1})$, $\zeta_t = \Delta_{2t}(w_t, \xi_{2t})$:

$$\begin{array}{rl} & \gamma_t\langle g_t + \zeta_t, x_{t+1} - x\rangle \le V_{x_t}(x) - V_{x_{t+1}}(x) - V_{x_t}(x_{t+1}) \text{ [see (2) in Transparencies]} \\ & \gamma_t\langle \widehat{g}_t + \eta_t, w_t - x_{t+1}\rangle \le V_{x_t}(x_{t+1}) - V_{w_t}(x_{t+1}) - V_{x_t}(w_t) \text{ [see (2) in Transparencies]} \\ \Rightarrow & \gamma_t\langle g_t + \zeta_t, w_t - x\rangle \le V_{x_t}(x) - V_{x_{t+1}}(x) + [\gamma_t\langle g_t + \zeta_t, w_t - x_{t+1}\rangle - V_{x_t}(x_{t+1})] \\ \Rightarrow & \gamma_t\langle g_t + \zeta_t, w_t - x\rangle \le V_{x_t}(x) - V_{x_{t+1}}(x) + \gamma_t\langle g_t + \zeta_t - \widehat{g}_t - \eta_t, w_t - x_{t+1}\rangle + \langle \widehat{g}_t + \eta_t, w_t - x_{t+1}\rangle - V_{x_t}(x_{t+1}) \\ \le & V_{x_t}(x) - V_{x_{t+1}}(x) + \gamma_t\langle g_t + \zeta_t - \widehat{g}_t - \eta_t, w_t - x_{t+1}\rangle + V_{x_t}(x_{t+1}) - V_{w_t}(x_{t+1}) - V_{x_t}(w_t) - V_{x_t}(x_{t+1}) \\ \Rightarrow & \gamma_t\langle g_t + \zeta_t, w_t - x\rangle \le V_{x_t}(x) - V_{x_{t+1}}(x) + [\gamma_t\langle g_t + \zeta_t - \widehat{g}_t - \eta_t, w_t - x_{t+1}\rangle - V_{w_t}(x_{t+1}) - V_{x_t}(w_t)] \\ \Rightarrow & \gamma_t\langle g_t + \zeta_t, w_t - x\rangle \le V_{x_t}(x) - V_{x_{t+1}}(x) + \gamma_t\|g_t - \widehat{g}_t\|_*\|w_t - x_{t+1}\| + \gamma_t\|\zeta_t - \eta_t\|_*\|w_t - x_{t+1}\| - \frac{1}{2}\|x_t - w_t\|^2 - \frac{1}{2}\|x_{t+1} - w_t\|^2 \end{array}$$

Assuming

$$\gamma_t\mathcal{M} \le 1/2, \tag{23}$$

we have

$$\begin{array}{rl} & \gamma_t\|g_t - \widehat{g}_t\|_*\|w_t - x_{t+1}\| + \gamma_t\|\zeta_t - \eta_t\|_*\|w_t - x_{t+1}\| - \frac{1}{2}\|x_t - w_t\|^2 - \frac{1}{2}\|x_{t+1} - w_t\|^2 \\ \le & \gamma_t\mathcal{M}\|w_t - x_t\|\|w_t - x_{t+1}\| + \gamma_t L\|w_t - x_{t+1}\| + \gamma_t\|\zeta_t - \eta_t\|_*\|w_t - x_{t+1}\| \\ & \qquad -\frac{1}{2}\|x_t - w_t\|^2 - \frac{1}{2}\|x_{t+1} - w_t\|^2 \\ \le & \frac{1}{2}[-\|x_t - w_t\|^2 - \frac{1}{4}\|x_{t+1} - w_t\|^2 + \|w_t - x_t\|\|w_t - x_{t+1}\|] \\ & \qquad +[\gamma_t[L + \|\zeta_t - \eta_t\|_*]\|w_t - x_{t+1}\| - \frac{3}{8}\|w_t - x_{t+1}\|^2] \\ \le & \gamma_t^2[L + \|\zeta_t\|_* + \|\eta_t\|_*]^2 \le 3\gamma_t^2[L^2 + \|\zeta_t\|_*^2 + \|\eta_t\|_*^2] \end{array}$$

Thus, for every $x \in X$ and every t we have

$$\gamma_t\langle g_t + \zeta_t, w_t - x\rangle \le V_{x_t}(x) - V_{x_{t+1}}(x) + 3\gamma_t^2[L^2 + \|\zeta_t\|_*^2 + \|\eta_t\|_*^2],$$

whence

$$\sum_{t=1}^T \gamma_t\langle g_t, w_t - x\rangle \le \frac{1}{2}\Omega^2 + 3\sum_{t=1}^T \gamma_t^2[L^2 + \|\zeta_t\|_*^2 + \|\eta_t\|_*^2] + \sum_{t=1}^T \gamma_t\langle\zeta_t, x - w_t\rangle.$$

Therefore with $\lambda_t := \lambda_t^T = \gamma_t/\sum_{s=1}^T \gamma_s$ we have

$$\epsilon_T := \max_{x\in X}\sum_{t=1}^T \lambda_t\langle g_t, w_t - x\rangle \le \frac{\frac{1}{2}\Omega^2 + 3\sum_{t=1}^T \gamma_t^2[L^2 + \|\zeta_t\|_*^2 + \|\eta_t\|_*^2] + \max\limits_{x\in X}\sum_{t=1}^T \gamma_t\langle\zeta_t, x - w_t\rangle}{\sum_{\tau=1}^T \gamma_\tau} \tag{24}$$

Our goal is to prove the following

Theorem 2.1 (i) *Assuming* (22), (23), *one has for every* $T = 1, 2, ...$

$$\mathbf{E}\{\epsilon_T\} \le \Xi := \frac{\Omega^2 + [3L^2 + 13\sigma^2/2]\sum_{t=1}^T \gamma_t^2}{\sum_{t=1}^T \gamma_t} + 2\mu\Omega. \tag{25}$$

(ii) *Strengthening* (22.*b*) *to*

$$\mathbf{E}_\xi\{\exp\{\|\Delta_s(x,\xi)\|_*^2/\sigma^2\}\} \le \exp\{1\}\ \forall(x \in X, s = 1, 2, ...) \tag{26}$$

we have for every $\theta > 0$:

$$\begin{array}{c}\mathrm{Prob}\{\epsilon_T > \Xi + \theta\Sigma\} \le 6\exp\{-\theta\} + \exp\{-\theta^2/4\},\\ \Sigma = \frac{7\sigma^2\sum_{t=1}^T\gamma_t^2 + 2\sigma\Omega\sqrt{\sum_{t=1}^T\gamma_t^2}}{\sum_{t=1}^T\gamma_t}.\end{array} \tag{27}$$

Note that with the stepsizes

$$\gamma_t = \min\left[\frac{1}{2\mathcal{M}}, \frac{\Omega}{\sqrt{3L^2 + 13\sigma^2/2}\sqrt{T}}\right],\ 1 \le t \le T \tag{28}$$

one has

$$\Xi = \frac{2\Omega^2\mathcal{M}}{T} + \frac{2\Omega\sqrt{3L^2 + 13\sigma^2/2}}{\sqrt{T}} + 2\mu\Omega,\ \Sigma \le 9\frac{\Omega\sigma}{\sqrt{T}}. \tag{29}$$

Proof repeats word by word the one of Theorem 1.1, with (24) in the role of (3).

3 Proximal Setup for ℓ_1/ℓ_2 Ball

Let

$$X = \{x = [x^1; ...; x^n] \subset E = \mathbf{R}^{k_1} \times ...\mathbf{R}^{k_n} : \sum_{j=1}^n \|x^j\|_2 \le 1\}$$

and

$$\omega(x) = \frac{1}{p\gamma}\sum_{j=1}^p \|x^j\|_2^p,\ p = \left\{\begin{array}{ll} 2, & n \le 2\\ 1 + \frac{1}{\ln n}, & n \ge 3\end{array}\right.,\ \gamma = \left\{\begin{array}{ll} 1, & n = 1\\ \frac{1}{2}, & n = 2\\ \frac{1}{\mathrm{e}\ln(n)}, & n > 2\end{array}\right.$$

We have for $x \in X' = \{x \in X : x^j \neq 0 \, \forall j\}$:

$$
\begin{array}{l}
\gamma D\omega(x)[h] = \sum_{j=1}^n \|x^j\|_2^{p-2} \langle x^j, h^j \rangle \\
\gamma D^2\omega(x)[h,h] = -(2-p) \sum_{j=1}^n \|x^j\|_2^{p-4} [\langle x^j, h^j \rangle]^2 + \sum_{j=1}^n \|x^j\|_2^{p-2} \|h^j\|_2^2 \\
\geq \sum_{j=1}^n \|x^j\|_2^{p-2} \|h^j\|_2^2 - (2-p) \sum_{j=1}^n \|x^j\|_2^{p-4} \|x^j\|_2^2 \|h^j\|_2^2 \\
\geq (p-1) \sum_{j=1}^n \|x^j\|_2^{p-2} \|h^j\|_2^2 \\
\Rightarrow \left[\sum_j \|h^j\|_2\right]^2 = \left[\sum_{j=1}^n [\|h^j\|_2 \|x^j\|_2^{\frac{p-2}{2}}] \|x^j\|_2^{\frac{2-p}{2}}\right]^2 \leq \left[\sum_{j=1}^n \|h^j\|_2^2 \|x^j\|_2^{p-2}\right] \left[\sum_{j=1}^n \|x^j\|_2^{2-p}\right] \\
\Rightarrow \left[\sum_j \|h^j\|_2\right]^2 \leq \left[\sum_{j=1}^n \|x^j\|_2^{2-p}\right] \frac{\gamma}{p-1} D^2\omega(x)[h,h]
\end{array}
$$

Setting $t_j = \|x^j\|_2 \geq 0$, we have $\sum_j t_j \leq 1$, whence due to $0 \leq 2-p \leq 1$ it holds $\sum_j t_j^{2-p} \leq n n^{-(2-p)} = n^{p-1}$. Thus,

$$
\left[\sum_j \|h^j\|_2\right]^2 \leq n^{p-1} \frac{\gamma}{p-1} D^2\omega(x)[h,h]
$$

while

$$
\max_{x \in X} \omega(x) - \min_{x \in X} \omega(x) \leq \frac{1}{\gamma p} \tag{30}
$$

With p, γ as above, when $n \geq 3$ we get $\frac{\gamma}{p-1} n^{p-1} = \frac{1}{\mathrm{e} \ln(n)/\ln(n)} n^{1/\ln(n)} = 1$, and similarly for $n = 1, 2$. Consequently,

$$
\forall (x \in X', h) : \left[\sum_{j=1}^n \|\Delta_j\|_2\|\right]^2 \leq D^2\omega(x)[h,h]. \tag{31}
$$

Since $\omega(\cdot)$ is continuously differentiable and the complement of X' in X is the union of finitely many proper linear subspaces of E, (31) implies that ω is strongly convex on X, modulus 1, w.r.t. the ℓ_1/ℓ_2 norm. Besides this, we have

$$
\frac{1}{\gamma p} = \left\{ \begin{array}{ll} \frac{1}{2}, & n = 1 \\ 2, & n = 2 \\ \leq \mathrm{e} \ln(n), & n \geq 3 \end{array} \right\} \leq O(1) \ln(n+1).
$$

which combines with (30) to imply that the ω-radius of X is $\leq O(1)\sqrt{\ln(n+1)}$.

4 Proximal Setup for Nuclear Norm Ball

For $y \in \mathbf{S}^n$, let $\lambda(y)$ be the vector of eigenvalues of y (taken with their multiplicities in the non-ascending order), and let $|y|_1 = \|\lambda(y)\|_1$ be the trace norm.

Proposition 4.1 *Let $N \geq M \geq 2$, and let E be a linear subspace in $\mathbf{S}^N$ such that every matrix $y \in E$ has at most M nonzero eigenvalues. Let $q = \frac{1}{2\ln(M)}$, so that $0 < q < 1$, and let*

$$\widehat{\omega}(y) = \frac{8\mathrm{e}\ln(M)}{1+q} \sum_{j=1}^{N} |\lambda_j(y)|^{1+q} : \mathbf{S}^N \to \mathbf{R}.$$

The function $\widehat{\omega}(\cdot)$ is continuously differentiable, convex, and its restriction on the set $Y_E = \{y \in E : |y|_1 \leq 1\}$ is strongly convex, modulus 1, w.r.t. $|\cdot|_1$. Besides this,

$$\forall (y \in Y, h \in \mathbf{S}^N) : |\langle \omega'(y), h\rangle| \leq 8\mathrm{e}\ln(M)|h|_1. \tag{32}$$

Proof. 1^0. Let $0 < q < 1$. Consider the following function of $y \in \mathbf{S}^N$:

$$\chi(y) = \frac{1}{1+q} \sum_{i=1}^{N} |\lambda_i(y)|^{1+q} = \mathrm{Tr}(f(y)),\ f(s) = \frac{1}{1+q}|s|^{1+q}.$$

2^0. Function $f(s)$ is continuously differentiable on the axis and twice continuously differentiable outside of the origin; consequently, we can find a sequence of polynomials $f_k(s)$ converging, as $k \to \infty$, to f along with their first derivatives uniformly on every compact subset of $\mathbf{R}$ and, besides this, converging to f uniformly along with the first and the second derivative on every compact subset of $\mathbf{R}\backslash\{0\}$. Now let $y, h \in \mathbf{S}^N$, let $y = u\mathrm{Diag}\{\lambda\}u^T$ be the eigenvalue decomposition of y, and let $h = u\widehat{h}u^T$. For a polynomial $p(s) = \sum_{k=0}^{K} p_k s^k$, setting

$P(w) = \mathrm{Tr}(\sum_{k=0}^K p_k w^k) : \mathbf{S}^N \to \mathbf{R}$, and denoting by γ a closed contour in $\mathbb{C}$ encircling the spectrum of y, we have

$$
\begin{array}{ll}
(a) & P(y) = \mathrm{Tr}(p(y)) = \sum_{j=1}^N p(\lambda_j(y)) \\
(b) & DP(y)[h] = \mathrm{Tr}(\sum_{k=0}^K k p_k \mathrm{Tr}(y^{k-1}h)) = \mathrm{Tr}(p'(y)h) = \sum_{j=1}^N p'(\lambda_j(y))\widehat{h}_{ii} \\
(c) & D^2P(y)[h,h] = \frac{d}{dt}\big|_{t=0} DP(y+th)[h] = \frac{d}{dt}\big|_{t=0}\mathrm{Tr}(p'(y+th)h) \\
& = \frac{q}{dt}\big|_{t=0} \frac{1}{2\pi i}\oint\limits_\gamma \mathrm{Tr}(h(zI-(y+th))^{-1})p'(z)dz = \frac{1}{2\pi i}\oint\limits_\gamma \mathrm{Tr}(h(zI-y)^{-1}h(zI-y)^{-1})p'(z)dz \\
& = \frac{1}{2\pi i}\oint\limits_\gamma \sum_{i,j=1}^N \widehat{h}_{ij}^2 \frac{p'(z)}{(z-\lambda_i(y))(z-\lambda_j(y))}dz = \sum_{i,j=1}^n \widehat{h}_{ij}^2 \Gamma_{ij}, \\
& \Gamma_{ij} = \begin{cases} \frac{p'(\lambda_i(y))-p'(\lambda_j(y))}{\lambda_i(y)-\lambda_j(y)}, & \lambda_i(y) \neq \lambda_j(y) \\ p''(\lambda_i(y)), & \lambda_i(y) = \lambda_j(y) \end{cases}
\end{array}
$$

We conclude from (a,b) that as $k \to \infty$, the real-valued polynomials $F_k(\cdot) = \mathrm{Tr}(f_k(\cdot))$ on $\mathbf{S}^N$ converge, along with their first order derivatives, uniformly on every bounded subset of $\mathbf{S}^N$, and the limit of the sequence, by (a), is exactly $\chi(\cdot)$. Thus, $\chi(\cdot)$ is continuously differentiable, and (b) says that

$$
D\chi(y)[h] = \sum_{j=1}^N f'(\lambda_j(y))\widehat{h}_{jj}. \tag{33}
$$

Besides this, $(a$-$c)$ say that if U is a closed convex set in $\mathbf{S}^N$ which does not contain singular matrices, then $F_k(\cdot)$, as $k \to \infty$, converge along with the first and the second derivative uniformly on every compact subset of U, so that $\chi(\cdot)$ is twice continuously differentiable on U, and at every point $y \in U$ we have

$$
D^2\chi(y)[h,h] = \sum_{i,j=1}^N \widehat{h}_{ij}^2 \Gamma_{ij},\ \Gamma_{ij} = \begin{cases} \frac{f'(\lambda_i(y))-f'(\lambda_j(y))}{\lambda_i(y)-\lambda_j(y)}, & \lambda_i(y) \neq \lambda_j(y) \\ f''(\lambda_i(y)), & \lambda_i(y) = \lambda_j(y) \end{cases} \tag{34}
$$

and in particular $\chi(\cdot)$ is convex on U.

3^0. We intend to prove that (i) $\chi(\cdot)$ is convex, and (ii) its restriction on the unit ball Y of the trace norm is strongly convex, with certain modulus $\alpha > 0$, w.r.t. the trace norm $|\cdot|_1$. Since χ is continuously differentiable, all we need to prove (i) is to verify that

$$
\langle \chi'(y') - \chi'(y''), y' - y'' \rangle \geq 0 \tag{$*$}
$$

for a dense in $\mathbf{S}^n \times \mathbf{S}^N$ set of pairs (y', y''), e.g., those with nonsingular $y' - y''$. For a pair of the latter type, the polynomial $q(t) = \mathrm{Det}(y' + t(y'' - y'))$ of $t \in \mathbf{R}$ is not identically zero and thus has finitely many roots on $[0, 1]$. In other words, we can find finitely many points $t_0 = 0 < t_1 < ... < t_n = 1$ such that all "matrix intervals" $\Delta_i = (y_i, y_{i+1})$, $y_k = y' + t_k(y'' - y')$, $1 \le i \le n-1$, are comprised of nonsingular matrices. Therefore χ is convex on every compact subset of every interval Δ_i, and since χ is continuously differentiable, $(*)$ follows.

4^0. Now let us prove that with properly defined $\alpha > 0$ one has

$$\langle \chi'(y') - \chi'(y''), y' - y'' \rangle \ge \alpha |y' - y''|_1^2 \ \forall y', y'' \in Y_E$$

Let $\epsilon > 0$, and let Y^ϵ be a convex open in $Y = \{y : |y|_1 \le 1\}$ neighbourhood of Y_E such that for all $y \in N^\epsilon$ at most M eigenvalues of y are of magnitude $> \epsilon$. We intend to prove that for some $\alpha_\epsilon > 0$ one has

$$\langle \chi'(y') - \chi'(y''), y' - y'' \rangle \ge \alpha_\epsilon |y' - y''|_1^2 \ \forall y', y'' \in Y^\epsilon. \tag{35}$$

Same as above, it suffices to verify this relation for a dense in $Y^\epsilon \times Y^\epsilon$ set of pairs $y', y'' \in Y^\epsilon$, e.g., for those pairs $y', y'' \in Y^\epsilon$ for which $y' - y''$ is nonsingular. Defining matrix intervals Δ_i as above and taking into account continuous differentiability of χ, it suffices to verify that if $y \in \Delta_i$ and $h = y' - y''$, then $D^2\chi(y)[h, h] \ge \alpha_\epsilon |\Delta_1|^2$. To this end observe that by (34) all we have to prove is that

$$D^2\chi(y)[h, h] = \sum_{i,j=1}^N \widehat{h}_{ij}^2 \Gamma_{ij} \ge \alpha_\epsilon |h|_1^2. \tag{\#}$$

Setting $\lambda_j = \lambda_j(y)$, observe that $\lambda_i \ne 0$ for all i due to the origin of y, and if $|\lambda_i| \ge |\lambda_j|$, then $\Gamma_{ij} \ge q|\lambda_i|^{q-1}$. Indeed, the latter relation definitely holds true when $\lambda_i = \lambda_j$. Now, if λ_i and λ_j are of the same sign, then $\Gamma_{ij} = \frac{|\lambda_i|^q - |\lambda|_j^q}{|\lambda_i| - |\lambda_j|} \ge q|\lambda_i|^{q-1}$, since the derivative of the concave (recall that $0 < q \le 1$) function t^q of $t > 0$ is positive and nonincreasing. If λ_i and λ_j are of different signs, then $\Gamma_{ij} = \frac{|\lambda_i|^q + |\lambda_j|^q}{|\lambda_i| + |\lambda_j|} \ge |\lambda_i|^{q-1}$ due to $|\lambda_j|^q \ge |\lambda_j||\lambda_i|^{q-1}$, and therefore $\Gamma_{ij} \ge q|\lambda_i|^{q-1}$. Without loss of generality, we can assume that the positive reals $\mu_i = |\lambda_i|$, $i = 1, ..., N$, form a nondecreasing sequence, so that, by above, $\Gamma_{ij} \ge q\mu_j^{q-1}$ when $i \le j$. Besides this, at most M of μ_j are $\ge \epsilon$, since $y', y'' \in Y^\epsilon$ and therefore $y \in Y^\epsilon$ by convexity of Y^ϵ. By the above,

$$D^2\chi(y)[h, h] = 2q \sum_{i<j\le N} \widehat{h}_{ij}^2 \mu_j^{q-1} + q \sum_{j=1}^N \widehat{h}_{jj}^2 \mu_j^{q-1},$$

or, equivalently by symmetry of $\widehat{h}$, if

$$h^j = \begin{bmatrix} & & & \widehat{h}_{1j} \\ & & & \widehat{h}_{2j} \\ & & & \vdots \\ \widehat{h}_{j1} & \widehat{h}_{j2} & \cdots & \widehat{h}_{jj} \\ & & & \end{bmatrix}$$

and H_j is the Frobenius norm of h^j, then

$$D^2\chi(y)[h,h] = q\sum_{j=1}^{N} H_j^2\mu_j^{q-1} \geq q\epsilon^{q-1}\sum_{j=1}^{N-M} H_j^2 + q\sum_{j=N-M+1}^{N} H_j^2\mu_j^{q-1}.$$

Now note that $\mu_j > 0$ and $\sum_{j=N-M+1}^{N}\mu_j \leq 1$ due to $y \in Y$. It follows that setting $\eta = [H_{N-M+1}; H_{N-M+2}; ...; H_N]$, we have

$$\begin{array}{l} \sum_{j=N-M+1}^{N} H_j^2\mu_j^{q-1} \geq \min\limits_{\nu_j>0:\sum_{j=N-M+1}^{N}\nu_j\leq 1} \sum_{j=N-M+1}^{N} H_j^2\nu_j^{q-1} = \left[\sum_{j=N-M+1}^{N} H_j^{\frac{2}{2-q}}\right]^{2-q} \\ = \|\eta\|^2_{\frac{2}{2-q}} \geq M^{-2[1-\frac{2-q}{2}]}\|\eta\|_1^2 = M^{-2q}\|\eta\|_1^2, \end{array}$$

(when computing the minimum, take into account that $0 < q < 1$). Besides this, setting $\zeta = [H_1; H_2; ...; H_{N-M}]$, we have

$$\|\zeta\|_1^2 \leq (N-M)\sum_{j=1}^{N-M} H_j^2 \leq [\epsilon^{1-q}(N-M)]\epsilon^{q-1}\sum_{j=1}^{N-M} H_j^2$$

We see that for every positive δ one has

$$\begin{array}{l} \left[\sum_{j=1}^{N} H_j\right]^2 = [\|\zeta\|_1 + \|\eta\|_1]^2 \leq (1+\delta)\|\eta\|_1^2 + (1+\delta^{-1})\|\zeta\|_1^2 \\ \leq (1+\delta)M^{2q}\sum_{j=N-M+1}^{N} H_j^2\mu_j^{q-1} + (1+\delta^{-1})[\epsilon^{1-q}(N-M)]\epsilon^{q-1}\sum_{j=1}^{N-M} H_j^2 \\ \leq \max[(1+\delta)M^{2q}, (1+\delta^{-1})\epsilon^{1-q}(N-M)]\left[\epsilon^{1-q}\sum_{j=1}^{N-M} H_j^2 + \sum_{j=N-M+1}^{N} H_j^2\mu_j^{q-1}\right] \\ \leq q^{-1}\max[(1+\delta)M^{2q}, (1+\delta^{-1})\epsilon^{1-q}(N-M)]D^2\chi(y)[h,h]. \end{array}$$

Now observe that $\widehat{h} = \sum_{j=1}^N h^j$ and h^j is of rank ≤ 2, so that $|h^j|_1$ is at most twice the Frobenius norm H_j of h^j. Therefore

$$|h|_1^2 = |\widehat{h}|_1^2 \le 4\left[\sum_{j=1}^N H_j\right]^2 \le 4q^{-1}\max[(1+\delta)M^{2q}, (1+\delta^{-1})\epsilon^{1-q}(N-M)]D^2\chi(y)[h,h].$$

This inequality holds true for all δ. Setting

$$\alpha_\epsilon^{-1} = \min_{\delta>0} 4q^{-1}\max[(1+\delta)M^{2q}, (1+\delta^{-1})\epsilon^{1-q}(N-M)],$$

we ensure the validity of (#), and consequently the validity of (35). The latter relation, combined with $\alpha_\epsilon \to \alpha = qM^{-2q}/4$ as $\epsilon \to +0$ due to $q < 1$, implies that

$$\langle \chi'(y') - \chi'(y''), y' - y''\rangle \ge \alpha |y' - y''|_1^2 \ \forall (y', y'' \in Y_E),\ \alpha = qM^{-2q}/4.$$

Setting $q = \frac{1}{2\ln(M)}$ and observing that with this q, $\alpha = [8\mathrm{e}\ln(M)]^{-1}$, so that $\widehat{\omega}(\cdot) = \alpha^{-1}\chi(\cdot)$, we see that $\widehat{\omega}$ indeed is continuously differentiable convex function on $\mathbf{S}^N$ which is strongly convex, modulus 1 w.r.t. $|\cdot|_1$, on Y_E. It remains to note that by (33) for $y \in Y$ and $h \in \mathbf{S}^N$ we have

$$\begin{array}{l} |\langle \omega'(y), h\rangle = 8\mathrm{e}\ln(M)|\langle \chi'(y), h\rangle| \le 8\mathrm{e}\ln(M)\sum_{j=1}^N |\lambda_j(y)|^q |\widehat{h}_{jj}| \\ \le 8\mathrm{e}\ln(M)\sum_j |\widehat{h}_{jj}| \le 8\mathrm{e}\ln(M)|\widehat{h}|_1 = 8\mathrm{e}\ln(M)|h|_1. \end{array} \qquad \square$$

Now let m, n be positive integers with $2 \le m \le n$, and let $N = m + n$, $M = m$. For $x \in \mathbf{R}^{m\times n}$, let $\sigma_i(x)$, $1 \le i \le m$, be the singular values of x, let $\|x\|_{\mathrm{nuc}}$ be the nuclear norm of x, and let $\mathcal{A}x = \frac{1}{2}\begin{bmatrix} & x \\ x^T & \end{bmatrix} \in \mathbf{S}^N$. Observe that the image space E of $\mathcal{A}$ is a linear subspace of $\mathbf{S}^N$, and that the eigenvalues of $y = \mathcal{A}x$ are the $2m$ reals $\pm\sigma_i(x)/2$, $1 \le i \le m$, and $N - m$ zeros, so that $\|x\|_{\mathrm{nuc}} \equiv |\mathcal{A}x|_1$ and $M = m, E$ satisfy the premise of Proposition 4.1. Setting

$$\omega(x) = \widehat{\omega}(\mathcal{A}x) = \frac{8\mathrm{e}\ln(m)}{2^q(1+q)}\sum_i \sigma_i^{1+q}(x),\ q = \frac{1}{2\ln(m)},$$

and invoking Proposition 4.1, we see that ω is a convex continuously differentiable function on $\mathbf{R}^{m\times n}$ which, due to the identity $\|x\|_{\mathrm{nuc}} \equiv |\mathcal{A}x|_1$, is strongly convex, modulus 1 w.r.t. $\|\cdot\|_{\mathrm{nuc}}$, on the $\|\cdot\|_{\mathrm{nuc}}$-unit ball X. Observe that

$$\Omega[X, \omega(\cdot)] \le 4\sqrt{\mathrm{e}\ln(m)} \le 7\sqrt{\ln(m)},$$

and that by (32) we have

$$\begin{array}{l}\Omega^+[X,\omega(\cdot)] := \sqrt{2\sup_{x,x'\in X}[\omega(x')-\omega(x)-\langle\omega'(x),x-x'\rangle]}\\ \le \sqrt{2[8\mathrm{e}\ln(m)+8\mathrm{e}\ln(m)\max_{x,x'\in X}\|x-x'\|_{\rm nuc}} = \sqrt{48\mathrm{e}\ln(m)} \le 12\sqrt{\ln(m)}.\end{array} \tag{36}$$

5 Mirror Descent in Semi-Bounded Case

Theorem 5.1 *Let $\|\cdot\|,\omega$ be a proximal setup for $X\subset E$, and assume that X contains $\|\cdot\|$-ball of positive radius $\theta\Omega$ centered at some point c. Consider MD trajectory*

$$x_1 = x_\omega; x_{\tau+1} = \mathrm{Prox}_{x_\tau}(\gamma_\tau g_\tau(x_\tau)) \tag{37}$$

with stepsizes

$$\gamma_\tau = \begin{cases} \nu_\tau/\|g_\tau(x_\tau)\|_*, & g_\tau(x_\tau)\neq 0\\ \nu_\tau, & g_\tau(x_\tau)=0\end{cases} \tag{38}$$

Assume the vector fields $g_\tau(\cdot)$ are uniformly semi-bounded on X:

$$\sup_{x,x'\in X,\tau}\langle g_\tau(x),x'-x\rangle \le L<\infty \tag{39}$$

Then with x^t defined as

$$x^t = \textstyle\sum_{\tau=1}^t \lambda_\tau^t x_\tau,\ \lambda_\tau^t = \gamma_\tau/\sum_{s=1}^t\gamma_s$$

when $g_\tau(x_\tau)\neq 0$ for all $\tau\le t$, otherwise defined as (any) x_τ such that $g_\tau(x_\tau)=0$, the following holds true: when

$$\Delta_t := \frac{\Omega^2+\sum_{\tau=1}^t\nu_\tau^2}{2\sum_{\tau=1}^t\nu_\tau} < \theta\Omega,$$

one has

$$\epsilon_t := \max_{x\in X}\sum_{\tau=1}^t\lambda_\tau^t\langle g_\tau(x_\tau),x_\tau-x\rangle \le \frac{L\Delta_t}{\theta\Omega-\Delta_t}. \tag{40}$$

In particular, whet $t \geq 4\theta^{-2}$ *and* $\nu_\tau = \frac{\Omega}{\sqrt{t}}$, $1 \leq \tau \leq t$, *one has*

$$\epsilon_t \leq \frac{2L}{\theta\sqrt{t}}.$$

Proof. There is nothing to prove when $x^t = x_\tau$ such that $g_\tau(x_\tau) = 0$; thus assume that $g_\tau(x_\tau) \neq 0$ for all $\tau \leq t$. Let $h_\tau(x) = g_\tau(x)/\|g_\tau(x)\|_*$ when $g_\tau(x) \neq 0$, and $h_\tau(x) = 0$ when $g_\tau(x) = 0$. Then the recurrence (37) reads

$$x_1 = x_\omega; x_{\tau+1} = \mathrm{Prox}_{x_\tau}(\nu_\tau h_\tau(x_\tau)) \tag{41}$$

and $\|h_\tau(x_\tau)\|_* \leq 1$, whence

$$\begin{array}{c}\max_{x\in X} \sum_{\tau=1}^t \mu_\tau \langle h_\tau(x_\tau), x_\tau - x\rangle \leq \Delta_t := \frac{\Omega^2+\sum_{\tau=1}^t \nu_\tau^2}{2\sum_{\tau=1}^t \nu_\tau}, \\ \mu_\tau = \nu_\tau/\sum_{s=1}^t \nu_s\end{array} \tag{42}$$

or

$$\max_{x\in X} \sum_{\tau\leq t} \frac{\mu_\tau}{\|g_\tau(x_\tau)\|_*} \langle g_\tau(x_\tau), x_\tau - x\rangle \leq \Delta_t. \tag{43}$$

Assuming w.l.o.g. $c = 0$, we have by (39)

$$\forall (x, \|x\| \leq r := \theta\Omega, \tau) : \ \langle g_\tau(x_\tau), x - x_\tau\rangle \leq L,$$

whence

$$\|g_\tau(x_\tau)\|_* \leq r^{-1}[L + \langle g_\tau, x_\tau\rangle] \, \forall \tau$$

or, equivalently,

$$\forall (\tau \leq t) : \frac{1}{\|g_\tau(x_\tau)\|_*} \geq \frac{r}{L} - \frac{1}{L}\frac{\langle g_\tau(x_\tau), x_\tau\rangle}{\|g_\tau(x_\tau)\|_*}$$

and therefore

$$\sum_{\tau\leq t} \frac{\mu_\tau}{\|g_\tau(x_\tau)\|_*} \geq \frac{r}{L}\sum_{\tau\leq t}\mu_\tau - \frac{1}{L}\sum_{\tau\leq t}\mu_\tau \frac{\langle g_\tau(x_\tau), x_\tau\rangle}{\|g_\tau(x_\tau)\|_*}.$$

In other words,

$$\frac{r}{L} \leq \sum_{\tau\leq t} \frac{\mu_\tau}{\|g_\tau(x_\tau)\|_*} + \frac{1}{L}\sum_{\tau\leq t}\mu_\tau \frac{\langle g_\tau, x_\tau\rangle}{\|g_\tau(x_\tau)\|_*} \leq \sum_{\tau\leq t} \frac{\mu_\tau}{\|g_\tau(x_\tau)\|_*} + \frac{1}{L}\Delta_t,$$

where the concluding $\leq$ is due to (43) and $0 \in X$. We see that

$$\sum_{\tau \leq t} \frac{\mu_\tau}{\|g_\tau(x_\tau)\|_*} \geq \frac{r - \Delta_t}{L}.$$

Assuming the right hand side in this inequality positive and taking into account that

$$\lambda_\tau^t = \frac{\mu_\tau / \|g_\tau(x_\tau)\|_*}{\sum_{s \leq t} \mu_s / \|g_s(x_s)\|_*},$$

we get

$$\max_{x \in X} \sum_{\tau \leq t} \lambda_\tau^t \langle g_\tau(x_\tau), x_\tau - x \rangle \leq \frac{L \Delta_t}{r - \Delta_t},$$

as claimed in (40). $\square$