How Good Are Sparse Cutting-Planes?*

Santanu S. Dey, Marco Molinaro, and Qianyi Wang

School of Industrial and Systems Engineering, Georgia Institute of Technology

Abstract. Sparse cutting-planes are often the ones used in mixed-integer programing (MIP) solvers, since they help in solving the linear programs encountered during branch-&-bound more efficiently. However, how well can we approximate the integer hull by just using sparse cutting-planes? In order to understand this question better, given a polyope $\mathbf{P}$ (e.g. the integer hull of a MIP), let $\mathbf{P}^k$ be its best approximation using cuts with at most k non-zero coefficients. We consider $\mathrm{d}(\mathbf{P}, \mathbf{P}^k) = \max_{x \in \mathbf{P}^k} (\min_{y \in \mathbf{P}} \|x - y\|)$ as a measure of the quality of sparse cuts. In our first result, we present general upper bounds on $\mathrm{d}(\mathbf{P}, \mathbf{P}^k)$ which depend on the number of vertices in the polytope and exhibits three phases as k increases. Our bounds imply that if $\mathbf{P}$ has polynomially many vertices, using half sparsity already approximates it very well. Second, we present a lower bound on $\mathrm{d}(\mathbf{P}, \mathbf{P}^k)$ for random polytopes that show that the upper bounds are quite tight. Third, we show that for a class of hard packing IPs, sparse cutting-planes do not approximate the integer hull well. Finally, we show that using sparse cutting-planes in extended formulations is at least as good as using them in the original polyhedron, and give an example where the former is actually much better.

1 Introduction

Most successful mixed integer linear programming (MILP) solvers are based on branch-&-bound and cutting-plane (cut) algorithms. Since MILPs belong to the class of NP-hard problems, one does not expect the size of branch-&-bound tree to be small (polynomial is size) for every instance. In the case where the branch-&-bound tree is not small, a large number of linear programs must be solved. It is well-known that dense cutting-planes are difficult for linear programming solvers to handle. Therefore, most commercial MILPs solvers consider sparsity of cuts as an important criterion for cutting-plane selection and use [4, 1, 7].

Surprisingly, very few studies have been conducted on the topic of sparse cutting-planes. Apart from cutting-plane techniques that are based on generation of cuts from single rows (which implicitly lead to sparse cuts if the underlying row is sparse), to the best of our knowledge only the paper [2] explicitly discusses methods to generate sparse cutting-planes.

The use of sparse cutting-planes may be viewed as a compromise between two competing objectives. As discussed above, on the one hand, the use of sparse

* Santanu S. Dey and Qianyi Wang were partially supported by NSF grant CMMI-1149400.

cutting-planes aids in solving the linear programs encountered in the branch-&-bound tree faster. On the other hand, it is possible that 'important' facet-defining or valid inequalities for the convex hull of the feasible solutions are dense and thus without adding these cuts, one may not be able to attain significant integrality gap closure. This may lead to a larger branch-&-bound tree and thus result in the solution time to increase.

It is challenging to simultaneously study both the competing objectives in relation to cutting-plane sparsity. Therefore, a first approach to understanding usage of sparse cutting-planes is the following: *If we are able to separate and use valid inequalities with a given level of sparsity (as against completely dense cuts), how much does this cost in terms of loss in closure of integrality gap?*

Considered more abstractly, the problem reduces to a purely geometric question: Given a polytope $\mathbf{P}$ (which represents the convex hull of feasible solutions of a MILP), how well is $\mathbf{P}$ approximated by the use of sparse valid inequalities. In this paper we will study polytopes contained in the $[0,\ 1]^n$ hypercube. This is without loss of generality since one can always translate and scale a polytope to be contained in the $[0,\ 1]^n$ hypercube.

1.1 Preliminaries

A cut $ax \leq b$ is called *k-sparse* if the vector a has at most k nonzero components. Given a set $\mathbf{P} \subseteq \mathbb{R}^n$, define $\mathbf{P}^k$ as the best outer-approximation obtained from k-sparse cuts, that is, it is the intersection of all k-sparse cuts valid for $\mathbf{P}$.

For integers k and n, let $[n] := \{1, \ldots, n\}$ and let $\binom{[n]}{k}$ be the set of all subsets of $[n]$ of cardinality k. Given a k-subset of indices $I \subseteq [n]$, define $\mathbb{R}^{\bar{I}} = \{x \in \mathbb{R}^n : x_i = 0 \text{ for all } i \in I\}$. An equivalent and handy definition of $\mathbf{P}^k$ is the following: $\mathbf{P}^k = \bigcap_{I \in \binom{[n]}{k}} \left(\mathbf{P} + \mathbb{R}^{\bar{I}}\right)$. Thus, if $\mathbf{P}$ is a polytope, $\mathbf{P}^k$ is also a polytope.

1.2 Measure of Approximation

There are several natural measures to compare the quality of approximation provided by $\mathbf{P}^k$ in relation to $\mathbf{P}$. For example, one may consider objective value ratio: maximum over all costs c of expression $\frac{z^{c,k}}{z^c}$, where $z^{c,k}$ is the value of maximizing c over $\mathbf{P}^k$, and z^c is the same for $\mathbf{P}$. We discard this ratio, since this ratio can become infinity and not provide any useful information[1]. Similarly, we may compare the volumes of $\mathbf{P}$ and $\mathbf{P}^k$. However, this ratio is not useful if $\mathbf{P}$ is not full-dimensional and $\mathbf{P}^k$ is.

In order to have a useful measure that is well-defined for all polytopes contained in $[0,1]^n$, we consider the following *distance measure*:

$$\mathrm{d}(\mathbf{P}, \mathbf{P}^k) := \max_{x \in \mathbf{P}^k} \left(\min_{y \in \mathbf{P}} \|x - y\| \right),$$

where $\|\cdot\|$ is the ℓ_2 norm. It is easily verified that there is a vertex of $\mathbf{P}^k$ attaining the maximum above. Thus, alternatively the distance measure can be interpreted as the Euclidean distance between $\mathbf{P}$ and the farthest vertex of $\mathbf{P}^k$ from $\mathbf{P}$.

[1] Take $\mathbf{P} = \mathrm{conv}\{(0,0),(0,1),(1,1)\}$ and compare with $\mathbf{P}^1$ wrt $c = (1,-1)$.

Observation 1 ($\mathbf{d(P, P^k)}$ is an upper bound on depth of cut) *Suppose $\alpha x \leq \beta$ is a valid inequality for $\mathbf{P}$ where $\|\alpha\| = 1$. Let the* depth *of this cut be the smallest $\gamma \geq 0$ such that $\alpha x \leq \beta + \gamma$ is valid for $\mathbf{P}^k$. It is straightforward to verify that $\gamma \leq d(\mathbf{P}, \mathbf{P}^k)$. Therefore, the distance measure gives an upper bound on additive error when optimizing a (normalized) linear function over $\mathbf{P}$ and $\mathbf{P}^k$.*

Observation 2 (Comparing $\mathbf{d(P, P^k)}$ to $\sqrt{n}$) *Notice that the largest distance between any two points in the $[0,\ 1]^n$ hypercube is at most $\sqrt{n}$. Therefore in the rest of the paper we will compare the value of $d(\mathbf{P}, \mathbf{P}^k)$ to $\sqrt{n}$.*

1.3 Some Examples

In order to build some intuition we begin with some examples in this section. Let $\mathbf{P} := \{x \in [0,\ 1]^n : ax \leq b\}$ where a is a non-negative vector. It is straightforward to verify that in this case, $\mathbf{P}^k := \{x \in [0,\ 1]^n : a^I x \leq b\ \forall I \in \binom{[n]}{k}\}$, where $a^I_j := a_j$ if $j \in I$ and $a^I_j = 0$ otherwise.

Example 1: Consider the simplex $\mathbf{P} = \{x \in [0,1]^n : \sum_{i=1}^n x_i \leq 1\}$. Using the above observation, we have that $\mathbf{P}^k = \text{conv}\{e^1, e^2, \ldots, e^n, \frac{1}{k}e\}$, where e^j is the unit vector in the direction of the j^{th} coordinate and e is the all ones vector. Therefore the distance measure between $\mathbf{P}$ and $\mathbf{P}^k$ is $\sqrt{n}(\frac{1}{k} - \frac{1}{n}) \approx \frac{\sqrt{n}}{k}$, attained by the points $\frac{1}{n}e \in \mathbf{P}$ and $\frac{1}{k}e \in \mathbf{P}^k$. This is quite nice because with $k \approx \sqrt{n}$ (which is pretty reasonably sparse) we get a constant distance. Observe also that the *rate of change of the distance measure* follows a 'single pattern' - we call this a *single phase example*. See Figure 1(a) for $\text{d}(\mathbf{P}, \mathbf{P}^k)$ plotted against k (in blue) and $k \cdot \text{d}(\mathbf{P}, \mathbf{P}^k)$ plotted against k (in green).

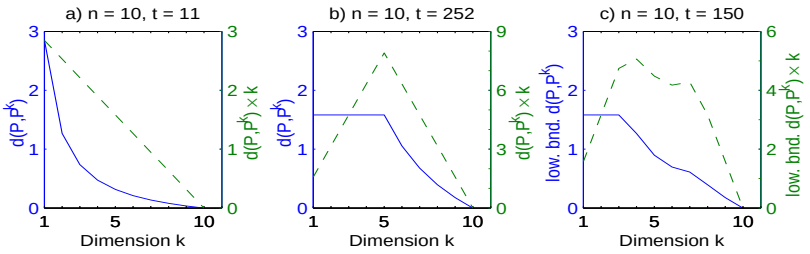

Fig. 1. (a) Sparsity is good. (b) Sparsity is not so good. (c) Example with three phases.

Example 2: Consider the set $\mathbf{P} = \{x \in [0,1]^n : \sum_i x_i \leq \frac{n}{2}\}$. We have that $\mathbf{P}^k := \{x \in [0,\ 1]^n : \sum_{i \in I} x_i \leq \frac{n}{2},\ \forall I \in \binom{[n]}{k}\}$. Therefore, for all $k \in \{1, \ldots, n/2\}$ we have $\mathbf{P}^k = [0,\ 1]^n$ and hence $\text{d}(\mathbf{P}, \mathbf{P}^k) = \sqrt{n}/2$. Thus, we stay with distance $\Omega(\sqrt{n})$ (the worst possible for polytopes in $[0,1]^n$) even with $\Theta(n)$ sparsity. Also observe that for $k > \frac{n}{2}$, we have $\text{d}(\mathbf{P}, \mathbf{P}^k) = \frac{n\sqrt{n}}{2k} - \frac{\sqrt{n}}{2}$. Clearly the rate of change of the distance measure has *two phases*, first phase of k between 1 and $\frac{n}{2}$ and the second phase of k between $\frac{n}{2}$ and n. See Figure 1(b) for the plot of $\text{d}(\mathbf{P}, \mathbf{P}^k)$ against k (in blue) and of $k \cdot \text{d}(\mathbf{P}, \mathbf{P}^k)$ against k (in green).

Example 3: We present an experimental example in dimension $n = 10$. The polytope $\mathbf{P}$ is now set as the convex hull of 150 binary points randomly selected from the hyperplane $\{x \in \mathbb{R}^{10} : \sum_{i=1}^{10} x_i = 5\}$. We experimentally computed lower bounds on $\mathrm{d}(\mathbf{P}, \mathbf{P}^k)$ which are plotted in Figure 1(c) as the blue line (details appear in the full version of the paper). Notice that there are now three phases, which are more discernible in the plot between the lower bound on $k \cdot \mathrm{d}(\mathbf{P}, \mathbf{P}^k)$ and k (in green).

The above examples serve to illustrate the fact that different polytopes, behave very differently when we try and approximate them using sparse inequalities. We note here that in all our additional experiments, albeit in small dimensions, we have usually found at most three phases as in the previous examples.

2 Main Results

2.1 Upper Bounds

Surprisingly, it appears that the complicated behavior of $\mathrm{d}(\mathbf{P}, \mathbf{P}^k)$ as k changes can be described to some extent in closed form. Our first result is nontrivial upper bounds on $\mathrm{d}(\mathbf{P}, \mathbf{P}^k)$ for general polytopes. This is proven in Section 3.

Theorem 3 (Upper Bound on $\mathrm{d}(\mathbf{P}, \mathbf{P}^k)$). *Let $n \geq 2$. Let $\mathbf{P} \subseteq [0,1]^n$ be the convex hull of points $\{p^1, \ldots, p^t\}$. Then*

1. $d(\mathbf{P}, \mathbf{P}^k) \leq 4 \max\left\{ \frac{n^{1/4}}{\sqrt{k}} \sqrt{8 \max_{i \in [t]} \|p^i\| \sqrt{\log 4tn}}, \frac{8\sqrt{n}}{3k} \log 4tn \right\}$
2. $d(\mathbf{P}, \mathbf{P}^k) \leq 2\sqrt{n}\left(\frac{n}{k} - 1\right)$.

Since $\max_{i \in \{1,\ldots,t\}} ||p^i|| \leq \sqrt{n}$ and the first upper bound yields nontrivial values only when $k \geq 8 \log 4tn$, a simpler (although weaker) expression for the first upper bound is $4\frac{\sqrt{n}}{\sqrt{k}}\sqrt{\log 4tn}$. We make two observations based on Theorem 3.

Consider polytopes with 'few' vertices, say n^q vertices for some constant q. Suppose we decide to use cutting-planes with half sparsity (i.e. $k = \frac{n}{2}$), a reasonable assumption in practice. Then plugging in these values, it is easily verified that $\mathrm{d}(\mathbf{P}, \mathbf{P}^k) \leq 4\sqrt{2}\sqrt{(q+1)\log n} \approx c\sqrt{\log n}$ for a constant c, which is a significantly small quantity in comparison to $\sqrt{n}$. In other words, *if the number of vertices is small, independent of the location of the vertices, using half sparsity cutting-planes allows us to approximate the integer hull very well.* We believe that as the number of vertices increase, the structure of the polytope becomes more important in determining $\mathrm{d}(\mathbf{P}, \mathbf{P}^k)$ and Theorem 3 only captures the worst-case scenario. Overall, Theorem 3 presents a theoretical justification for the use of sparse cutting-planes in many cases.

Theorem 3 supports the existence of three phases in the behavior of $\mathrm{d}(\mathbf{P}, \mathbf{P}^k)$ as k varies: **(Small k)** When $k \leq 16 \log 4tn$ the (simplified) upper bounds are larger than $\sqrt{n}$, indicating that 'no progress' is made in approximating the shape of $\mathbf{P}$ (this is seen Examples 2 and 3). **(Medium k)** When $16 \log 4tn \leq k \lesssim n - \sqrt{n \log 4tn}$ the first upper bound in Theorem 3 dominates. **(Large k)** When $k \gtrsim n - \sqrt{n \log 4tn}$ the upper bound $2\sqrt{n}\left(\frac{n}{k} - 1\right)$ dominates. In particular, in this phase, $k \cdot \mathrm{d}(\mathbf{P}, \mathbf{P}^k) \leq 2n^{3/2} - 2\sqrt{n}k$, i.e., the upper bound times k is a linear function of k. All the examples in Section 1 illustrate this behaviour.

2.2 Lower Bounds

How good is the quality of the upper bound presented in Theorem 3? Let us first consider the second upper bound in Theorem 3. Then observe that for the second example in Section 1, this upper bound is tight up to a constant factor for k between the values of $\frac{n}{2}$ and n.

We study lower bounds on $\mathrm{d}(\mathbf{P}, \mathbf{P}^k)$ for random polytopes in Section 4 that show that the first upper bound in Theorem 3 is also quite tight.

Theorem 4. *Let $X^1, X^2, \ldots, X^t$ be independent uniformly random points in $\{0,1\}^n$, and let $\mathbf{P} = conv(X^1, X^2, \ldots, X^t)$. Then for t and k satisfying $(2k^2 \log n + 2)^2 \leq t \leq e^n$ we have with probability at least $1/4$*

$$d(\mathbf{P}, \mathbf{P}^k) \geq \min\left\{\frac{\sqrt{n}}{\sqrt{k}}\frac{\sqrt{\log(t/2)}}{78\sqrt{\log n}}, \frac{\sqrt{n}}{8}\right\}\left(\frac{1}{2} - \frac{1}{k^{3/2}}\right) - 3\sqrt{\log t}.$$

Let us compare this lower bound with the simpler expression $4\frac{\sqrt{n}}{\sqrt{k}}\sqrt{\log tn}$ for the first part of the upper bound of Theorem 3. We focus on the case where the minimum in the lower bound is acheived by the first term. Then comparing the leading term $\sqrt{\frac{n}{k}}\frac{\sqrt{\log t}}{2\cdot 78\sqrt{\log n}}$ in the lower bound with the upper bound, we see that these quantities match up to a factor of $624\frac{\sqrt{\log(tn)}\sqrt{\log n}}{\sqrt{\log t}}$, showing that for many 0/1 polytopes the first upper bound of Theorem 3 is quite tight. We also remark that the in order to simplify the exposition we did not try to optimize constants and lower order terms in our bounds.

The main technical tool for proving this lower bound is a new anticoncentration result for linear combinations aX, where the X_i's are independent Bernoulli random variables. The main difference from standard anticoncentration results is that the latter focus on variation around the standard deviation; in this case, standard tools such as the Berry-Esseen Theorem or the Paley-Zygmund Inequality can be used to obtain constant-probability anticoncentration. However, we need to control the behavior of aX much further from its standard deviation, where we cannot hope to get constant-probability anticoncentration.

Lemma 1. *Let $X_1, X_2, \ldots, X_n$ be independent random variables with X_i taking value 0 with probability 1/2 and value 1 with probability 1/2. Then for every $a \in [-1,1]^n$ and $\alpha \in [0, \frac{\sqrt{n}}{8}]$,*

$$\Pr\left(aX \geq \mathbb{E}[aX] + \frac{\alpha}{2\sqrt{n}}\left(1 - \frac{1}{n^2}\right)\|a\|_1 - \frac{1}{2n^2}\right) \geq \left(e^{-50\alpha^2} - e^{-100\alpha^2}\right)^{60\log n}.$$

2.3 Hard Packing Integer Programs

We also study well-known, randomly generated, hard packing integer program instances (see for instance [5]). Given parameters $n, m, M \in \mathbb{N}$, the convex hull of the packing IP is given by $\mathbf{P} = \mathrm{conv}(\{x \in \{0,1\}^n : A^j x \leq \frac{\sum_i A^j_i}{2},\ \forall j \in [m]\})$,

where the A_i^j's are chosen independently and uniformly in the set $\{0, 1, \ldots, M\}$. Let (n, m, M)-PIP denote the distribution over the generated $\mathbf{P}$'s.

The following result shows the limitation of sparse cuts for these instances.

Theorem 5. *Consider $n, m, M \in \mathbb{N}$ such that $n \geq 50$ and $8 \log 8n \leq m \leq n$. Let $\mathbf{P}$ be sampled from the distribution (n, m, M)-PIP. Then with probability at least $1/2$, $d(\mathbf{P}, \mathbf{P}^k) \geq \frac{\sqrt{n}}{2}\left(\frac{2}{\max\{\alpha,1\}}(1-\epsilon)^2 - (1+\epsilon')\right)$, where $c = k/n$ and*

$$\frac{1}{\alpha} = \frac{M}{2(M+1)}\left\lfloor \frac{n - 2\sqrt{n \log 8m}}{c((2-c)n+1) + 2\sqrt{10cnm}} \right\rfloor, \quad \epsilon = \frac{24\sqrt{\log 4n^2 m}}{\sqrt{n}}, \quad \epsilon' = \frac{3\sqrt{\log 8n}}{\sqrt{m} - 2\sqrt{\log 8n}}.$$

Notice that when m is sufficiently large, and n reasonably larger than m, we have ϵ and ϵ' approximately 0, and the above bound reduces to approximately $\frac{\sqrt{n}}{2}\left(\left(\frac{M}{M+1}\right)\left(\frac{n}{k(2-n/k)}\right) - 1\right) \approx \frac{\sqrt{n}}{2}\left(\frac{n}{k(2-n/k)} - 1\right)$, which is within a constant factor of the upper bound from Theorem 3. The poor behavior of sparse cuts gives an indication for the hardness of these instances and suggests that denser cuts should be explored in this case.

One interesting feature of this result is that it works directly with the IP formulation, not relying on an explicit linear description of the convex hull.

2.4 Sparse Cutting-Planes and Extended Formulations

Let $\text{proj}_x : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^n$ denote the projection operator onto the first n coordinates. We say that a set $\mathbf{Q} \subseteq \mathbb{R}^n \times \mathbb{R}^m$ is an *extended formulation* of $\mathbf{P} \subseteq \mathbb{R}^n$ if $\mathbf{P} = \text{proj}_x(\mathbf{Q})$.

As our final result we remark that using sparse cutting-planes in extended formulations is at least as good as using them in the original polyhedron, and sometime much better; proofs are provided in the full version of the paper.

Lemma 2. *Consider a polyhedron $\mathbf{P} \subseteq \mathbb{R}^n$ and an extended formulation $\mathbf{Q} \subseteq \mathbb{R}^n \times \mathbb{R}^m$ for it. Then $\text{proj}_x(\mathbf{Q}^k) \subseteq (\text{proj}_x(\mathbf{Q}))^k = \mathbf{P}^k$.*

Lemma 3. *Consider $n \in \mathbb{N}$ and assume it is a power of 2. Then there is a polytope $\mathbf{P} \subseteq \mathbb{R}^n$ such that: 1) $d(\mathbf{P}, \mathbf{P}^k) = \sqrt{n/2}$ for all $k \leq n/2$; 2) there is an extended formulation $\mathbf{Q} \subseteq \mathbb{R}^n \times \mathbb{R}^{2n-1}$ of $\mathbf{P}$ such that $\text{proj}_x(\mathbf{Q}^3) = \mathbf{P}$.*

3 Upper Bound

In this section we prove Theorem 3. In fact we prove the same bound for polytopes in $[-1, 1]^n$, which is a slightly stronger result. The following well-known property is crucial for the constructions used in both parts of the theorem.

Observation 6 (Section 2.5.1 of [3]) *Consider a compact convex set $S \subseteq \mathbb{R}^n$. Let $\bar{x}$ be a point outside S and let $\bar{y}$ be the closest point to $\bar{x}$ in S. Then setting $a = \bar{x} - \bar{y}$, the inequality $ax \leq a\bar{y}$ is valid for S and cuts $\bar{x}$ off.*

3.1 Proof of First Part of Theorem 3

Consider a polytope $\mathbf{P} = \text{conv}\{p^1, p^2, \ldots, p^t\}$ in $[-1,1]^n$. Define

$$\lambda^* = \max\left\{\frac{n^{1/4}}{\sqrt{k}}\sqrt{8\max_i \|p^i\|\sqrt{\log 4tn}}, \frac{8\sqrt{n}}{3k}\log 4tn\right\}.$$

In order to show that $\mathrm{d}(\mathbf{P}, \mathbf{P}^k)$ is at most $4\lambda^*$ we show that every point at distance more than $4\lambda^*$ from $\mathbf{P}$ is cut off by a valid inequality for $\mathbf{P}^k$. Assume until the end of this section that $4\lambda^*$ is at most $\sqrt{n}$, otherwise the result is trivial; in particular, this implies that the second term in the definition of λ^* is at most $\sqrt{n}/4$ and hence $k \geq 8\log 4tn$.

So let $u \in \mathbb{R}^n$ be a point at distance more than $4\lambda^*$ from $\mathbf{P}$. Let $v \in \mathbf{P}$ be the closest point in $\mathbf{P}$ to $\mathbf{P}^k$. We can write $u = v + \lambda d$ for some vector d with $\|d\|_2 = 1$ and $\lambda > 4\lambda^*$. From Observation 6, inequality $dx \leq dv$ is valid for $\mathbf{P}$, so in particular $dp^i \leq dv$ for all $i \in [t]$; in addition, it that this inequality cuts off u: $du = dv + \lambda > dv$. The idea is to use this extra slack factor λ in the previous equation to show we can 'sparsify' the inequality $dx \leq dv$ while maintaining separation of $\mathbf{P}$ and u. It then suffices to prove the following lemma.

Lemma 4. *There is a k-sparse vector $\tilde{d} \in \mathbb{R}^n$ such that $\tilde{d}p^i \leq \tilde{d}v + \frac{\lambda}{2}$ for all $i \in [t]$, and $\tilde{d}u > \tilde{d}v + \frac{\lambda}{2}$.*

To prove the lemma we construct a random vector $\tilde{D} \in \mathbb{R}^n$ which, with non-zero probability, is k-sparse and satisfies the two other requirements of the lemma. Let $\alpha = \frac{k}{2\sqrt{n}}$. Define $\tilde{D}$ as the random vector with independent coordinates, where $\tilde{D}_i$ is defined as follows: if $\alpha|d_i| \geq 1$, then $\tilde{D}_i = d_i$ with probability 1; if $\alpha|d_i| < 1$, then $\tilde{D}_i$ takes value $\text{sign}(d_i)/\alpha$ with probability $\alpha|d_i|$ and takes value 0 with probability $1 - \alpha|d_i|$. (For convenience we define $\text{sign}(0) = 1$.)

The next observation follows directly from the definition of $\tilde{D}$.

Observation 7 *For every vector $a \in \mathbb{R}^n$ the following hold:*
1. $\mathbb{E}[\tilde{D}a] = da$
2. $Var(\tilde{D}a) \leq \frac{1}{\alpha}\sum_{i\in[n]} a_i^2|d_i|$
3. $|\tilde{D}_i a_i - \mathbb{E}[\tilde{D}_i a_i]| \leq \frac{\lceil a_i \rceil}{\alpha}$.

Claim. With probability at least $1 - 1/4n$, $\tilde{D}$ is k-sparse.

Proof. Construct the vector $a \in \mathbb{R}^n$ as follows: if $\alpha|d_i| \geq 1$ then $a_i = 1/d_i$, and if $\alpha|d_i| < 1$ then $a_i = \alpha/\text{sign}(d_i)$. Notice that $\tilde{D}a$ equals the number of non-zero coordinates of $\tilde{D}$ and $\mathbb{E}[\tilde{D}a] \leq \alpha\|d\|_1 \leq k/2$. Also, from Observation 7 we have

$$\text{Var}(\tilde{D}a) \leq \frac{1}{\alpha}\sum_{i\in[n]} a_i^2|d_i| \leq \alpha\|d\|_1 \leq \frac{k}{2}.$$

Then using Bernstein's inequality ([6], Appendix A.2) we obtain

$$\Pr(\tilde{D}a \geq k) \leq \exp\left(-\min\left\{\frac{k^2}{8k}, \frac{3k}{8}\right\}\right) \leq \frac{1}{4n},$$

where the last inequality uses our assumption that $k \geq 8\log 4tn$. □

We now show that property 1 required by Lemma 4 holds for $\tilde{D}$ with high probability.

Claim. $\Pr(\max_{i\in[t]}[\tilde{D}(p^i - v) - d(p^i - v)] > 2\lambda^*) \leq 1/4n$.

Proof. Define the centered random variable $Z = \tilde{D} - d$. To make the analysis cleaner, notice that $\max_{i\in[t]} Z(p^i - v) \leq 2\max_{i\in[t]} |Zp^i|$; this is because $\max_{i\in[t]} Z(p^i - v) \leq \max_{i\in[t]} |Zp^i| + |Zv|$, and because for all $a \in \mathbb{R}^n$ we have $|av| \leq \max_{p\in\mathbf{P}} |ap| = \max_{i\in[t]} |ap^i|$ (since $v \in \mathbf{P}$).

Therefore our goal is to upper bound the probability that the process $\max_{i\in[t]} |Zp^i|$ is larger then λ^*. Fix $i \in [t]$. By Bernstein's inequality,

$$\Pr(|Zp^i| \geq \lambda^*) \leq \exp\left(-\min\left\{\frac{(\lambda^*)^2}{4\mathrm{Var}(|Zp^i|)}, \frac{3\lambda^*}{4M}\right\}\right), \tag{1}$$

where M is an upper bound on $\max_j |Z_j p^i_j|$.

To bound the terms in the right-hand side, from Obersvation 7 we have

$$\mathrm{Var}(Zp^i) = \mathrm{Var}(\tilde{D}p^i) \leq \frac{1}{\alpha}\sum_j (p^i_j)^2|d_j| \leq \frac{1}{\alpha}\sum_j p^i_j|d_j| \leq \frac{1}{\alpha}\|p^i\|\|d\| = \frac{1}{\alpha}\|p^i\|,$$

where the second inequality follows from the fact $p^i \in [0,1]^n$, and the third inequality follows from the Cauchy-Schwarz inequality. Moreover, it is not difficulty to see that for every random variable W, $\mathrm{Var}(|W|) \leq \mathrm{Var}(W)$. Using the first term in the definition of λ^*, we then have

$$\frac{(\lambda^*)^2}{\mathrm{Var}(|Zp^i|)} \geq 4\log 4tn.$$

In addition, for every coordinate j we have $|Z_j p^i_j| = |\tilde{D}_j p^i_j - \mathbb{E}[\tilde{D}_j p^i_j]| \leq 1/\alpha$, where the inequality follows from Observation 7. Then we can set $M = 1/\alpha$ and using the second term in the definition of λ^* we get $\frac{\lambda^*}{M} \geq \frac{4}{3}\log 4tn$. Therefore, replacing these bounds in inequality (1) gives $\Pr(|Zp^i| \geq \lambda^*) \leq \frac{1}{4tn}$.

Taking a union bound over all $i \in [t]$ gives that $\Pr(\max_{i\in[t]} |Zp^i| \geq \lambda^*) \leq 1/4n$. This concludes the proof of the claim. □

Claim. $\Pr(\tilde{D}(u - v) \leq \lambda/2) \leq 1 - 1/(2n-1)$.

Proof. Recall $u - v = \lambda d$, hence it is equivalent to bound $\Pr(\tilde{D}d \leq 1/2)$. First, $\mathbb{E}[\tilde{D}d] = dd = 1$. Also, from Observation 7 we have $\tilde{D}d \leq |\tilde{D}d - dd| + |dd| \leq \frac{1}{\alpha}\sum_i |d_i| + 1 \leq \frac{2n}{k} + 1 \leq n$, where the last inequality uses the assumption $k \geq 8\log 4tn$. Then employing Markov's inequality to the non-negative random variable $n - \tilde{D}d$, we get $\Pr(\tilde{D}d \leq 1/2) \leq 1 - \frac{1}{2n-1}$. This concludes the proof. □

Proof of Lemma 4. Employ the previous three claims and union bound to find a realization of $\tilde{D}$ that is k-sparse and satisfies requirements 1 and 2 of the lemma.

This concludes the proof of the first part of Theorem 3.

Observation 8 *Notice that in the above proof λ^* is set by Claim 3.1, and need to be essentially $\mathbb{E}[\max_{i\in[t]}(\tilde{D}-d)p^i]$. There is a vast literature on bounds on the supremum of stochastic processes [6], and improved bounds for structured $\mathbf{P}$'s are possible (for instance, via the* generic chaining *method).*

3.2 Proof of Second Part of Theorem 3

The main tool for proving this upper bound is the following lemma, which shows that when $\mathbf{P}$ is 'simple', and we have a stronger control over the distance of a point $\bar{x}$ to $\mathbf{P}$, then there is a k-sparse inequality that cuts $\bar{x}$ off.

Lemma 5. *Consider a hyperplane $H = \{x \in \mathbb{R}^n : ax \leq b\}$ and let $\mathbf{P} = H \cap [-1,1]^n$. Let $\bar{x} \in [-1,1]^n$ be such that $d(\bar{x}, H) > 2\sqrt{n}(\frac{n}{k} - 1)$. Then $\bar{x} \notin \mathbf{P}^k$.*

Proof. Assume without loss of generality that $\|a\|_2 = 1$. Let $\bar{y}$ be the point in H closes to $\bar{x}$, and notice that $\bar{x} = \bar{y} + \lambda a$ where $\lambda > \sqrt{n}(\frac{n}{k} - 1)$.

For any set $I \in \binom{[n]}{k}$, the inequality $\sum_{i\in I} a_i x_i \leq b + \sum_{i\notin I: a_i \geq 0} a_i - \sum_{i\notin I: a_i<0} a_i$ is valid for $\mathbf{P}$; since it is k-sparse, it is also valid for $\mathbf{P}^k$. Averaging out this inequality over all $I \in \binom{[n]}{k}$, we get that the following is valid for $\mathbf{P}^k$:

$$\frac{k}{n} ax \leq b + \left(1 - \frac{k}{n}\right)\left(\sum_{i:a_i\geq 0} a_i - \sum_{i:a_i<0} a_i\right) \equiv ax \leq b + \left(\frac{n}{k} - 1\right)(b + \|a\|_1).$$

We claim that $\bar{x}$ violates this inequality. First notice that $a\bar{x} = a\bar{y} + \lambda = b + \lambda > b + 2\sqrt{n}\left(\frac{n}{k} - 1\right)$, hence it suffices to show $b + \|a\|_1 \leq 2\sqrt{n}$. Our assumption on $\bar{x}$ implies that $\mathbf{P} \neq [-1,1]^n$, and hence $b < \max_{x\in[-1,1]} ax = \|a\|_1$; this gives $b + \|a\|_1 \leq 2\|a\|_1 \leq 2\sqrt{n}\|a\|_2 = 2\sqrt{n}$, thus concluding the proof. □

To prove the second part of Theorem 3 consider a point $\bar{x}$ of distance greater than $2\sqrt{n}(\frac{n}{k} - 1)$ from $\mathbf{P}$; we show $\bar{x} \notin \mathbf{P}^k$. Let $\bar{y}$ be the closest point to $\bar{x}$ in $\mathbf{P}$. Let $a = \bar{x} - \bar{y}$. From Observation 6 we have that $ax \leq a\bar{y}$ is valid for $\mathbf{P}$. Define $H' = \{x \in \mathbb{R}^n : ax \leq a\bar{y}\}$ and $\mathbf{P}' = H' \cap [-1,1]^n$. Notice that $\mathrm{d}(\bar{x}, H') = \mathrm{d}(\bar{x}, \bar{y}) > 2\sqrt{n}(\frac{n}{k} - 1)$. Then Lemma 5 guarantees that $\bar{x}$ does not belong to $\mathbf{P}'^k$. But $\mathbf{P} \subseteq \mathbf{P}'$, so by monotonicity of the k-sparse closure we have $\mathbf{P}^k \subseteq \mathbf{P}'^k$; this shows that $\bar{x} \notin \mathbf{P}^k$, thus concluding the proof.

4 Lower Bound

In this section we prove Theorem 4. The proof is based on the 'bad' polytope of Example 2 and proceeds in two steps. First, for a random 0/1 polytope $\mathbf{P}$ we show that with good probability the facets $dx \leq d_0$ for $\mathbf{P}^k$ have d_0 being large, namely $d_0 \gtrsim \left(\frac{1}{2} + \frac{\sqrt{\log t}}{\sqrt{k}}\right)\sum_i d_i$; therefore, with good probability the point $\bar{p} \approx (\frac{1}{2} + \frac{\sqrt{\log t}}{\sqrt{k}})e$ belongs to $\mathbf{P}^k$. In the second step, we show that with good

probability the distance from $\bar{p}$ to $\mathbf{P}$ is at least $\approx \sqrt{\frac{n}{k}}\sqrt{\log t}$, by showing that the inequality $\sum_i x_i \lesssim \frac{n}{2} + \sqrt{n}$ is valid for $\mathbf{P}$.

We now proceed with the proof. Consider the random set $\mathbf{X} = \{X^1, X^2, \ldots, X^t\}$ where the X^i's are independent uniform random points in $\{0,1\}^n$, and define the random 0/1 polytope $\mathbf{P} = \text{conv}(\mathbf{X})$.

We say that a 0/1 polytope $\mathbf{P}$ is *α-tough* if for every facet $dx \leq d_0$ of $\mathbf{P}^k$ we have $d_0 \geq \frac{\sum_i d_i}{2} + \frac{\alpha}{2\sqrt{k}}(1 - \frac{1}{k^2})\|d\|_1 - \|d\|_\infty/2k^2$. To get a handle on α-toughness of random 0/1 polytopes, define $\mathcal{D}$ as the set of all integral vectors $\ell \in \mathbb{R}^n$ that are k-sparse and satisfy $\|\ell\|_\infty \leq (k+1)^{(k+1)/2}$. The following claim, shows that all the facets of $\mathbf{P}^k$ come from the set $\mathcal{D}$; it follows directly from applying Corollary 26 in [8] to each term $\mathbf{P} + \mathbb{R}^{\bar{I}}$ in the definition of $\mathbf{P}^k$ from Section 1.1.

Lemma 6. *Let $\mathbf{Q} \subseteq \mathbb{R}^n$ be a 0/1 polytope. Then there is a subset $\mathcal{D}' \subseteq \mathcal{D}$ such that $\mathbf{Q}^k = \{x : dx \leq \max_{y \in \mathbf{P}^k} dy,\ d \in \mathcal{D}'\}$.*

Now we can analyze the probability that $\mathbf{P}$ is α-tough.

Lemma 7. *If $1 \leq \alpha^2 \leq \min\left\{\frac{\log(t/2)}{6000 \log n}, \frac{k}{64}\right\}$ and $k \leq n-1$, then $\mathbf{P}$ is α-tough with probability at least $1/2$.*

Proof. Let E be the event that for all $d \in \mathcal{D}$ we have $\max_{i \in [t]} dX^i \geq \frac{1}{2}\sum_j d_j + \frac{\alpha}{2\sqrt{k}}(1 - \frac{1}{k^2})\|d\|_1 - \|d\|_\infty/2k^2$. Because of Lemma 6, whenever E holds we have that $\mathbf{P}$ is α-tough and thus it suffices to show $\Pr(E) \geq 1/2$.

Fix $d \in \mathcal{D}$. Since d is k-sparse, we can apply Lemma 1 to $d/\|d\|_\infty$ restricted to the coordinates in its support to obtain that

$$\Pr\left(dX^i \geq \frac{\sum_i d_i}{2} + \frac{\alpha}{2\sqrt{k}}\left(1 - \frac{1}{k^2}\right)\|d\|_1 - \frac{\|d\|_\infty}{2k^2}\right) \geq \left(e^{-50\alpha^2} - e^{-100\alpha^2}\right)^{60 \log n}$$
$$\geq e^{-100\alpha^2 \cdot 60 \log n} \geq \frac{1}{t^{1/2}},$$

where the second inequality follows from our lower bound on α and the last inequality follows from our upper bound on α. By independence of the X^i's,

$$\Pr\left(\max_{i \in [t]} dX^i < \frac{\sum_i d_i}{2} + \frac{\alpha}{2\sqrt{k}}\left(1 - \frac{1}{k^2}\right)\|d\|_1 - \frac{\|d\|_\infty}{2k^2}\right) \leq \left(1 - \frac{1}{t^{1/2}}\right)^t \leq e^{-t^{1/2}},$$

where the second inequality follows from the fact that $(1-x) \leq e^{-x}$ for all x.

Finally notice that $|\mathcal{D}| = \binom{n}{k}(k+1)^{(k+1)^2/2}$ and that, by our assumption on the size of t and $k \leq n-1$, $e^{-t^{1/2}} \leq (1/2)|\mathcal{D}|$. Therefore, taking a union bound over all $d \in \mathcal{D}$ of the previous displayed inequality gives $\Pr(E) \geq 1/2$, concluding the proof of the lemma. □

The next lemma takes care of the second step of the argument; its simple proof is based on Bernstein's and is deferred to the full version of the paper.

Lemma 8. *With probability at least $3/4$, the inequality $\sum_j x_j \leq \frac{n}{2} + 3\sqrt{n \log t}$ is valid for $\mathbf{P}$.*

Lemma 9. *Suppose that the polytope* $\mathbf{Q}$ *is* α*-tough for* $\alpha \geq 1$ *and that the inequality* $\sum_i x_i \leq \frac{n}{2} + 3\sqrt{n \log t}$ *is valid for* $\mathbf{Q}$*. Then we have* $d(\mathbf{Q}, \mathbf{Q}^k) \geq \sqrt{n}\left(\frac{\alpha}{2\sqrt{k}} - \frac{\alpha}{k^2} - \frac{3\sqrt{\log t}}{\sqrt{n}}\right)$.

Proof. We first show that the point $\bar{q} = (\frac{1}{2} + \frac{\alpha}{2\sqrt{k}} - \frac{\alpha}{k^2})e$ belongs to $\mathbf{P}$. Let $dx \leq d_0$ be facet for $\mathbf{P}$. Then we have

$$\begin{aligned} d\bar{q} &= \frac{\sum_i d_i}{2} + \alpha\left(\frac{1}{2\sqrt{k}} - \frac{1}{k^2}\right)\sum_i d_i \leq \frac{\sum_i d_i}{2} + \alpha\left(\frac{1}{2\sqrt{k}} - \frac{1}{k^2}\right)\|d\|_1 \\ &\leq \frac{\sum_i d_i}{2} + \alpha\left(\frac{1}{2\sqrt{k}} - \frac{1}{2k^2}\right)\|d\|_1 - \frac{\|d\|_\infty}{2k^2}, \end{aligned}$$

where the first inequality uses the fact that $\frac{1}{2\sqrt{k}} - \frac{1}{k^2} \geq 0$ and the second inequality uses $\alpha \geq 1$ and $\|d\|_1 \geq \|d\|_\infty$. Since $\mathbf{Q}$ is α-tough it follows that $\bar{q}$ satisfies $dx \leq d_0$; since this holds for all facets of $\mathbf{Q}$, we have $\bar{q} \in \mathbf{Q}$.

Now define the halfspace $H = \{x : \sum_i x_i \leq \frac{n}{2} + 3\sqrt{n \log t}\}$. By assumption $\mathbf{Q} \subseteq H$, and hence $\mathrm{d}(\mathbf{Q}, \mathbf{Q}^k) \geq \mathrm{d}(H, \mathbf{Q}^k)$. But it is easy to see that the point in H closest to $\bar{q}$ is the point $\tilde{q} = (\frac{1}{2} + \frac{3\sqrt{\log t}}{\sqrt{n}})e$. This gives that $\mathrm{d}(\mathbf{Q}, \mathbf{Q}^k) \geq \mathrm{d}(H, \mathbf{Q}^k) \geq \mathrm{d}(\bar{q}, \tilde{q}) \geq \sqrt{n}\left(\frac{\alpha}{2\sqrt{k}} - \frac{\alpha}{k^2} - \frac{3\sqrt{\log t}}{\sqrt{n}}\right)$. This concludes the proof. □

We now conclude the proof of Theorem 4. Set $\bar{\alpha}^2 = \min\left\{\frac{\log(t/2)}{6000 \log n}, \frac{k}{64}\right\}$. Taking union bound over Lemmas 7 and 8, with probability at least 1/4 $\mathbf{P}$ is $\bar{\alpha}$-tough and the inequality inequality $\sum_i x_i \leq \frac{n}{2} + 3\sqrt{n \log t}$ is valid for it. Then from Lemma 9 we get that with probability at least 1/4, $\mathrm{d}(\mathbf{P}, \mathbf{P}^k) \geq \sqrt{n}\left(\frac{\bar{\alpha}}{2\sqrt{k}} - \frac{\bar{\alpha}}{k^2} - \frac{3\sqrt{\log t}}{\sqrt{n}}\right)$, and the result follows by plugging in the value of $\bar{\alpha}$.

5 Hard Packing Integer Programs

In this section we prove Theorem 5; missing proof are presented in the full version of the paper. With overload in notation, we use $\binom{[n]}{k}$ to denote the set of vectors in $\{0,1\}^n$ with exactly k 1's.

Let $\mathbf{P}$ be a random polytope sampled from the distribution (n, m, M)-PIP and consider the corresponding random vectors A^j's. The idea of the proof is to show that with constant probability $\mathbf{P}$ behaves like Example 2, by showing that the cut $\sum_i x_i \lesssim \frac{n}{2}$ is valid for it and that it approximately contains 0/1 points with many 1's.

We start with a couple of lemmas that are proved via Bernstein's inequality.

Lemma 10. *With probability at least* $1 - \frac{1}{8}$ *we have* $|\sum_{i=1}^n A_i^j - \frac{nM}{2}| \leq M\sqrt{n \log 8m}$ *for all* $j \in [m]$.

Lemma 11. *With probability at least* $1 - \frac{1}{4}$ *the cut* $(1 - \frac{2\sqrt{\log 8n}}{\sqrt{m}})\sum_i x_i \leq \frac{n}{2} + \frac{\sqrt{n \log 8}}{\sqrt{m}}$ *is valid for* $\mathbf{P}$.

The next lemma shows that with constant probability $\mathbf{P}$ almost contains *all* 0/1 points with many 1's.

Lemma 12. *With probability at least* $1 - \frac{1}{8}$ *we have*

$$A^j \bar{x} \leq \frac{(M+1)c(2n - cn + 1)}{2} + (M+1)\sqrt{10cnm}, \qquad \forall j \in [m], \forall \bar{x} \in \binom{[n]}{cn}.$$

Lemma 13. *Consider a 0/1 polytope* $\mathbf{Q} = conv(\{x \in \{0,1\}^n : a^j x \leq b_j,\ j = 1, 2, \ldots, m\})$ *where* $n \geq 20$, $m \leq n$, $a_i^j \in [0, M]$ *for all* i, j, *and* $b_j \geq \frac{nM}{12}$ *for all* i. *Consider* $1 < \alpha \leq 2\sqrt{n}$ *and let* $\bar{x} \in \{0,1\}^n$ *be such that for all* j, $a^j \bar{x} \leq \alpha b_j$. *Then the point* $\frac{1}{\alpha}(1-\epsilon)^2 \bar{x}$ *belongs to* $\mathbf{Q}$ *as long as* $\frac{12\sqrt{\log 4n^2 m}}{\sqrt{n}} \leq \epsilon \leq \frac{1}{2}$.

Proof of Theorem 5. Recall the definitions of $\alpha, \epsilon, \epsilon'$, and $c = k/n$ from the statement of the theorem. Let E be the event that Lemmas 10, 11 and 12 hold; notice that $\Pr(E) \geq 1/2$. For the rest of the proof we fix a $\mathbf{P}$ (and the associated A^j's) where E holds and prove a lower bound on $\mathrm{d}(\mathbf{P}, \mathbf{P}^k)$.

Consider a set $I \in \binom{[n]}{cn}$ and let $\bar{x}$ be the incidence vector of I (i.e. $\bar{x}_i = 1$ if $i \in I$ and $\bar{x}_i = 0$ if $i \notin I$). Since the bounds from Lemmas 10 and 12 hold for our $\mathbf{P}$, straightforward calculations show that $A^j \bar{x} \leq \alpha \frac{1}{2} \sum_i A_i^j$ for all $j \in [m]$. Therefore, from Lemma 13 we have that the point $\frac{1}{\max\{\alpha,1\}}(1-\epsilon)^2 \bar{x}$ belongs to $\mathbf{P}$. This means that the point $\tilde{x} = \frac{1}{\max\{\alpha,1\}}(1-\epsilon)^2 e$ belongs to $\mathbf{P} + \mathbb{R}^{\bar{I}}$ (see Section 1.1). Since this holds for every $I \in \binom{[n]}{cn}$, we have $\tilde{x} \in \mathbf{P}^k$.

Let $\tilde{y}$ be the point in $\mathbf{P}$ closest to $\tilde{x}$. Let $a = (1 - \frac{2\sqrt{\log 8n}}{\sqrt{m}})$ and $b = \frac{n}{2} + \sqrt{n \log 8m}$, so that the cut in Lemma 11 is given by $aex \leq b$. From Cauchy-Schwarz we have that $\mathrm{d}(\tilde{x}, \tilde{y}) \geq \frac{ae\tilde{x} - ae\tilde{y}}{\|ae\|} = \frac{e\tilde{x}}{\sqrt{n}} - \frac{ae\tilde{y}}{a\sqrt{n}}$.

By definition of $\tilde{x}$ we have $e\tilde{x} = \frac{1}{\max\{\alpha,1\}}(1-\epsilon)^2 n$. From the fact the cut $aex \leq b$ is valid for $\mathbf{P}$ and $\tilde{y} \in \mathbf{P}$, we have $ae\tilde{y} \leq b$. Simple calculations show that $\frac{b}{a\sqrt{n}} \leq \frac{n}{2}(1+\epsilon')$. Plugging these values in we get that $\mathrm{d}(\mathbf{P}, \mathbf{P}^k) = \mathrm{d}(\tilde{x}, \tilde{y}) \geq \frac{\sqrt{n}}{2}\left(\frac{2(1-\epsilon)^2}{\max\{\alpha,1\}} - (1+\epsilon')\right)$. Theorem 5 follows from the definition of α, ϵ and ϵ'.

References

1. Tobias Achterberg. Personal communication.
2. Kent Andersen and Robert Weismantel. Zero-coefficient cuts. In *IPCO*, 2010.
3. Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.
4. Zonghao Gu. Personal communication.
5. Konstantinos Kaparis and Adam N. Letchford. Separation algorithms for 0-1 knapsack polytopes. *Mathematical Programming*, 124(1-2):69–91, 2010.
6. Vladimir Koltchinskii. *Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems*. Springer-Verlag, 2011.
7. Amar Narisetty. Personal communication.
8. Günter M. Ziegler. *Lectures on Polytopes*. Springer, 1995.