Math. Stat. Learn. 5 (2022), 273–310
DOI 10.4171/MSL/35

The smoothed complexity of Frank–Wolfe methods via conditioning of random matrices and polytopes

Luis Rademacher and Chang Shu

Abstract. Frank–Wolfe methods are popular for optimization over a polytope. One of the reasons is because they do not need projection onto the polytope but only linear optimization over it. To understand its complexity, a fruitful approach in many works has been the use of condition measures of polytopes. Lacoste-Julien and Jaggi introduced a condition number for polytopes and showed linear convergence for several variations of the method. The actual running time can still be exponential in the worst case (when the condition number is exponential). We study the smoothed complexity of the condition number, namely the condition number of small random perturbations of the input polytope and show that it is polynomial for any simplex and exponential for general polytopes. Our results also apply to other condition measures of polytopes that have been proposed for the analysis of Frank–Wolfe methods: vertex-facet distance (Beck and Shtern) and facial distance (Peña and Rodríguez).

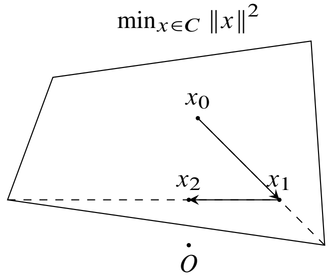
Our argument for polytopes is a refinement of an argument that we develop to study the conditioning of random matrices. The basic argument shows that for $c > 1$ a d-by-n random Gaussian matrix with $n \geq cd$ has a d-by-d submatrix with minimum singular value that is exponentially small with high probability. This also has consequences on known results about the robust uniqueness of tensor decompositions, the complexity of the simplex method and the diameter of polytopes.

1. Introduction

Frank–Wolfe methods (FWMs) [23] are a family of algorithms that attempt to minimize a differentiable function over a convex set. For concreteness we start by describing the basic Frank–Wolfe method to minimize a differentiable function $f: C \mapsto \mathbb{R}$, where $C \subseteq \mathbb{R}^d$ is a compact convex set. It is an iterative method and proceeds as follows:

2020 Mathematics Subject Classification. Primary 68W40; Secondary 60B20, 60D05.
Keywords. Smoothed analysis, Frank–Wolfe methods, random matrix, random polytope, condition number.

Let $x_0 \in C$.
for $k = 0, \dots, K$ **do**
Compute $y \in \operatorname{argmin}_{x \in C} (\nabla f(x_k))^T x$.
Let $x_{k+1} = x_k + \alpha^*(y - x_k)$, where α^* is a suitable step size.
end for



Some of our results are about Wolfe's method [46], which is a variation of Frank–Wolfe methods specialized to the minimum norm point problem in a polytope (that is, a bounded convex polyhedron).

1.1. Our contributions and related work

In this paper we are interested in the complexity of FWMs. The time complexity of Wolfe's method is know to be exponential in the worst case (by an upper bound in [46] and a lower bound in [19]). There is a large body of work proving linear convergence of several variations of FWMs [8, 24, 25, 30, 31, 34–36]. We are particularly interested in [8, 30, 31, 35, 36] which prove *global* linear convergence of certain variations of FWMs: F-W with away steps, pairwise F-W and Wolfe's method when the feasible region is a polytope $C = \operatorname{conv}(A)$ for finite $A \subseteq \mathbb{R}^d$. In these results the upper bound on the running time (actual speed of linear convergence) depends on a condition number[1] of C. Informally speaking, the dependence is of the following kind: if x_t is the current point after t iterations, then the function value satisfies

$$f(x_t) - f^* \le (1 - \kappa)^t (f(x_0) - f^*),$$

where f^* is the optimal value, x_0 is the initial point and $0 \le \kappa \le 1$ is a measure of conditioning. If κ is small, then convergence is slow. In the previously mentioned papers, κ is of the form "something"/ $\operatorname{diam}(C)$, where "something" can be:

- [31] minimum width, $\operatorname{minwidth}(A) = \min_{S \subseteq A} \operatorname{width}(S)$ (width is standard, see Section 2.8.1);

[1] Informally, in this paper we use the term *condition measure* to denote any number that partially describes the conditioning of an object. We reserve the term *condition number* for condition measures were larger values denote worse conditioning. For example, $\sigma_{\max}/\sigma_{\min}$ is a condition number for a matrix, while $\sigma_{\min}$ is a condition measure. This is so that we can describe numbers such as $\sigma_{\min}$ where smaller values mean worse conditioning without contradicting the usual meaning of condition number.

- [31] pyramidal width, $\operatorname{PWidth}(A)$ (essentially the same as $\Phi(C)$, see discussion below);
- [8] vertex-facet distance, $\operatorname{vf}(C) = \min_{F \in \operatorname{facets}(C)} d(\operatorname{aff} F, \operatorname{vertices}(C) \setminus F)$; or
- [35] facial distance, $\Phi(C) = \min_{\substack{F \in \operatorname{faces}(C) \\ \emptyset \subsetneq F \subsetneq C}} d(F, \operatorname{conv}(\operatorname{vertices}(C) \setminus F))$.

We do not provide a definition of pyramidal width at this point as it is complicated and it was shown in [35] that $\operatorname{PWidth}(A) = \Phi(C)$ (Theorem 2.21 here). It is also known that $\operatorname{minwidth}(A) \leq \operatorname{PWidth}(A)$ (see [31, Section 3.1]). We start with the observation that $\Phi(C) \leq \operatorname{vf}(C)$ (Theorem 2.22). (Note that the reverse inequality was claimed in [35], but the cube $[0, 1]^d$ is a counterexample: $\Phi([0, 1]^d) = 1/\sqrt{d}$, while $\operatorname{vf}([0, 1]^d) = 1$.) This implies that all four quantities lie between $\operatorname{minwidth}(A)$ and $\operatorname{vf}(C)$ (Theorem 2.22). It follows from [19] that all of them can be exponentially small as a function of the bit-length of A. In fact, a stronger result follows from the work of Alon and Vu [2] combined with the stated inequalities. Alon and Vu showed that there is a 0/1-simplex S such that $\operatorname{vf}(S)$ is sub-exponentially small in the dimension (Corollary 3.3) The connection between polytope conditioning for FWMs and the Alon and Vu result was observed in [31].

The main contributions of this paper are about the smoothed analysis of FWMs and the condition measures of matrices and polytopes. Smoothed analysis [42] is an approach to understand the behavior of algorithms that are efficient in practice but are inefficient in the worst case. The main idea is to study small random perturbations of any given instance of a problem. Suppose that the instance is described by a vector $x \in \mathbb{R}^n$. Then one aims to understand $T(x + g)$, where $g \in \mathbb{R}^n$ is a random vector with distribution $N(0, \sigma^2 \mathbf{I}_n)$ and T is a measure of complexity (for example, $T(x)$ could be the running time of a particular algorithm on input x). We adopt a definition that first appeared in [9, 10].

Definition 1.1 ([39], [40, Section 1.1]). We say T has *(probabilistic) polynomial smoothed complexity* if there is a polynomial p such that

$$\max_{x \in \mathbb{R}^n, \|x\| \leq 1} \mathbb{P}_g\big(T(x + g) \geq p(n, 1/\sigma, 1/\delta)\big) \leq \delta.$$

Note that having probabilistic polynomial smoothed complexity does not imply that the expected running time is polynomially bounded, but this definition is more robust with respect to changes in the machine model (see [40, 43] for a discussion).

Our first smoothed analysis result concerns FWMs minimizing a convex function on a simplex (Section 3). We show that minwidth has good smoothed complexity (Lemma 3.6). This implies the following result on polytope conditioning that can be combined with results in [31] to show polynomial smoothed time complexity of several FWMs for the minimization of a convex function in any simplex:

Theorem 1.2. *Let $A = \{A_1, \ldots, A_{d+1}\}$ be a set of independent Gaussian random vectors with means μ_i, $\|\mu_i\| \leq 1$, $i \in [d+1]$, and covariance matrix $\sigma^2 \mathbf{I}_d$. Then for $\delta > 0$, with probability at least $1 - \delta$, the measure of conditioning $\kappa = \frac{\mathrm{PWidth}(A)}{\mathrm{diam}(A)}$ of A is at least some inverse polynomial in d, $1/\sigma$ and $1/\delta$.*

Note that even the problem of finding the minimum norm point in a simplex is not known to have a simple polynomial time algorithm. All polynomial time algorithms we know for such a special case are general purpose convex programming algorithms such as the ellipsoid method. Moreover, [19] shows that the linear programming problem reduces in strongly polynomial time to the minimum norm point in a simplex problem. This suggests that to find a simple polynomial time algorithm for the minimum norm point in a simplex is hard and, in particular, to find a strongly polynomial time algorithm would imply the existence of a strongly polynomial time algorithm for linear programming, which would solve a major open problem.

Our second smoothed analysis result concerns condition measures of general polytopes (Section 7). We show that the standard global linear convergence results for FWMs mentioned above based on polytope conditioning cannot guarantee polynomial complexity for general polytopes in the average or smoothed sense. More specifically, for V-polytopes $\mathrm{conv}(A)$ with $|A|$ and d large and comparable, $d \approx \delta|A|$, $\delta \in (0, 1)$, we show that vertex-facet distance does not have polynomial smoothed complexity. Given that the complexity here increases as $\mathrm{vf}(A)$ gets smaller, in the context of Definition 1.1 one sets $T = 1/\mathrm{vf}$. It is enough to take $x = 0$ there and we show:

Theorem 1.3. *Let $\delta \in (0, 1)$. Suppose $A = \{A_1, \ldots, A_{n+1}\}$ is a set of iid. standard Gaussian random vectors in $\mathbb{R}^d$ and $d = \lfloor \delta n \rfloor$. Let $P_{n+1} = \mathrm{conv}(A_1, \ldots, A_{n+1})$. Then*

$$\mathbb{P}\left(\mathrm{diam}(P_{n+1}) \geq \sqrt{d}\right) \geq 1 - e^{-\frac{nd}{32}},$$

and there exist constants $0 < c < 1$ and $0 < c' < 1$ (that depend only on δ) such that,

$$\lim_{n \to \infty} \mathbb{P}\left(\mathrm{vf}(P_{n+1}) \leq c^d\right) \geq c'.$$

Hence, the measure of conditioning $\kappa = \frac{\mathrm{vf}(P_{n+1})}{\mathrm{diam}(P_{n+1})}$ of A is exponentially small in d with constant probability.

Theorem 1.3 combined with Theorem 2.22 implies that none of the four measures of polytope conditioning (minwidth, PWidth, Φ, vf) have polynomial smoothed complexity.

A way of interpreting Theorem 1.3 is that the standard conditioning measures of polytopes for FWMs are somewhat pessimistic and can appear ill-conditioned even when the polytope is bad only locally. For example, vertex-facet distance can be small even if one vertex and one facet are bad while the rest of the polytope is good. In other words, it may still be possible to show smoothed polynomial complexity of FWMs in a different way.

Theorem 1.3 is a statement about the minimum distance between the affine hull of d points that form a facet and a vertex not on that facet. In order to understand this problem we study first a simplified version where we replace affine hull by span and we remove the restriction that the $d-1$ points form a facet. Namely, we study the following question: given n standard Gaussian random points in $\mathbb{R}^d$, how close can one of the points be to the span of some $d-1$ others when n is somewhat larger than d, say, $n = 2d$? This question is easier to understand than the polytope version and it relates to conditioning of random matrices and the restricted isometry property in compressive sensing. The relation starts from the known observation (Lemma 2.6) that the minimum point-hyperplane distance is, up to polynomial factors, the same as the smallest singular value of a matrix. Given this, our question is essentially equivalent to: given a d-by-n random matrix with iid. standard Gaussian entries, what is the minimum of the smallest singular values over d-by-d submatrices? We answer this question by showing that when $n/d \geq c > 1$ the minimum smallest singular value above (and, equivalently, minimum point-hyperplane distance) is exponentially small:

Theorem 1.4. *Let A be a d-by-n random matrix with iid. standard Gaussian entries with $d \geq 2$ and $\frac{n}{d} \geq c_0 > 1$. Then, there exist constants $c_2, c_4 > 1$, $0 < c_6 < 1$ (that depend only on c_0) such that with probability at least $1 - 2c_4 c_6^d$,*

$$\min_{S \subseteq [n], |S| = d} \sigma_d(A_S) \leq \frac{1}{c_4 c_2^{d-1}}.$$

Theorem 1.5. *Let A be a d-by-n random matrix with iid. standard Gaussian entries with $d \geq 2$ and $1 < \frac{n}{d-1} \leq C_0$. Then, there exist constants $C_1 > 1$, $0 < C_2 < 1$ (that depend only on C_0) such that with probability at least $1 - nC_2^{d-1}$,*

$$\min_{S \subseteq [n], |S| = d} \sigma_d(A_S) \geq \frac{1}{C_1^{d-1}}.$$

While Theorems 1.4 and 1.5 are new as far as we know, there is a large body of work, partly motivated by compressive sensing, that studies questions related to them. In that area one is generally interested in showing that all d-by-k submatrices of A are well-conditioned, say, σ_1/σ_k is no more than a constant (the *restricted isometry property* of Candès and Tao [15, 16]). This can only happen when k is much smaller than d, a regime very different from our case, $k = d$. The standard analyses in compressive sensing as well as recent results such as [14] do not seem to be able to clarify the behavior in our regime. This is because Theorem 1.4 informally shows that some submatrix is ill-conditioned, the reverse of what one wants in compressed sensing.

The idea of the proof of Theorem 1.4 (Section 4) is the following: Consider the case $n = 2d$ for concreteness and aim to show that with constant probability one point is exponentially close to the span of $d-1$ others. Let $\mathcal{S}$ be the family of sets of $d-1$ columns of A. For $S \in \mathcal{S}$, let $\mathcal{B}_S$ be the set of points in $\mathbb{R}^d$ within distance ε of span S. Let $V = \bigcup_{S \in \mathcal{S}} \mathcal{B}_S$. It is enough to show that for $\varepsilon = 1/c^d$, $c > 1$, the Gaussian volume $\mathcal{G}(V)$ is at least a constant. We do this by lower bounding it using the *first two terms of the inclusion-exclusion principle (Bonferroni inequality)*:

$$\mathcal{G}(V) \geq \sum_S \mathcal{G}(\mathcal{B}_S) - \frac{1}{2} \sum_{S,T:S \neq T} \mathcal{G}(\mathcal{B}_S \cap \mathcal{B}_T).$$

Note that $\mathcal{B}_S \cap \mathcal{B}_T$ can be large if S and T share many columns. To deal with this difficulty, replace $\mathcal{S}$ above with a large subfamily $\mathcal{T} \subseteq \mathcal{S}$ of subsets of columns where each pair of subsets has few columns in common by picking separated subsets *greedily (Gilbert–Varshamov bound)*. See [38], [28, Lemma 19.3] for another instance of Bonferroni's inequality with almost pairwise independence.

Our aim with Theorem 1.5 is to provide a matching lower bound for Theorem 1.4 for completeness. While it may be possible to deduce it from a union bound and estimates for the smallest singular value of a single matrix (without taking submatrices) in [21] and [41, Section 3], our proof is self-contained and follows from a union bound and elementary estimates.

While Theorems 1.4 and 1.5 are results about random matrices, they have direct implications in the analysis of algorithms: In Section 5 we discuss how Theorem 1.4 conditions the applicability of the robustness of tensor decomposition result by Bhaskara, Charikar and Vijayaraghavan [7]. In Section 6 we discuss how Theorem 1.4 conditions the applicability of results about the complexity of the simplex method and the diameter of polytopes in [12, 13, 18, 22].

2. Preliminaries

2.1. Notation

For $v \in \mathbb{R}^d$ and $i \in [d]$, let v_{-i} denote vector v with coordinate v_i removed, that is

$$v_{-i} := (v_1, \ldots, v_{i-1}, v_{i+1}, \ldots, v_d).$$

If $v \neq 0$, let $\hat{v} := v/\|v\|_2$. Let $B(x, \varepsilon) := \{y \in \mathbb{R}^d : \|y - x\|_2 \leq \varepsilon\}$. Let $\mathcal{S}^{d-1}$ denote the $(d-1)$-dimensional unit sphere in $\mathbb{R}^d$. For $v \in \mathcal{S}^{d-1}$, denote the spherical cap centered at v with angle α as $\mathcal{C}_\alpha(v) := \{x \in \mathcal{S}^{d-1} : v \cdot x \geq \cos \alpha\}$. For $A \subseteq \mathbb{R}^d$, let aff A be the affine hull of A, and define

$$A_\varepsilon := \{x \in \mathbb{R}^d : \operatorname{dist}(x, A) \leq \varepsilon\}, \quad A_{-\varepsilon} := \{x \in \mathbb{R}^d : B(x, \varepsilon) \subset A\}.$$

Let $\mathcal{N}(\mu, \sigma^2)$ denote the normal distribution with mean μ and standard deviation σ. We write $X \sim \mathcal{N}(\mu, \sigma^2)$ if X is normally distributed with mean μ and standard deviation σ. This notion also generalizes to the multivariate normal distribution with the first argument as mean vector and the second as covariance matrix. Let $\mathcal{G}$ denote the standard multivariate Gaussian probability measure. For random variables or distributions X, Y, notation $X \overset{d}{=} Y$ states that X and Y have the same distribution.

2.2. Comparison inequality for the Gaussian distribution

We need the following known comparison inequality for the Gaussian distribution. It is a special case of Anderson's lemma [3].

Lemma 2.1. *Let $\mu \in \mathbb{R}$. Let $X_i \sim \mathcal{N}(0, \sigma^2)$, $Y_i \sim \mathcal{N}(\mu, \sigma^2)$, $i \in [k]$, be independent. Let $t \geq 0$. Then*

$$\mathbb{P}\Big(\sum_{i=1}^{k} X_i^2 \geq t\Big) \leq \mathbb{P}\Big(\sum_{i=1}^{k} Y_i^2 \geq t\Big).$$

In the proof of Lemma 7.4, we will need the following comparison inequality, which follows from Lemma 2.1.

Lemma 2.2. *Let $\mu \in \mathbb{R}$. Let $X_0, X_i, Y_0 \sim \mathcal{N}(0, \sigma^2)$, $Y_i \sim \mathcal{N}(\mu, \sigma^2)$, $i \in [n]$ and be independent. Then for any $t \in (0, 1)$, we have*

$$\mathbb{P}\Big(\frac{Y_0^2}{Y_0^2 + \sum_{i=1}^n Y_i^2} \geq t^2\Big) \leq \mathbb{P}\Big(\frac{X_0^2}{X_0^2 + \sum_{i=1}^n X_i^2} \geq t^2\Big).$$

Proof. Let f denote the probability density function of X_0^2 and Y_0^2. By the law of total expectation,

$$\begin{aligned}
\mathbb{P}\Big(\frac{Y_0^2}{Y_0^2 + \sum_{i=1}^n Y_i^2} \geq t^2\Big) &= \int_0^\infty \mathbb{P}\Big(\frac{y}{y + \sum_{i=1}^n Y_i^2} \geq t^2\Big) f(y)\,\mathrm{d}y \\
&= \int_0^\infty \mathbb{P}\Big(\sum_{i=1}^n Y_i^2 \leq (1/t^2 - 1)y\Big) f(y)\,\mathrm{d}y \\
&\leq \int_0^\infty \mathbb{P}\Big(\sum_{i=1}^n X_i^2 \leq (1/t^2 - 1)x\Big) f(x)\,\mathrm{d}x \quad \text{(Lemma 2.1)} \\
&= \int_0^\infty \mathbb{P}\Big(\frac{x}{x + \sum_{i=1}^n X_i^2} \geq t^2\Big) f(x)\,\mathrm{d}x \\
&= \mathbb{P}\Big(\frac{X_0^2}{X_0^2 + \sum_{i=1}^n X_i^2} \geq t^2\Big).
\end{aligned}$$

■

2.3. Concentration and tail inequalities

Lemma 2.3 ([32]). *Let $(X_1, \dots, X_n)$ be iid. standard Gaussian variables. Let $\alpha_1, \dots, \alpha_n$ be non-negative. Let $Z = \sum_{i=1}^n \alpha_i (X_i^2 - 1)$. Then, the following inequalities hold for any positive t:*

$$\mathbb{P}\big(Z \geq 2\|\alpha\|_2 \sqrt{t} + 2\|\alpha\|_\infty t\big) \leq \exp(-t),$$
$$\mathbb{P}\big(Z \leq -2\|\alpha\|_2 \sqrt{t}\big) \leq \exp(-t).$$

2.4. Gilbert–Varshamov bound

We need the following well-known bound on the number of binary vectors satisfying a minimum distance condition.

Lemma 2.4. *Let $A(n,t,w)$ be the maximum number of binary n-vectors with exactly w ones and pairwise Hamming distance greater than or equal to t. Then for any $c_0 > 1$, there exist constants $c_1 > 0$ and $c_2 > 1$ (that depend only on c_0) such that for all $d \geq 1$ and $n/d \geq c_0$ we have $A(n, c_1 d, d) \geq c_2^d$.*

2.5. Generalization of Archimedes' formula

Lemma 2.5. *Let $d \geq 3$. Let U be a uniformly random d-dimensional unit vector. Then $(U_1, \dots, U_{d-2})$ is uniform in B^{d-2} and $\mathbb{P}(\|(U_1, \dots, U_{d-2})\| \leq t) = t^{d-2}$.*

Proof. The first part is well known, a proof can be found in [5, Corollary 4]. The second part follows immediately from the first part. ∎

2.6. One-off-distance vs sigma min

Lemma 2.6 (see e.g. [6, Lemma 3.5] for a proof). *If $A \in \mathbb{R}^{m \times n}$ has columns $a_1, \dots, a_n$ and $m \geq n$, then denoting $a_{-i} = \operatorname{span}(a_j : j \neq i)$, we have*

$$\frac{1}{\sqrt{n}} \min_{i \in [n]} \operatorname{dist}(a_i, a_{-i}) \leq \sigma_n(A) \leq \min_{i \in [n]} \operatorname{dist}(a_i, a_{-i}).$$

2.7. Facts about Gaussian random polytopes

2.7.1. Gaussian ε-neighborhood.

Corollary 2.7. *Let Q be a convex set in $\mathbb{R}^d$. Then there exists an absolute constant $c > 0$ such that $\mathcal{G}(Q \setminus Q_{-\varepsilon}) \leq c\varepsilon d^{1/4}$.*

Proof. The proof follows immediately from [17, Lemma A.2] and the fact $\|I\|_{HS} = \sqrt{d}$ (Hilbert–Schmidt norm). Their proof is based on [4, 33]. ∎

2.7.2. Distances of facets.

Lemma 2.8 ([45, Theorem 4.4.5]). *Let X be an $m \times n$ random matrix whose entries are iid. standard Gaussian random variables. Then for $t > 0$, we have*

$$\mathbb{P}\big(\sigma_{\max}(X) > c(\sqrt{m} + \sqrt{n} + t)\big) \le 2e^{-t^2},$$

where c is some absolute positive constant.

Lemma 2.9. *Let $X_1, \dots, X_n$ be iid. standard Gaussian random vectors in $\mathbb{R}^d$. For $S \subseteq [n]$, $|S| = d$, define V_S as the shortest vector in $\operatorname{aff}(X_S)$. Then there exists a constant $c > 0$ such that*

$$\mathbb{P}\big(\max_{S \subseteq [n], |S| = d} \|V_S\| \le c(2 + \sqrt{n/d})\big) \ge 1 - 2e^{-d}.$$

Proof. Let X be the matrix whose column vectors are $X_1, \dots, X_n$. For any $S \subseteq [n], |S| = d$, X_S is linearly independent with probability 1. Using that the norm of the average of the columns of X_S is at least the norm of V_S,

$$\|V_S\| \le \Big\| \frac{1}{d} \sum_{i \in S} X_i \Big\| = \frac{1}{d} \|X_S \mathbb{1}\| \le \frac{\sigma_{\max}(X_S)}{\sqrt{d}} \le \frac{\sigma_{\max}(X)}{\sqrt{d}}. \tag{1}$$

From Lemma 2.8 we know $\mathbb{P}(\sigma_{\max}(X) > c(\sqrt{d} + \sqrt{n} + t)) \le 2e^{-t^2}$. The claim follows by letting $t = \sqrt{d}$ and applying (1). ∎

Note that Lemma 2.9 directly generalizes to Gaussian random vectors with mean zero and covariance matrix $\sigma^2 \mathbf{I}_d$ by scaling by σ.

2.7.3. Number of facets. We will need the fact that the number of facets of the convex hull of n Gaussian random points in $\mathbb{R}^d$ is exponential in d with high probability when $n = cd$, $c > 1$. We could not find such a result in the literature and we do not see how to deduce it from results on the asymptotic number of facets in stochastic geometry [1, 11, 26, 27, 37] (the difficulties are: either they only determine the expectation or variance of the number of facets, or the bounds are as n goes to infinity for fixed d). Nevertheless, it is easy to deduce what we want from the work of Donoho and Tanner on compressive sensing and the neighborliness of random polytopes. We build on top of basic polytope theory from [47].

Definition 2.10 (Neighborliness). A polytope P is k-neighborly if every subset of k vertices forms a $(k-1)$-face.

Let $f_l(P)$ denote the number of l-faces of polytope P.

Theorem 2.11 ([20, Corollary 1.1, Lemma 3.2]). *There exists a function (threshold) $\rho(\delta)\colon (0,1) \to \mathbb{R}$, $\rho(\delta) > 0$ with the following property: Let $\delta \in (0,1)$. Let $d = \lfloor \delta n \rfloor$.*

Let $\rho < \rho(\delta)$. Let $X_1, \ldots, X_n$ be iid. samples from a Gaussian distribution in $\mathbb{R}^d$ with non-singular covariance. Let $P = \operatorname{conv}\{X_1, \ldots, X_n\}$. Then

$$\lim_{n\to\infty} \mathbb{P}\big(f_1(P) = n \text{ and } P \text{ is } \lfloor \rho d \rfloor\text{-neighborly}\big) = 1.$$

The above theorem demonstrates, given its assumptions, that when n is large enough, P has $\binom{n}{\lfloor \rho d \rfloor}$ many $\lfloor \rho d \rfloor$-faces with high probability. Note also that P is simplicial (every facet is a simplex) a.s. Thus, a.s. each facet of P provides at most $\binom{d}{\lfloor \rho d \rfloor}$ many $\lfloor \rho d \rfloor$-faces, and the number of facets is at least

$$\frac{\binom{n}{\lfloor \rho d \rfloor}}{\binom{d}{\lfloor \rho d \rfloor}} \geq \left(\frac{n}{d}\right)^{\lfloor \rho d \rfloor} \geq \left(\frac{1}{\delta}\right)^{\lfloor \rho d \rfloor} \geq c^d,$$

for some $c > 1$ (and d large enough). We conclude:

Corollary 2.12. *Let $\delta \in (0,1)$. Let P be the convex hull of n iid. standard Gaussian random points in $\mathbb{R}^d$, $d = \lfloor \delta n \rfloor$. Then there exists a constant $c > 1$ (that depends only on δ) such that $\lim_{n\to\infty} \mathbb{P}(f_d(P) \geq c^d) = 1$.*

Corollary 2.12 can probably also be proven directly from different but related neighborliness results by Vershik and Sporyshev [44], [20, Theorem 2].

2.8. Condition measures of polytopes

2.8.1. Width and minwidth.

Definition 2.13 (Directional width and width). The *directional width* of a set $A \subseteq \mathbb{R}^d$ with respect to a direction $r \in \mathbb{R}^d$ is defined as $\operatorname{dirW}(A, r) := \sup_{s,v\in\mathcal{A}} \langle \frac{r}{\|r\|}, s - v\rangle$. The *width* of A, denoted $\operatorname{width}(A)$ is the infimum of the directional width over all directions on its affine hull.

Definition 2.14 (Minwidth, [31, Section 3.1]). The minwidth of a finite set $A \subseteq \mathbb{R}^d$, denoted $\operatorname{minwidth}(A)$, is the minimum width over all subsets of A.

2.8.2. Pyramidal width.

Definition 2.15 (Pyramidal directional width, [31]). We define the *pyramidal directional width* of a finite set $A \subseteq \mathbb{R}^d$ with respect to a direction $r \in \mathbb{R}^d$ and a base point $x \in \operatorname{conv}(A)$ to be

$$\operatorname{PDirW}(A, r, x) := \min_{S\in S_x} \operatorname{dirW}(S \cup \{s(A, r)\}, r) = \min_{S\in S_x} \max_{s\in A, v\in S} \left\langle \frac{r}{\|r\|}, s - v\right\rangle,$$

where $S_x := \{T \subseteq A : x \text{ is a proper convex combination of all the elements in } T\}$ and $s(A, r) := \operatorname{argmax}_{v\in A}\langle r, v\rangle$.

Definition 2.16 (Feasible direction, [31]). A direction r is feasible for A from x if it points inwards $\mathrm{conv}(A)$, i.e. $r \in \mathrm{cone}(A - x)$. A direction r is feasible for A if it is feasible for A from some $x \in A$.

Definition 2.17 (Pyramidal width, [31]). We define the *pyramidal width* of a finite set $A \subseteq \mathbb{R}^d$ to be the smallest pyramidal directional width of all its faces,

$$\mathrm{PWidth}(A) := \min_{\substack{K \in \mathrm{faces}(\mathrm{conv}(A)) \\ x \in K \\ r \in \mathrm{cone}(K - x) \setminus \{0\}}} \mathrm{PDirW}(K \cap A, r, x).$$

2.8.3. Vertex-facet distance. The *vertex-facet distance* polytope conditioning parameter for the analysis of FWMs was introduced in [8]. We adopt here the slightly specialized definition in [35], which is defined as a property of a polytope independent of the representation, while the original version in [8] can depend on the numbers used to represent a polytope.

Definition 2.18 (Vertex-facet distance, [8, 35]). Let $P \subseteq \mathbb{R}^d$ be a polytope with $\dim(\mathrm{aff}(P)) \geq 1$. The vertex-facet distance of P is

$$\mathrm{vf}(P) := \min_{F \in \mathrm{facets}(P)} \mathrm{dist}\big(\mathrm{aff}(F), \mathrm{vertices}(P) \setminus F\big).$$

2.8.4. Relation between vertex-facet distance and pyramidal width. We show $\mathrm{vf}(\mathrm{conv}(A)) \geq \mathrm{PWidth}(A)$. It seems that this result may have already been known to [35, comment before Theorem 1, combined with Theorem 2], but it is claimed there in the wrong direction. That direction is impossible as the example of a unit cube shows: $\mathrm{PWidth}([0,1]^d) = 1/\sqrt{d}$ (see [31, Lemma 4]), but $\mathrm{vf}([0,1]^d) = 1$.

Proposition 2.19. *Let $A \subseteq \mathbb{R}^d$ be a finite set with at least two points. Then*

$$\mathrm{vf}\big(\mathrm{conv}(A)\big) \geq \mathrm{PWidth}(A).$$

Proof. Let $P = \mathrm{conv}(A)$. Let F be a facet of P and pick $v \in \mathrm{vertices}(P) \setminus F$ so that

$$\mathrm{dist}\big(v, \mathrm{aff}(F)\big) = \varepsilon := \mathrm{vf}(P).$$

Pick $x \in \mathrm{relint}(\mathrm{conv}(F \cup \{v\}))$ and let r be the unit outer normal vector to F (in $\mathrm{aff}(P)$ if P is not full-dimensional). We set $K = P$ as in Definition 2.17 so that $r \in \mathrm{cone}(K - x) = \mathrm{aff}(P)$ and

$$\mathrm{PWidth}(A) \leq \mathrm{PDirW}(K \cap A, r, x) = \mathrm{PDirW}(A, r, x).$$

Now, set $S = A \cap (F \cup \{v\})$ as in Definition 2.15 so that, with these choices,

$$\mathrm{PDirW}(A, r, x) \leq \mathrm{dirW}(S, r) \leq \varepsilon.$$

The claim follows. ∎

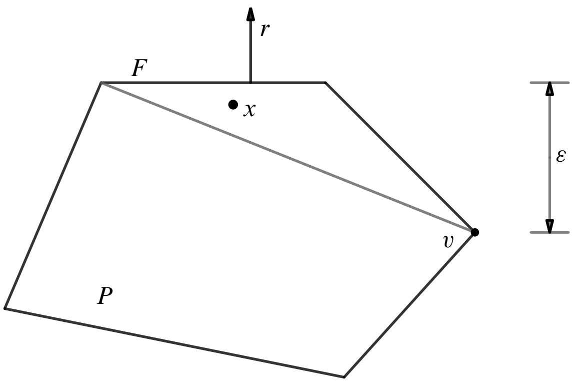


Figure 1. Proof of Proposition 2.19.

2.8.5. Facial distance.

Definition 2.20 ([35]). Let $C \subseteq \mathbb{R}^d$ be a polytope with $\dim(\operatorname{aff}(C)) \geq 1$. The *facial distance* of C is

$$\Phi(C) := \min_{\substack{F \in \operatorname{faces}(C) \\ \emptyset \subsetneq F \subsetneq C}} d\big(F, \operatorname{conv}(\operatorname{vertices}(C) \setminus F)\big).$$

2.8.6. Relation between facial distance and pyramidal width. One of the motivations of [35] to introduce parameter Φ is that it is the same as PWidth (except in degenerate cases) while the definition of Φ is simpler to use in many cases. We quote their result next.

Theorem 2.21 ([35, Theorem 2]). *Let $A \subseteq \mathbb{R}^d$ be a finite set with at least two points. Then*

$$\Phi\big(\operatorname{conv}(A)\big) = \operatorname{PWidth}(A).$$

2.8.7. Summary result.

Theorem 2.22. *Let $A \subseteq \mathbb{R}^d$ be a finite set with at least two points. Then*

$$\operatorname{minwidth}(A) \leq \Phi\big(\operatorname{conv}(A)\big) = \operatorname{PWidth}(A) \leq \operatorname{vf}\big(\operatorname{conv}(A)\big).$$

Proof. Immediate from [31, Section 3.1], Theorem 2.21 and Proposition 2.19. ∎

3. Conditioning of simplices

In this section we show that the smoothed conditioning of any simplex is polynomial. This implies that several FWMs have smoothed polynomial complexities on the minimum norm point in a simplex problem and the minimization of many convex functions on a simplex. To put this result in context, we first argue (based on known results) that even a simplex with vertices having 0/1-coordinates can have bad conditioning. Another relevant context to keep in mind is the fact that linear programming reduces in strongly polynomial time to the minimum norm point in a simplex [19].

3.1. Equality of width and minwidth of a simplex

We start with the observation that the minwidth of a simplex is the same as its width.

Lemma 3.1. *Let A be the vertex set of a simplex in $\mathbb{R}^d$ and $A_0 \subset A$ which includes more than one vertex. Then* $\operatorname{width}(A) \leq \operatorname{width}(A_0)$. *In particular,* $\operatorname{minwidth}(A) = \operatorname{width}(A)$.

Proof. We prove by induction in d. The width of a polytope is the minimum distance between parallel supporting hyperplanes in its affine hull. Width of a 2-simplex is the minimum height of triangle, which is smaller than the length of any edge. For a k-simplex A, suppose the width of one of its facet is given by the distance between two parallel $(k-2)$-dimensional planes, p_1^{k-2} and p_2^{k-2}. One can extend p_1^{k-2} and p_2^{k-2} to parallel hyperplanes in $\mathbb{R}^k$ that enclose A. Suppose extensions p_1^{k-1} and p_2^{k-1} give the minimum distance. Then,

$$\begin{aligned}\operatorname{dist}\big(p_1^{k-1}, p_2^{k-1}\big) &= \min_{a \in p_1^{k-1}, b \in p_2^{k-1}} \|a-b\| \\ &\leq \min_{a \in p_1^{k-2}, b \in p_2^{k-2}} \|a-b\| = \operatorname{dist}\big(p_1^{k-2}, p_2^{k-2}\big),\end{aligned}$$

which shows that the width of a k-simplex is less than the width of any of its facets. The claim then follows by induction. ∎

3.2. Bad worst case conditioning of a 0/1-simplex

Lacoste-Julien and Jaggi [31] observed that the minwidth of the unit cube in $\mathbb{R}^d$ is exponentially small in d. This example was one of their motivations for introducing PWidth, which is $1/\sqrt{d}$ for the cube. Their observation is based on the following result by Alon and Vu:

Theorem 3.2 ([2, Theorem 3.2.2], [48, Corollary 27]). *There are $d+1$ vectors in $\{0,1\}^d$ that form the vertices of a d-dimensional simplex S so that*

$$\frac{2^{d-1}}{d^{d/2}} \leq \operatorname{vf}(S) \leq \frac{2^{d(2+o(1))}}{d^{d/2}}.$$

The authors of [19] observed that PWidth can be exponentially small in the size (bitlength) of a set of points with integer coordinates. Using Theorem 3.2 and the relationships between polytope condition measures, we can immediately strengthen this result and show that this is not just a "large numbers" phenomenon, namely, all condition measures are exponentially small even for a 0/1-simplex:

Corollary 3.3. *There are $d+1$ vectors in $\{0,1\}^d$ that form the vertices of a d-dimensional simplex S so that*

$$\begin{aligned}\text{width}(\text{vertices}(S)) &= \text{minwidth}(\text{vertices}(S))\\ &\le \text{PWidth}(\text{vertices}(S)) = \Phi(S) \le \text{vf}(S) \le \frac{2^{d(2+o(1))}}{d^{d/2}}.\end{aligned}$$

Proof. Let S be the d-dimensional simplex given by Theorem 3.2. Lemma 3.1 gives the leftmost equality. The rightmost inequality is one of the conclusions of Theorem 3.2. The other relations follow from Theorem 2.22. ■

3.3. Polynomial smoothed complexity of FWMs on a simplex

Now we start analyzing smoothed complexity of FWMs on the minimization of a strongly convex function with Lipschitz gradient on a simplex.

Definition 3.4. A differentiable function f is said to have L-Lipschitz gradient if for some $L>0$ and for all x, y in its domain, we have $\|\nabla f(x) - \nabla f(y)\| \le L\|x-y\|$.

Definition 3.5. A differentiable function f is μ-strongly convex if for some $\mu > 0$ and for all x, y in its domain, we have

$$f(y) \ge f(x) + \nabla f(x)^T (y-x) + \frac{\mu}{2}\|y-x\|^2.$$

In [31, Theorem 1], Lacoste-Julien and Jaggi proved the global linear convergence of FWMs on the minimization of a strongly convex function with Lipschitz gradient: suppose u_t is the current point after t good iterations[2], $f(u_t)$ satisfies

$$f(u_t) - f^* \le \left(1 - \frac{\mu}{4L}\left(\frac{\text{PWidth}(A)}{\text{diam}(A)}\right)^2\right)^t (f(u_0) - f^*), \tag{2}$$

where f^* is the optimal value and u_0 is the initial point. To show polynomial smoothed complexity, we need to prove that the measure of conditioning $\kappa = \frac{\text{PWidth}(A)}{\text{diam}(A)}$ is at least inverse polynomial in $d, 1/\sigma, 1/\delta$. We are going to get this by giving a polynomial lower bound on PWidth(A) and a polynomial upper bound on diam(A).

[2]The number of good iterations depends on variants of FWMs being used. It is always lower bounded by some linear function of the actual number of iterations. See details in [31, Theorem 1].

3.3.1. Inverse polynomial smoothed minwidth. We know from Theorem 2.22 that minwidth $\leq$ PWidth, and from Lemma 3.1 that minwidth $=$ width for any simplex. Thus, we instead find a lower bound on width, namely the diameter of a ball contained in the simplex, which is also a lower bound on PWidth. In the next lemma, we prove that a random simplex contains a ball of radius $\Omega(d^{-2})$ with probability close to 1.

Lemma 3.6. *Let $A = \{A_1, \ldots, A_{d+1}\}$ be a set of independent Gaussian random vectors with means μ_i, $\|\mu_i\| \leq 1$, $i \in [d+1]$, and covariance matrix $\sigma^2 \mathbf{I}_d$. Then for $\delta > 0$, we have*

$$\mathbb{P}\big(\text{minwidth}\big(\text{conv}(A)\big) \geq \sqrt{2\pi}\sigma\delta(d+1)^{-2}\big) \geq 1 - \delta.$$

Moreover,

$$\mathbb{P}\big(\text{PWidth}\big(\text{conv}(A)\big) \geq \sqrt{2\pi}\sigma\delta(d+1)^{-2}\big) \geq 1 - \delta.$$

Proof. It is easy to see that A forms a simplex with probability 1. From Lemma 3.1, we know the minwidth of a simplex is its width. Let D_i be the distance from A_i to the affine hull of its opposite facet, $\text{aff}\{A_j : j \neq i\}$. Conditioning on $\text{aff}\{A_j : j \neq i\}$, by the rotational invariance of Gaussian distribution, D_i is equal in distribution to the absolute value of a Gaussian random variable with mean $\mu \in \mathbb{R}$ (not necessarily zero) and variance σ^2. Let $X \sim \mathcal{N}(0, \sigma^2)$. By Lemma 2.1, we have

$$\mathbb{P}(D_i < t) \leq \mathbb{P}\big(\|X\| < t\big)$$

for all t. The right-hand side is upper bounded by $2t/\sqrt{2\pi}\sigma$, which is the product of the maximal Gaussian density and the length of the interval. Apply union bound to get

$$\mathbb{P}\left(\bigcap_{i=1}^{d+1}\{D_i \geq t\}\right) \geq 1 - \frac{2t(d+1)}{\sqrt{2\pi}\sigma}.$$

Let C_i be the distance between the center of mass of $\text{conv}(A)$ and $\text{aff}(A_j : j \neq i)$. Note that $C_i = D_i/(d+1)$. Then

$$\mathbb{P}\left(\bigcap_{i=1}^{d+1}\left\{C_i \geq \frac{t}{d+1}\right\}\right) \geq 1 - \frac{2t(d+1)}{\sqrt{2\pi}\sigma}.$$

The above expression states that with some probability the ball centered at the center of mass of radius $t/(d+1)$ lies inside $\text{conv}(A)$. Setting $t = \frac{\delta\sigma\sqrt{\pi}}{\sqrt{2}(d+1)}$ and using the fact that the width of the simplex is at least the diameter of the inscribed ball, we get

$$\mathbb{P}\big(\text{width}\big(\text{conv}(A)\big) \geq \sqrt{2\pi}\sigma\delta(d+1)^{-2}\big) \geq 1 - \delta.$$

The claim follows immediately from Lemma 3.1 and Theorem 2.22. ∎

3.3.2. Smoothed diameter.

Lemma 3.7. *Let $A = \{A_1, \dots, A_{d+1}\}$ be a set of independent Gaussian random vectors with means μ_i, $\|\mu_i\| \le 1$, $i \in [d+1]$, and covariance matrix $\sigma^2 \mathbf{I}_d$. Then for $\delta > 0$, we have*

$$\mathbb{P}\left(\operatorname{diam}(A) \le 2\left(\sigma\sqrt{2d + 3\ln\Big(\frac{d+1}{\delta}\Big)} + 1\right)\right) \ge 1 - \delta.$$

Proof. Let $A_i = \mu_i + X_i$, where $X_i \sim \mathcal{N}(0, \sigma^2 \mathbf{I}_d)$. Let $t > 0$. Triangle inequality gives that

$$\mathbb{P}\big(\|A_i\| > t + 1\big) = \mathbb{P}\big(\|X_i + \mu_i\| > t + 1\big) \le \mathbb{P}\big(\|X_i\| > t\big).$$

Apply Lemma 2.3 with $\alpha = (\sigma^2, \dots, \sigma^2)$, we have

$$\mathbb{P}\left(\|A_i\| > \sigma\sqrt{d + 2\sqrt{dt} + 2t} + 1\right) \le \mathbb{P}\left(\|X_i\| \ge \sigma\sqrt{d + 2\sqrt{dt} + 2t}\right) \le e^{-t},$$

which shows that every A_i is contained in a ball of radius $\sigma\sqrt{d + 2\sqrt{dt} + 2t} + 1 \le \sigma\sqrt{2d + 3t} + 1$ with high probability. With union bound, we see the diameter of the ball is an upper bound of the diameter of convex hull of A:

$$\mathbb{P}\Big(\operatorname{diam}\big(\operatorname{conv}(A)\big) \le 2\big(\sigma\sqrt{2d + 3t} + 1\big)\Big) \ge 1 - (d+1)e^{-t}.$$

The claim then follows by setting $t = \ln((d+1)/\delta)$. ∎

Next we restate and prove our main theorem for this section:

Theorem 1.2. *Let $A = \{A_1, \dots, A_{d+1}\}$ be a set of independent Gaussian random vectors with means μ_i, $\|\mu_i\| \le 1$, $i \in [d+1]$, and covariance matrix $\sigma^2 \mathbf{I}_d$. Then for $\delta > 0$, with probability at least $1 - \delta$, the measure of conditioning $\kappa = \frac{\operatorname{PWidth}(A)}{\operatorname{diam}(A)}$ of A is at least some inverse polynomial in d, $1/\sigma$ and $1/\delta$.*

Proof. We proved in Lemma 3.6 and Lemma 3.7 that

$$\mathbb{P}\Big(\operatorname{PWidth}\big(\operatorname{conv}(A)\big) \ge \sqrt{2\pi}\sigma\delta(d+1)^{-2}\Big) \ge 1 - \delta$$

and

$$\mathbb{P}\left(\operatorname{diam}(A) \le 2\left(\sigma\sqrt{2d + 3\ln\Big(\frac{d+1}{\delta}\Big)} + 1\right)\right) \ge 1 - \delta.$$

Thus, with probability at least $1 - 2\delta$, we have

$$\begin{aligned}\frac{\mathrm{PWidth}(A)}{\mathrm{diam}(A)} &\geq \frac{\sqrt{2\pi}\sigma\delta(d+1)^{-2}}{2\big(\sigma\sqrt{2d+3\ln((d+1)/\delta)}+1\big)}\\ &= \frac{\delta\sqrt{\pi/2}}{(d+1)^2\big(\sqrt{2d+3\ln((d+1)/\delta)}+\frac{1}{\sigma}\big)}\\ &\geq 1/\rho(d,1/\sigma,1/\delta),\end{aligned}$$

where ρ is a polynomial function of $d, 1/\sigma, 1/\delta$. ■

Going back to (2), let $h_t = f(u_t) - f^*$. We have

$$h_t \leq \left(1 - \frac{\mu}{4L}\left(\frac{\mathrm{PWidth}(A)}{\mathrm{diam}(A)}\right)^2\right)^t h_0.$$

Based on our smoothed analysis on the measure of conditioning in Theorem 1.2, with probability at least $1 - 2\delta$,

$$h_t \leq \left(1 - \frac{\mu}{4L\rho^2}\right)^t h_0 \leq e^{-\frac{\mu t}{4L\rho^2}} h_0.$$

Hence, one needs at most $\frac{4L\rho^2\ln(1/\varepsilon)}{\mu}$ good iterations to get a solution whose value is within distance $\varepsilon(f_0 - f^*)$ of f^*. Let T denote the number of good iterations, we have (using the notation from Definition 1.1)

$$\max_{\substack{A\subseteq B(0,1)\subseteq\mathbb{R}^d\\ |A|=d+1}} \mathbb{P}_g\left(T(A+g) \geq \frac{4L\rho(d,1/\sigma,1/\delta)^2\ln(1/\varepsilon)}{\mu}\right) \leq 2\delta.$$

4. Conditioning of random matrices

In this section we prove that the smallest singular value of some square submatrix of a d-by-n Gaussian random matrix is exponentially small with probability exponentially close to 1 when $n/d \geq c > 1$. From Lemma 2.6, we know that the smallest singular value of a square matrix is comparable to the minimum distance between one column vector and the span of the other column vectors (one-off-distance). If we consider exponentially narrow bands around each span of $d-1$ column vectors of a rectangular matrix, the matrix will have exponentially small minimum singular value if some other column vector falls in one of those bands. We lower bound the Gaussian measure of the union of bands by a constant using the first two terms of the inclusion-exclusion principle (Bonferroni inequality). See Section 1 for a high level overview of the proof.

We start by giving an upper bound of the intersection of two bands in Gaussian measure, which appears in the second term of the inclusion-exclusion principle. The following lemma shows that the Gaussian measure of the intersection depends on the width of bands and the angle between two bands.

Lemma 4.1. *Let $u, v \in \mathbb{R}^d$ be unit length vectors, let $\varepsilon > 0$, and let $c_S, c_T \in \mathbb{R}$. Let*

$$\mathcal{B}_S = \{x \in \mathbb{R}^d : c_S \le x \cdot u \le c_S + \varepsilon\},$$
$$\mathcal{B}_T = \{x \in \mathbb{R}^d : c_T \le x \cdot v \le c_T + \varepsilon\}.$$

Then

$$\mathcal{G}(\mathcal{B}_S \cap \mathcal{B}_T) \le \frac{\varepsilon^2}{2\pi\sqrt{1-(u\cdot v)^2}}.$$

Proof. If u and v are parallel then the claim holds. If they are not parallel, then by the structure of the Gaussian measure $\mathcal{G}$ this is a 2-dimensional problem in the plane spanned by u, v. Identify this plane with $\mathbb{R}^2$. $\mathcal{G}(\mathcal{B}_S \cap \mathcal{B}_T)$ is at most the maximum density $1/(2\pi)$ multiplied by the area of the parallelogram

$$P' := \{x \in \mathbb{R}^2 : c_S \le x \cdot u \le c_S + \varepsilon, c_T \le x \cdot u \le c_T + \varepsilon\}.$$

One can see that P' has the same area as

$$P := \{x \in \mathbb{R}^2 : |x \cdot u| \le \varepsilon/2, |x \cdot v| \le \varepsilon/2\}.$$

Defining A to be the matrix with rows u, v, we have $P = \{x : \|Ax\|_\infty \le \varepsilon/2\}$. This implies

$$\operatorname{area}(P) = \varepsilon^2 |\det A^{-1}| = \varepsilon^2/|\det A| = \varepsilon^2/\sqrt{\det AA^T} = \varepsilon^2/\sqrt{1-(u\cdot v)^2}.$$

The claim follows. ■

We now switch our focus to the random regime. The following lemma gives a probabilistic upper bound of the intersection of two bands around the spans of two (possibly not disjoint) subsets of random vectors in high-dimensional space. The bound is good when not too many points are shared by the subsets (so that the behavior is not very different from two independent bands).

Lemma 4.2. *Let $d \ge 1$. Let $0 \le k \le d-1$. Let $A_1, \dots, A_k$, $S_1, \dots, S_{d-k-1}$, $T_1, \dots,$ T_{d-k-1} be d-dimensional iid. standard Gaussian random vectors. Let*[3]

$$\mathcal{B}_S = \left(\operatorname{span}\{A_1, \dots, A_k, S_1, \dots, S_{d-k-1}\}\right)_{\varepsilon/2},$$
$$\mathcal{B}_T = \left(\operatorname{span}\{A_1, \dots, A_k, T_1, \dots, T_{d-k-1}\}\right)_{\varepsilon/2}.$$

[3]Recall that subscript ε denotes the ε-neighborhood of a set, see Section 2.

Then for any $t \geq 1$,

$$\mathbb{P}\Big(\mathcal{G}(\mathcal{B}_S \cap \mathcal{B}_T) \geq \frac{\varepsilon^2 t}{2\pi}\Big) \leq \frac{1}{t^{d-k-2}}.$$

Proof. If $d \leq 2$ or $k \geq d-2$, then the claim is immediate. Otherwise, $0 \leq k \leq d-3$ and we argue in the following way: By the structure of the Gaussian measure $\mathcal{G}$ this is a $(d-k)$-dimensional problem in $\{A_1, \ldots, A_k\}^\perp$. More precisely, let U, V be two $(d-k)$-dimensional iid. uniformly random unit-length vectors and define

$$\begin{aligned}\mathcal{B}'_S &= \{x \in \mathbb{R}^{d-k} : |x \cdot U| \leq \varepsilon/2\},\\ \mathcal{B}'_T &= \{x \in \mathbb{R}^{d-k} : |x \cdot V| \leq \varepsilon/2\}.\end{aligned}$$

Then $\mathcal{G}(\mathcal{B}_S \cap \mathcal{B}_T)$ has the same distribution as $\mathcal{G}(\mathcal{B}'_S \cap \mathcal{B}'_T)$.[4] From Lemma 4.1, we have

$$\mathcal{G}(\mathcal{B}'_S \cap \mathcal{B}'_T) \leq \frac{\varepsilon^2}{2\pi\sqrt{1-(U \cdot V)^2}}.$$

Using the rotational symmetry of the distribution of U and V and then Lemma 2.5, we get

$$\begin{aligned}\mathbb{P}\big(\sqrt{(1-(U\cdot V)^2)} \leq 1/t\big) &= \mathbb{P}\Big(\sqrt{U_1^2 + \cdots + U_{d-k-1}^2} \leq 1/t\Big)\\ &\leq \mathbb{P}\Big(\sqrt{U_1^2 + \cdots + U_{d-k-2}^2} \leq 1/t\Big)\\ &= 1/t^{d-k-2}.\end{aligned}$$

The claim follows. ∎

The main technical content of our singular value bound is the following lower bound on the Gaussian volume of the union of bands around any $d-1$ columns of a d-by-n Gaussian random matrix. We also include an upper bound on the volume.

Lemma 4.3. *Let $\varepsilon \geq 0$, $d \geq 2$. For $\{A_1, \ldots, A_n\} \subseteq \mathbb{R}^d$, define*

$$V = \mathcal{G}\Bigg(\Big(\bigcup_{S\subseteq[n],|S|=d-1} \operatorname{span} A_S\Big)_\varepsilon\Bigg).$$

(1) $V \leq (2\varepsilon/\sqrt{2\pi})\binom{n}{d-1}$.

[4] To see this, note that the Gaussian measure of $\mathcal{B}_S \cap \mathcal{B}_T$ is the same as the Gaussian measure of its projection onto $\{A_1, \ldots, A_k\}^\perp$ and the distribution of the projection of $\mathcal{B}_S$ (resp. $\mathcal{B}_T$) is the same as the distribution of $\mathcal{B}'_S$ (resp. $\mathcal{B}'_T$) after identifying $\{A_1, \ldots, A_k\}^\perp$ with $\mathbb{R}^{d-k}$.

(2) *Suppose $A_1, \dots, A_n$ are d-dimensional iid. standard Gaussian random vectors with $n/(d-1) \geq c_0 > 1$. Then there exist constants $c_2, c_4 > 1$ (that depend only on c_0) such that when $\varepsilon \leq 1/(c_4 c_2^{d-1})$ and with probability at least $1 - c_4 e^{-d}$ we have $V \geq (c_2^{d-1}/\sqrt{2\pi})\varepsilon$.*

Proof of part 1. The upper bound follows from the union bound and the fact that the 1-dimensional Gaussian density is upper bounded by $1/\sqrt{2\pi}$. ∎

Proof of part 2. Let $\mathcal{S} = \{S \subseteq [n], |S| = d-1\}$. We will use Lemma 2.4 (Gilbert–Varshamov bound). Recall that $A(n,t,w)$ denotes the maximum number of binary n-vectors with exactly w ones and pairwise Hamming distance greater than or equal to t. Use Lemma 2.4 to get the bound

$$A\big(n, c_1(d-1), d-1\big) \geq c_2^{d-1}.$$

We get a subfamily $\mathcal{T} \subseteq \mathcal{S}$ such that for all $S, T \in \mathcal{T}$ with $S \neq T$, we have

$$|S \cap T| \leq \left(1 - \frac{c_1}{2}\right)(d-1) \quad \text{and} \quad |\mathcal{T}| = c_2^{d-1}$$

for some constants $0 < c_1 < 1$, $c_2 > 1$ (that depend only on c_0), and any $d \geq 2$. Let $N = |\mathcal{T}|$.

Let $\mathcal{B}_S = (\operatorname{span} A_S)_\varepsilon$. Use the first two terms of the inclusion-exclusion principle (Bonferroni inequality) and Lemma 4.2 in a union bound applied to all pairs of sets in $\mathcal{T}$ to get $\mathcal{G}(\mathcal{B}_S \cap \mathcal{B}_T) \leq 2\varepsilon^2 t/\pi$ for all $S, T \in \mathcal{T}$, $S \neq T$. We get a bound on V that holds with probability at least

$$\begin{aligned} 1 - \frac{\binom{N}{2}}{t^{d-1-(1-c_1/2)(d-1)-2}} &= 1 - \frac{\binom{N}{2}}{t^{c_1(d-1)/2-1}} \\ &\geq 1 - \frac{N^2}{t^{c_1(d-1)/2-1}} \\ &= 1 - t\left(\frac{c_2^2}{t^{c_1/2}}\right)^{d-1} = 1 - c_3 e^{-d} \end{aligned}$$

(choosing a constant $t > 1$ that depends on $c_1(c_0)$ and $c_2(c_0)$ such that $c_2^2/t^{c_1/2} = 1/e$ and then setting $c_3 = t/e$, which ultimately depends only on c_0).The bound on V is

$$\begin{aligned} V &\geq \mathcal{G}\left(\left(\bigcup_{S \in \mathcal{T}} \operatorname{span} A_S\right)_\varepsilon\right) \\ &\geq \sum_{S \in \mathcal{T}} \mathcal{G}(\mathcal{B}_S) - \frac{1}{2} \sum_{S,T \in \mathcal{T}, S \neq T} \mathcal{G}(\mathcal{B}_S \cap \mathcal{B}_T) \\ &\geq \frac{2N\varepsilon}{\sqrt{2\pi}} e^{-\varepsilon^2/2} - \binom{N}{2} \frac{2\varepsilon^2 t}{\pi} \end{aligned}$$

$$\geq \frac{2N\varepsilon}{\sqrt{2\pi}}\Big(e^{-\varepsilon^2/2} - \frac{tN\varepsilon}{\sqrt{2\pi}}\Big)$$
$$\geq \frac{2N\varepsilon}{\sqrt{2\pi}}\Big(1 - \varepsilon^2/2 - \frac{tN\varepsilon}{\sqrt{2\pi}}\Big)$$
$$\geq \frac{N\varepsilon}{\sqrt{2\pi}} \qquad (\text{for } \varepsilon \leq \sqrt{2\pi}/(4tN)).$$

In other words, $V \geq c_2^{d-1}\varepsilon/\sqrt{2\pi}$ for $\varepsilon \leq \sqrt{2\pi}/4tN$. We finish our proof by taking $c_4 = 3ec_3/\sqrt{2\pi}$. ∎

We are ready now to restate and prove the main results of the section.

Theorem 1.4. *Let A be a d-by-n random matrix with iid. standard Gaussian entries with $d \geq 2$ and $\frac{n}{d} \geq c_0 > 1$. Then, there exist constants $c_2, c_4 > 1$, $0 < c_6 < 1$ (that depend only on c_0) such that with probability at least $1 - 2c_4c_6^d$,*

$$\min_{S\subseteq[n],|S|=d} \sigma_d(A_S) \leq \frac{1}{c_4c_2^{d-1}}.$$

Proof. Pick $c_1 \in (1, c_0)$. Let $m = \lfloor c_1 d\rfloor$. Note that

$$m \geq c_1 d - 1 \geq c_1 d - c_1 \geq c_1(d-1),$$

so that we can apply Lemma 4.3 to columns $A_1, \dots, A_m$ with $\varepsilon = 1/c_4c_2^{d-1}$. Then we get $V \geq 1/\sqrt{2\pi}c_4$ with probability greater than $1 - c_4e^{-d}$. This implies that with probability greater than

$$(1 - c_4e^{-d})\Big(1 - \Big(1 - \frac{1}{\sqrt{2\pi}c_4}\Big)^{n-m}\Big) \geq (1 - c_4e^{-d})\Big(1 - \Big(1 - \frac{1}{\sqrt{2\pi}c_4}\Big)^{(c_0-c_1)d}\Big)$$
$$\geq 1 - c_4e^{-d} - \Big(1 - \frac{1}{\sqrt{2\pi}c_4}\Big)^{(c_0-c_1)d}$$
$$\geq 1 - 2c_4c_6^d,$$

where

$$c_6 = \max\Big\{1/e, \Big(1 - \frac{1}{\sqrt{2\pi}c_4}\Big)^{(c_0-c_1)}\Big\},$$

at least one of $A_{m+1}, \dots, A_n$, say A_*, falls in V, that is, falls within distance $\varepsilon = 1/c_4c_2^{d-1}$ of $\mathrm{span}(A_S)$ for some $S \subseteq [m]$, $|S| = d-1$. Lemma 2.6 gives $\sigma_d(A_S, A_*) \leq 1/c_4c_2^{d-1}$. ∎

Theorem 1.5. *Let A be a d-by-n random matrix with iid. standard Gaussian entries with $d \geq 2$ and $1 < \frac{n}{d-1} \leq C_0$. Then, there exist constants $C_1 > 1$, $0 < C_2 < 1$ (that depend only on C_0) such that with probability at least $1 - nC_2^{d-1}$,*

$$\min_{S\subseteq[n],|S|=d} \sigma_d(A_S) \geq \frac{1}{C_1^{d-1}}.$$

Proof. Apply Lemma 4.3 to columns $A_1, \ldots, A_{n-1}$ to get

$$V \leq \frac{2\varepsilon}{\sqrt{2\pi}} \binom{n}{d-1} \leq \frac{2\varepsilon}{\sqrt{2\pi}} \left(\frac{en}{d-1}\right)^{d-1} \leq \frac{2\varepsilon}{\sqrt{2\pi}} (eC_0)^{d-1}.$$

By picking $\varepsilon = 1/C_1^{d-1}$ where $C_1 > eC_0$, there exists a constant $eC_0/C_1 < C_2 < 1$ such that $V \leq C_2^{d-1}$. This implies that, with probability at most C_2^{d-1}, column A_n is within distance $1/C_1^{d-1}$ of span A_S for some $S \subseteq [n-1]$, $|S| = d-1$. A similar claim holds for columns $A_1, \ldots, A_{n-1}$ as well. Applying the union bound, we get that no A_i falls within distance $1/C_1^{d-1}$ of span A_S for any $S \subseteq [n-1]$, $|S| = d-1$ with probability at least $1 - nC_2^{d-1}$. Lemma 2.6 gives $\sigma_d(A_S, A_n) \geq 1/C_1^{d-1}$ with probability at least $1 - nC_2^{d-1}$. ■

5. On the stability of tensor decomposition

Kruskal [29] showed a sufficient condition under which the component vectors a_i, b_i, c_i, $i = 1, \ldots, n$ of an order-3 tensor $T = \sum_{i=1}^n a_i \otimes b_i \otimes c_i$ are uniquely determined by the tensor (up to inherent ambiguities). The condition depends on a parameter now known as the Kruskal rank of a matrix: For a d-by-n matrix A, the Kruskal rank of A, denoted K-rank(A), is the maximum $r \in [n]$ such that any r columns of A are linearly independent. The condition is

$$\text{K-rank}(A) + \text{K-rank}(B) + \text{K-rank}(C) \geq 2n + 2,$$

where A, B, C are the matrices with columns (a_i), (b_i), (c_i), respectively. For concreteness, it is helpful to consider the symmetric case $A = B = C \in \mathbb{R}^{d \times n}$. Kruskal's condition becomes

$$3\,\text{K-rank}(A) \geq 2n + 2.$$

Informally, for a *generic* matrix A we have K-rank$(A) = d$, and so Kruskal's result guarantees uniqueness for generic A when $n \leq 3d/2 - 1$.

Bhaskara, Charikar and Vijayaraghavan [7, Theorem 5] extended Kruskal's uniqueness to a result that guarantees *robust decomposition*. That is, when the observed tensor is a small perturbation of the original tensor, the components of the perturbed tensor are uniquely determined and close to the components of the original tensor. Their condition for robust unique decomposition is a refinement of Kruskal's condition: Let $\tau > 0$. The *robust Kruskal rank (with threshold τ)* of A, denoted K-rank$_\tau(A)$, is the maximum $k \in [n]$ such that for any subset $S \subseteq [n]$ of size k, we have $\sigma_k(A_S) \geq 1/\tau$ (where σ_k denotes the kth largest singular value). The condition is

$$\text{K-rank}_\tau(A) + \text{K-rank}_\tau(B) + \text{K-rank}_\tau(C) \geq 2n + 2$$

and the error in the recovered components depends polynomially on τ.

In this context, Theorem 1.4 can be stated in the following equivalent way:

Theorem 5.1. *Let A be a d-by-n random matrix with iid. standard Gaussian entries with $d \geq 2$ and $n/d \geq c_0 > 1$. Then, there exist constants $c_4, c_5 > 1$, $0 < c_6 < 1$ (that depend only on c_0) such that with probability at least $1 - 2c_4 c_6^d$,*

$$\text{K-rank}_\tau(A) = d \Rightarrow \tau \geq c_4 c_5^{d-1}.$$

This has the following implication for Bhaskara, Charikar and Vijayaraghavan's result: Even though Kruskal's result guarantees uniqueness for generic A when $n = 3d/2 - 1$ (say, with probability 1 for a random Gaussian matrix, we have K-rank$(A) = d$), Bhaskara, Charikar and Vijayaraghavan's robust uniqueness can give a polynomial bound on the reconstruction error on no more than an exponentially small fraction of matrices A when the fraction is measured by the Gaussian measure. This rarity of sufficiently well-conditioned matrices A is somewhat surprising. Note that our result presents a clear limitation only as stated above: It should still be possible to apply their robust uniqueness results with K-rank$(A) = (1-\varepsilon)d$ for a small constant $\varepsilon > 0$ to guarantee robust uniqueness for $n \leq (1-\varepsilon)3d/2$.

6. On the complexity of the simplex method and the diameter of polytopes

In [13], Brunsch and Röglin introduced the following property of a matrix:

Definition 6.1 (δ-distance property, [12]). Let $A = (a_1, \ldots, a_m)^\top$ be an m-by-n matrix with unit rows. We say that A satisfies the δ-distance property if: for any $I \subseteq [m]$ and any $j \in [m]$ whenever $a_j \notin \operatorname{span}\{a_i : i \in I\}$, we have

$$d\big(a_j, \operatorname{span}\{a_i : i \in I\}\big) \geq \delta.$$

This property has been used in several papers [12,13,18,22] to study polytopes of the form $\{x \in \mathbb{R}^n : Ax \leq b\}$ to provide upper bounds of the form poly$(m, n, 1/\delta)$ on their diameter and the number of pivot steps of the simplex method. Our Theorem 1.4 combined with Lemma 2.6 and concentration of the length of a Gaussian random vector implies that, for $m/n \geq c' > 1$, matrices A with the δ-distance property for $\delta \geq c^n$, $0 < c < 1$, are "rare": they are exponentially unlikely when the rows are iid. random unit vectors. As in Section 5, this rarity of well-conditioned matrices A is somewhat surprising.

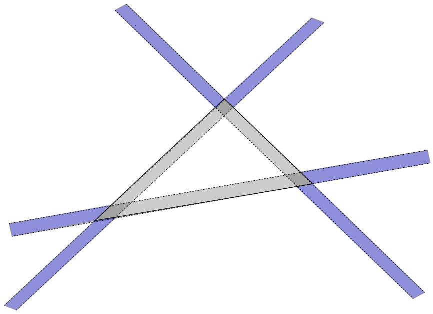


Figure 2. A polytope (triangle) and the region (blue) where a new point would create a small vertex-facet distance.

7. On the smoothed analysis of polytope conditioning

In this section we prove that the vertex-facet distance of the convex hull of a linear number of d-dimensional iid. Gaussian points can be exponentially small with probability at least some constant. The argument is a more elaborate version of the argument for the minimum singular value in Section 4 and works in the following way. Figure 2 shows a polytope, the convex hull of a partial sequence of random points, and ε-inner bands at all facets. If a new point falls into the blue region, then the new polytope, which is the convex hull of the old polytope plus the new point, will have vertex-facet distance no larger than ε: the new point is a vertex and its distance to the affine hull of the facet associated to the band where the point lies in is less than ε.

To get a lower bound on the Gaussian measure of the blue region (Lemma 7.5), we add the measures of the bands and then subtract the measures of pairwise intersections of bands and ε-inner neighborhood (grey region). Lemma 7.4 gives a bound on the measure of a pairwise intersection. Its proof is divided into two cases: Lemma 7.1 for the case where the two facets do not share vertices and Lemma 7.2 for the case where they do share vertices. This argument is a refinement of the proof of Lemma 4.2.

Lemma 7.1. *Let $S_1, \dots, S_d$, $T_1, \dots, T_d$ be iid. standard Gaussian random vectors in $\mathbb{R}^d$. Let*

$$\mathcal{B}_S = \left(\operatorname{aff}\{S_1, \dots, S_d\}\right)_{\varepsilon/2},$$
$$\mathcal{B}_T = \left(\operatorname{aff}\{T_1, \dots, T_d\}\right)_{\varepsilon/2}.$$

Then for $t \geq 1$,

$$\mathbb{P}\left(\mathcal{G}(\mathcal{B}_S \cap \mathcal{B}_T) \geq \frac{\varepsilon^2 t}{2\pi}\right) \leq \frac{1}{t^{d-2}}.$$

Proof. By the rotational invariance of the Gaussian distribution, unit normal vectors U, V to $\mathcal{B}_S, \mathcal{B}_T$ are independent and are uniformly distributed on $\mathcal{S}^{d-1}$. Define

$$\mathcal{B}'_S = \{x \in \mathbb{R}^d : |x \cdot U| \le \varepsilon/2\},$$
$$\mathcal{B}'_T = \{x \in \mathbb{R}^d : |x \cdot V| \le \varepsilon/2\}.$$

By a standard argument (say, using logconcavity), we have

$$\mathbb{P}\big(\mathcal{G}(\mathcal{B}_S \cap \mathcal{B}_T) \ge t\big) \le \mathbb{P}\big(\mathcal{G}(\mathcal{B}'_S \cap \mathcal{B}'_T) \ge t\big).$$

Then by the argument in the proof of Lemma 4.2, for any $t \ge 1$, we get that

$$\mathbb{P}\Big(\mathcal{G}(\mathcal{B}'_S \cap \mathcal{B}'_T) \ge \frac{\varepsilon^2 t}{2\pi}\Big) \le \frac{1}{t^{d-2}}.$$

The claim follows. ∎

Lemma 7.2. *Let $A_1, \dots, A_k$, $S_1, \dots, S_{d-k}$, $T_1, \dots, T_{d-k}$ be iid. standard Gaussian random vectors in $\mathbb{R}^d$, and $1 \le k \le d$. Let*

$$\mathcal{B}_S = \big(\mathrm{aff}\{A_1, \dots, A_k, S_1, \dots, S_{d-k}\}\big)_{\varepsilon/2},$$
$$\mathcal{B}_T = \big(\mathrm{aff}\{A_1, \dots, A_k, T_1, \dots, T_{d-k}\}\big)_{\varepsilon/2}.$$

Then for $0 < 2\alpha \le \beta < \pi/2$,

$$\mathbb{P}\Big(\mathcal{G}(\mathcal{B}_S \cap \mathcal{B}_T) \ge \frac{\varepsilon^2}{2\pi \sin\alpha}\Big) \le (\sin\beta)^{d-k-1} + 2\Big(\frac{\sin\alpha}{\sin(\beta-\alpha)}\Big)^{d-k-2}. \tag{3}$$

In particular, for $t > 2\pi$, we have

$$\mathbb{P}\Big(\mathcal{G}(\mathcal{B}_S \cap \mathcal{B}_T) \ge \frac{\varepsilon^2 t}{2\pi}\Big) \le 3\Big(\frac{\pi^{3/2}}{\sqrt{2t}}\Big)^{d-k-2}.$$

Proof. If $d-k \le 2$, then the bound holds immediately. Otherwise, $d-k > 2$ and we argue in the following way. By the structure of the Gaussian measure, this reduces to a $(d-k+1)$-dimensional problem: Conditioning on $A_i = a_i, i = 1, \dots, k$, we project onto the orthogonal complement of the linear subspace parallel to $\mathrm{aff}\{a_1, \dots, a_k\}$. We will then prove the bound claimed in (3) conditioning on $A_1, \dots, A_k$, which implies the claimed bound by total probability.

With a slight abuse of notation, we denote the projection of $\mathrm{aff}\{a_1, \dots, a_k\}$ as a_1 and the projections of S_i, T_i as $S_i, T_i, i = 1, \dots, d-k$. Using the fact that the Gaussian distribution is rotationally invariant, we may assume without loss of generality that

$a_1 = \mu e_1$ for some $\mu \geq 0$. A normal vector to $\text{aff}\{a_1, S_1, \ldots, S_{d-k}\}$ is[5]

$$U = \det \begin{pmatrix} e_1 & e_2 & \cdots & e_{d-k+1} \\ & & P_1^T & \\ & & \vdots & \\ & & P_{d-k}^T & \end{pmatrix}, \tag{4}$$

where $P_i := S_i - a_1$, $i \in [d-k]$. Define the matrix $P = \begin{pmatrix} P_1 \cdots P_{d-k} \end{pmatrix}$. Let V be a normal vector to $\text{aff}\{a_1, T_1, \ldots, T_{d-k}\}$, defined similarly.

Set $\begin{pmatrix} H_i \\ P_i' \end{pmatrix} = P_i$, where

$$H_i \sim \mathcal{N}(-\mu, 1) \quad \text{and} \quad P_i' \sim \mathcal{N}(0, \mathbf{I}_{d-k}).$$

Denote $H^T = \begin{pmatrix} H_1 \cdots H_{d-k} \end{pmatrix}$ as the first row of matrix P and $P' = \begin{pmatrix} P_1' \cdots P_{d-k}' \end{pmatrix}$ as the rest. H and P' are independent.

Note that $\|U\|^2 = \det(P^T P)$ (follows from (4) and the Cauchy–Binet formula). Also, $U_1 = U \cdot e_1 = \det(P')$. We now compute the distribution of the first coordinate of unit normal vector $\widehat{U}$ (using the *matrix determinant lemma* to compute the determinant of a rank-1 update):

$$\begin{aligned} \widehat{U}_1^2 &= \frac{\det(P'^T P')}{\det(P^T P)} \\ &= \frac{\det(P'^T P')}{\det(P'^T P' + H H^T)} \\ &= \frac{\det(P'^T P')}{(1 + H^T P'^{-1} P'^{-T} H) \det(P'^T P')} \\ &= \frac{1}{1 + H^T P'^{-1} P'^{-T} H}. \end{aligned}$$

Claim 7.3. We have

$$H^T P'^{-1} P'^{-T} H \overset{d}{=} \frac{\sum_{i=1}^{d-k} Y_i^2}{Y_0^2},$$

where $Y_0 \sim \mathcal{N}(0,1)$, $Y_i \sim \mathcal{N}(\mu, 1)$, $i \in [d-k]$ and $Y_0, Y_1, \ldots, Y_{d-k}$ are independent.

Proof of claim. Random variables P' and H are independent. Moreover, P' is a Gaussian matrix and therefore the distribution of P'^{-1} is invariant under any orthogonal transformation applied to rows or columns. Thus, it is enough to consider the case

[5] In the formula for U, the determinant should be interpreted as a formal cofactor expansion along the first row; the entries in the first row are the canonical vectors and the expansion gives the coefficients of these vectors (as subdeterminants).

$H = \|H\| e_1$. Note that $\|H\|^2 \overset{d}{=} \sum_{i=1}^{d-k} Y_i^2$, and

$$e_1^T P'^{-1} P'^{-T} e_1 = \|\text{first row of } P'^{-1}\|^2 \overset{d}{=} \frac{1}{Y_0^2}.$$

The claim follows. ∎

Recall that $\widehat{U}$, $\widehat{V}$ are unit normal vectors to $\mathcal{B}_S$, $\mathcal{B}_T$, respectively. We aim to show that $\mathbb{P}(\widehat{V} \in \mathcal{C}_\alpha(\widehat{U}) \cup \mathcal{C}_\alpha(-\widehat{U}))$, i.e. $\mathbb{P}(|\widehat{U} \cdot \widehat{V}| \geq \cos\alpha)$, is upper bounded by an expression of the form $c(\alpha)^d$ with $c(\alpha) \to 0$ as $\alpha \to 0$ (where $\mathcal{C}_\alpha(\widehat{U})$ denotes the spherical cap centered at $\widehat{U}$ with angle α). To see this, we divide the analysis into two cases, depending on whether the cap is close to e_1. The case analysis depends on a parameter β that will need to satisfy the constraint $\beta \geq 2\alpha$.

Case 1. $\mathcal{C}_\alpha(\widehat{U}) \subseteq \mathcal{C}_\beta(e_1) \cup \mathcal{C}_\beta(-e_1)$ (equivalently, $|\widehat{U}_1| \geq \cos(\beta - \alpha)$).

In this case, the α-cap around $\widehat{U}$ is contained in a larger cap centered at e_1:

$$\begin{aligned}\mathbb{P}\big(\{\widehat{V} \in \mathcal{C}_\alpha(\widehat{U}) \cup \mathcal{C}_\alpha(-\widehat{U})\} \cap \{\mathcal{C}_\alpha(\widehat{U}) \subseteq \mathcal{C}_\beta(e_1) \cup \mathcal{C}_\beta(-e_1)\}\big) \\ \leq \mathbb{P}\big(\widehat{V} \in \mathcal{C}_\beta(e_1) \cup \mathcal{C}_\beta(-e_1)\big) \\ = \mathbb{P}(\widehat{V}_1^2 \geq \cos^2\beta) \quad (\text{using } \beta \leq \pi/2). \end{aligned} \tag{5}$$

From Claim 7.3, we get

$$\widehat{V}_1^2 \overset{d}{=} \frac{Y_0^2}{Y_0^2 + \sum_{i=1}^n Y_i^2}.$$

To upper bound (5), we get from Lemma 2.2 that making $a_1 = 0$ (equivalently, $\mu = 0$) only makes the right-hand side larger and we then bound the case $a_1 = 0$ explicitly. More precisely, let W be a normal vector to $\text{span}\{T_1, \ldots, T_{d-k}\}$ defined similarly to U and V:

$$W = \det\begin{pmatrix} e_1 & e_2 & \cdots & e_{d-k+1} \\ & & T_1^T & \\ & & \vdots & \\ & & T_{d-k}^T & \end{pmatrix}.$$

Note that $\widehat{W}$ is a uniformly random unit vector. Following the same computation as for V, one can derive

$$\widehat{W}_1^2 \overset{d}{=} \frac{X_0^2}{X_0^2 + \sum_{i=1}^n X_i^2},$$

where $X_0, X_i \sim \mathcal{N}(0, 1)$, $i \in [d-k]$. Then by Lemma 2.2,

$$\mathbb{P}\big(\widehat{V}_1^2 \geq \cos^2\beta\big) \leq \mathbb{P}\big(\widehat{W}_1^2 \geq \cos^2\beta\big).$$

Hence,

$$\begin{aligned}
&\mathbb{P}\left(\left\{\hat{V} \in \mathcal{C}_\alpha(\hat{U}) \cup \mathcal{C}_\alpha(-\hat{U})\right\} \cap \left\{\mathcal{C}_\alpha(\hat{U}) \subseteq \mathcal{C}_\beta(e_1) \cup \mathcal{C}_\beta(-e_1)\right\}\right)\\
&\quad \le \mathbb{P}\left(\hat{W}_1^2 \ge \cos^2\beta\right)\\
&\quad = \mathbb{P}\left(\sum_{i=2}^{d-k+1} \hat{W}_i^2 \le \sin^2\beta\right)\\
&\quad \le \mathbb{P}\left(\sum_{i=2}^{d-k} \hat{W}_i^2 \le \sin^2\beta\right)\\
&\quad \le (\sin\beta)^{d-k-1} \quad \text{(by Lemma 2.5).}
\end{aligned}$$

Case 2. $\mathcal{C}_\alpha(\hat{U}) \nsubseteq \mathcal{C}_\beta(e_1) \cup \mathcal{C}_\beta(-e_1)$.

If $\mathcal{C}_\alpha(\hat{U})$ is not contained in $\mathcal{C}_\beta(e_1) \cup \mathcal{C}_\beta(-e_1)$, then $\hat{U}$ makes an angle at least $\beta - \alpha$ with e_1 and $-e_1$, that is

$$|\hat{U}_1| < \cos(\beta - \alpha). \tag{6}$$

Our goal here is to bound

$$\begin{aligned}
&\mathbb{P}\left(\left\{\hat{V} \in \mathcal{C}_\alpha(\hat{U}) \cup \mathcal{C}_\alpha(-\hat{U})\right\} \cap \left\{\mathcal{C}_\alpha(\hat{U}) \nsubseteq \mathcal{C}_\beta(e_1) \cup \mathcal{C}_\beta(-e_1)\right\}\right)\\
&\quad = \mathbb{P}\left(\left\{\hat{V} \in \mathcal{C}_\alpha(\hat{U})\right\} \cap \left\{\mathcal{C}_\alpha(\hat{U}) \nsubseteq \mathcal{C}_\beta(e_1) \cup \mathcal{C}_\beta(-e_1)\right\}\right)\\
&\qquad + \mathbb{P}\left(\left\{\hat{V} \in \mathcal{C}_\alpha(-\hat{U})\right\} \cap \left\{\mathcal{C}_\alpha(\hat{U}) \nsubseteq \mathcal{C}_\beta(e_1) \cup \mathcal{C}_\beta(-e_1)\right\}\right)\\
&\quad = 2 \cdot \mathbb{P}\left(\left\{\hat{V} \in \mathcal{C}_\alpha(\hat{U})\right\} \cap \left\{\mathcal{C}_\alpha(\hat{U}) \nsubseteq \mathcal{C}_\beta(e_1) \cup \mathcal{C}_\beta(-e_1)\right\}\right).
\end{aligned} \tag{7}$$

Observe that the distribution of $\hat{U}$ and the distribution of $\hat{V}$ are invariant under rotations orthogonal to e_1. Thus, if we let $\hat{U}_{-1}, \hat{V}_{-1}$ be the projections of $\hat{U}, \hat{V}$ orthogonal to e_1 and $\widehat{U_{-1}}, \widehat{V_{-1}}$ be their normalizations, respectively, then

$$\widehat{U_{-1}}, \widehat{V_{-1}} \sim \mathrm{Unif}(\mathcal{S}^{d-k}).$$

This observation motivates us to use the corresponding probability of projections to bound (7). We will show that under condition (6) of case 2, $\hat{V} \in \mathcal{C}_\alpha(\hat{U})$ implies that $\widehat{V_{-1}} \in \mathcal{C}_{f(\alpha)}(\widehat{U_{-1}})$, where $f(\alpha)$ is a bound (to be understood) on the angle that depends only on α. As events,

$$\begin{aligned}
\{\hat{V} \in \mathcal{C}_\alpha(\hat{U})\} &\subseteq \{\hat{V}_{-1} \in \mathrm{Proj}_{e_1^\perp} \mathcal{C}_\alpha(\hat{U})\}\\
&\subseteq \{\widehat{V_{-1}} \in \mathcal{C}_{f(\alpha)}(\widehat{U_{-1}})\}.
\end{aligned} \tag{8}$$

Bounding $f(\alpha)$ is a 3-dimensional problem since $\widehat{U_{-1}}, \widehat{V_{-1}}$ are in $\mathrm{span}\{e_1, \hat{U}, \hat{V}\}$. From now on, the analysis lives in the above 3-dimensional space to get an upper

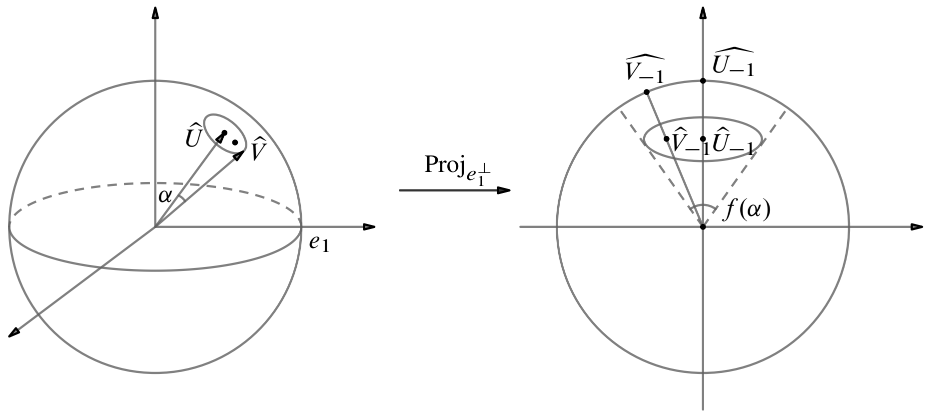


Figure 3. Case 2 of proof of Lemma 7.4.

bound on $f(\alpha)$. Let $\tilde{e}_2 = (\hat{U} - \hat{U} \cdot e_1)/\|\hat{U} - \hat{U} \cdot e_1\|$ (so that $\{e_1, \tilde{e}_2\}$ is an orthonormal basis of $\mathrm{span}\{e_1, \hat{U}\}$). Let $\{e_1, \tilde{e}_2, \tilde{e}_3\}$ be an orthonormal basis of $\mathrm{span}\{e_1, \hat{U}, \hat{V}\}$, and let $\hat{U} = (\hat{U}_1, \hat{U}_2, 0)$ be the coordinate tuple of $\hat{U}$ relative to $\{e_1, \tilde{e}_2, \tilde{e}_3\}$. Consider $x \in \mathcal{C}_\alpha(\hat{U})$ such that $x \cdot \hat{U} = \cos\gamma$. Note that its coordinates (x_1, x_2, x_3) in our chosen basis satisfy the following system of equations:

$$x_1^2 + x_2^2 + x_3^2 = 1,$$
$$x_1\hat{U}_1 + x_2\hat{U}_2 = \cos\gamma.$$

The projections of all such x (for fixed γ) onto $\mathrm{span}\{\tilde{e}_2, \tilde{e}_3\}$ form the ellipse:

$$\big(x_2 - \hat{U}_2\cos\gamma\big)^2 + x_3^2\hat{U}_1^2 = \hat{U}_1^2\sin^2\gamma.$$

If $\hat{U}_1 = 0$, then $\hat{U}_2 = 1$, and the projection is the line segment inside unit circle at $x_2 = \cos\gamma$. The angle between x_{-1} and $\widehat{U_{-1}}$ is upper bounded by γ. As γ ranges from 0 to α, $\widehat{U_{-1}}$ and $\widehat{V_{-1}}$ form an angle at most α when $\hat{U}_1 = 0$.

If $\hat{U}_1 \neq 0$, the projection is an ellipse inside the unit circle. As shown in Figure 3, angle between x_{-1} and $\widehat{U_{-1}}$ can be upper bounded by angle formed by $\widehat{U_{-1}}$ and tangent line

$$x_2 = \frac{\sqrt{\cos^2\gamma - \hat{U}_1^2}}{\sin\gamma} x_3.$$

Note that from (6), we know

$$\hat{U}_1^2 < \cos^2(\beta - \alpha) \le \cos^2\alpha \le \cos^2\gamma$$

(here we use $\beta \ge 2\alpha$ explicitly), so the tangent line always exists.

Hence, the angle between x_{-1} and $\widehat{U_{-1}}$ is at most

$$\arctan\left(\frac{\sin\gamma}{\sqrt{\cos^2\gamma - \hat{U}_1^2}}\right).$$

Furthermore, since $\arctan(\sin\gamma/\sqrt{\cos^2\gamma - \hat{U}_1^2})$ is increasing in γ, we can conclude that for any $\hat{V} \in \mathcal{C}_\alpha(\hat{U})$, its normalized projection orthogonal to e_1, $\widehat{V_{-1}}$, is contained in the spherical cap centered at $\widehat{U_{-1}}$ with polar angle at most

$$\arctan\left(\frac{\sin\alpha}{\sqrt{\cos^2\alpha - \hat{U}_1^2}}\right)$$

when $\hat{U}_1 \neq 0$.

Therefore, with (6), we can take

$$f(\alpha) = \max\left\{\arctan\left(\frac{\sin\alpha}{\sqrt{\cos^2\alpha - \cos^2(\beta-\alpha)}}\right), \alpha\right\}.$$

Combine with (7) and (8), to get

$$\begin{aligned}
&\mathbb{P}\big(\{\hat{V} \in \mathcal{C}_\alpha(\hat{U}) \cup \mathcal{C}_\alpha(-\hat{U})\} \cap \{\mathcal{C}_\alpha(\hat{U}) \not\subseteq \mathcal{C}_\beta(e_1) \cup \mathcal{C}_\beta(-e_1)\}\big) \\
&\quad \le 2 \cdot \mathbb{P}\big(\{\widehat{V_{-1}} \in \mathcal{C}_{f(\alpha)}(\widehat{U_{-1}})\} \cap \{\mathcal{C}_\alpha(\hat{U}) \not\subseteq \mathcal{C}_\beta(e_1) \cup \mathcal{C}_\beta(-e_1)\}\big) \\
&\quad \le 2 \cdot \mathbb{P}\big(|\widehat{U_{-1}} \cdot \widehat{V_{-1}}| \ge \cos\big(f(\alpha)\big)\big) \\
&\quad = 2 \cdot \mathbb{P}\Big(\sqrt{1 - (\widehat{U_{-1}} \cdot \widehat{V_{-1}})^2} \le \sin f(\alpha)\Big) \\
&\quad \le 2\big(\sin f(\alpha)\big)^{d-k-2} \qquad \text{(by Lemma 2.5)} \\
&\quad = 2\left(\max\left\{\frac{\sin\alpha}{\sqrt{\cos^2\alpha - \cos^2(\beta-\alpha) + \sin^2\alpha}}, \sin\alpha\right\}\right)^{d-k-2} \\
&\quad = 2\left(\max\left\{\frac{\sin\alpha}{\sin(\beta-\alpha)}, \sin\alpha\right\}\right)^{d-k-2} \\
&\quad = 2\left(\frac{\sin\alpha}{\sin(\beta-\alpha)}\right)^{d-k-2}.
\end{aligned}$$

Therefore,

$$\mathbb{P}\big(|\hat{U} \cdot \hat{V}| \ge \cos\alpha\big) \le (\sin\beta)^{d-k-1} + 2\left(\frac{\sin\alpha}{\sin(\beta-\alpha)}\right)^{d-k-2}. \tag{9}$$

Note that we proved bound (9) conditioning on A_i's, hence it is also a valid bound for random A_i's (unconditionally). By Lemma 4.1, (9) implies

$$\mathbb{P}\left(\mathcal{G}(\mathcal{B}_S \cap \mathcal{B}_T) \ge \frac{\varepsilon^2}{2\pi\sin\alpha}\right) \le (\sin\beta)^{d-k-1} + 2\left(\frac{\sin\alpha}{\sin(\beta-\alpha)}\right)^{d-k-2}. \tag{10}$$

Use inequalities $(2/\pi)x \le \sin x \le x$ for $0 \le x \le \pi/2$ to get

$$\mathbb{P}\left(\mathcal{G}(\mathcal{B}_S \cap \mathcal{B}_T) \ge \frac{\varepsilon^2}{4\alpha}\right) \le \beta^{d-k-1} + 2\left(\frac{\pi\alpha}{2(\beta-\alpha)}\right)^{d-k-2}.$$

Set $\beta = \sqrt{\alpha}$ and restrict $0 < \alpha < 1/4$ so that $\sqrt{\alpha} \le 1/2$. The above probabilistic bound simplifies to

$$\begin{aligned}\mathbb{P}\left(\mathcal{G}(\mathcal{B}_S \cap \mathcal{B}_T) \ge \frac{\varepsilon^2}{4\alpha}\right) &\le \alpha^{(d-k-1)/2} + 2\left(\frac{\pi\alpha}{2(\sqrt{\alpha}-\alpha)}\right)^{d-k-2}\\ &= \alpha^{(d-k-1)/2} + 2\left(\frac{\pi\sqrt{\alpha}}{2(1-\sqrt{\alpha})}\right)^{d-k-2}\\ &\le \alpha^{(d-k-1)/2} + 2(\pi\sqrt{\alpha})^{d-k-2} \quad (\text{use } 1-\sqrt{\alpha} > 1/2)\\ &\le 3(\pi\sqrt{\alpha})^{d-k-2}.\end{aligned}$$

The claim follows by setting $\alpha = \frac{\pi}{2t}$. ■

Combining Lemmas 7.1 and 7.2, we get

Lemma 7.4. *Let $A_1, \ldots, A_k$, $S_1, \ldots, S_{d-k}$, $T_1, \ldots, T_{d-k}$ be iid. standard Gaussian random vectors in $\mathbb{R}^d$ and $0 \le k \le d$. Let*

$$\begin{aligned}\mathcal{B}_S &= \left(\operatorname{aff}\{A_1, \ldots, A_k, S_1, \ldots, S_{d-k}\}\right)_{\varepsilon/2},\\ \mathcal{B}_T &= \left(\operatorname{aff}\{A_1, \ldots, A_k, T_1, \ldots, T_{d-k}\}\right)_{\varepsilon/2}.\end{aligned}$$

Then for $t > 2\pi$, we have

$$\mathbb{P}\left(\mathcal{G}(\mathcal{B}_S \cap \mathcal{B}_T) \ge \frac{\varepsilon^2 t}{2\pi}\right) \le 3\left(\frac{\pi^{3/2}}{\sqrt{2t}}\right)^{d-k-2}.$$

Suppose $P_n = \operatorname{conv}(A_1, \ldots, A_n)$ is a full-dimensional simplicial polytope in $\mathbb{R}^d$ and $\mathcal{F}_n$ is its set of facets. For $S \in \mathcal{F}_n$, we abuse notation so that S also denotes the index set of vertices of S. Let U_S be a unit inner normal vector of $\operatorname{aff}(A_S)$ to P_n. Fix $s \in S$. Define

$$(\operatorname{aff} A_S)_{\varepsilon^-} := \left\{x \in \mathbb{R}^d : 0 < d(x, \operatorname{aff} A_S) \le \varepsilon, U_S \cdot (x - A_s) \ge 0\right\}.$$

Note that the definition is independent of the choice of $s \in S$.

Lemma 7.5. *Let $\delta \in (0, 1)$. Suppose $A_1, \ldots, A_n$ are d-dimensional iid. standard Gaussian random vectors with $d = \lfloor \delta n \rfloor$. Let $P_n = \operatorname{conv}(A_1, \ldots, A_n)$, which is full-dimensional simplicial a.s. For $\varepsilon > 0$, define a.s.*

$$V_n = \mathcal{G}\left(\bigcup_{S \in \mathcal{F}_n} (\operatorname{aff} A_S)_{\varepsilon^-} \setminus P_n\right).$$

(1) $V_n \le (\varepsilon/\sqrt{2\pi})\binom{n}{d}$.

(2) *There exist* $c_2, c_7, c_8 > 1$ *(that depend only on* δ*) such that when* $\varepsilon = \varepsilon(d) \le 1/(c_8 c_2^d)$*, we have* $\lim_{n\to\infty} \mathbb{P}(V_n \ge (c_2^d/c_7)\varepsilon) = 1$.

Proof of part 1. The upper bound follows from the union bound of at most $\binom{n}{d}$ facets and the fact that the 1-dimensional Gaussian density is upper bounded by $1/\sqrt{2\pi}$. ■

Proof of part 2. From Corollary 2.12, there exists a constant $c_{\mathcal{F}} > 1$ (that depends only on δ) such that $\mathbb{P}(|\mathcal{F}_n| \ge c_{\mathcal{F}}^d) \to 1$ as $n \to \infty$. Since P_n is simplicial a.s., we may present $\mathcal{F}_n$ as a set of binary n-vectors with exactly d ones. Let $A_{\mathcal{F}_n}(t)$ be the maximum number of vectors in $\mathcal{F}_n$ with pairwise Hamming distance greater than or equal to t. Similarly to the proof of Lemma 2.4, one can pick vectors greedily (Gilbert–Varshamov bound) so that when $|\mathcal{F}_n| \ge c_{\mathcal{F}}^d$ and $c \in (0,1)$, and using $n/d < 2/\delta$ when $d \ge 2$, we have

$$A_{\mathcal{F}_n}(cd) \ge \frac{c_{\mathcal{F}}^d}{(ne/cd)^{cd}} \ge \frac{c_{\mathcal{F}}^d}{(2e/c\delta)^{cd}}.$$

Since $\lim_{c\to 0^+}(2e/c\delta)^c = 1$ and $(2e/c\delta)^c$ is increasing for $0 \le c \le 2/\delta$, we can pick $c_1 \in (0,1)$ such that $(2e/c_1\delta)^{c_1} < c_{\mathcal{F}}$. Let $c_2 = c_{\mathcal{F}}/(2e/c_1\delta)^{c_1} > 1$. Then we have,

$$\lim_{n\to\infty} \mathbb{P}\big(A_{\mathcal{F}_n}(c_1 d) \ge c_2^d\big) = 1. \tag{11}$$

Here we get a subset of facets $\mathcal{T} \subseteq \mathcal{F}_n$ such that any two different facets in $\mathcal{T}$ share no more than $(1-\frac{c_1}{2})d$ vertices, and $|\mathcal{T}| = c_2^d$ for some constants $0 < c_1 < 1, c_2 > 1$ (that depend only on δ). Let $N = |\mathcal{T}|$. Let $\mathcal{B}_S = (\operatorname{aff} A_S)_{\varepsilon^-}$, $S \in \mathcal{F}_n$. Using an argument similar to the proof of Lemma 4.3, we get

$$\begin{aligned}
V_n &= \mathcal{G}\Big(\bigcup_{S\in\mathcal{F}_n} (\operatorname{aff} A_S)_{\varepsilon^-} \setminus P_n\Big) \\
&= \mathcal{G}\Big(\bigcup_{S\in\mathcal{F}_n} (\operatorname{aff} A_S)_{\varepsilon^-}\Big) - \mathcal{G}\big(P_n \setminus (P_n)_{-\varepsilon}\big) \\
&\ge \mathcal{G}\Big(\bigcup_{S\in\mathcal{T}} (\operatorname{aff} A_S)_{\varepsilon^-}\Big) - \mathcal{G}\big(P_n \setminus (P_n)_{-\varepsilon}\big) \\
&\ge \sum_{S\in\mathcal{T}} \mathcal{G}(\mathcal{B}_S) - \frac{1}{2}\sum_{S,T\in\mathcal{T}, S\ne T} \mathcal{G}(\mathcal{B}_S \cap \mathcal{B}_T) - \mathcal{G}\big(P_n \setminus (P_n)_{-\varepsilon}\big).
\end{aligned}$$

We are going to bound each of the three terms in the last expression.

First term: $\sum_{S\in\mathcal{T}} \mathcal{G}(\mathcal{B}_S)$. From Lemma 2.9, there exists a constant $c_3 > 0$ (that depends only on δ) such that

$$\mathbb{P}\big(\max_{S\subseteq[n],|S|=d} \operatorname{dist}(\operatorname{aff} A_S, 0) \leq c_3\big) \geq 1 - 2e^{-d}.$$

Moreover, we increase c_3 so that $c_3 > 1$, which ensures that $c_3 \geq \varepsilon$. Recall that $\mathcal{B}_S = (\operatorname{aff} A_S)_{\varepsilon^-}$. We get

$$\mathbb{P}\left(\sum_{S\in\mathcal{T}} \mathcal{G}(\mathcal{B}_S) \geq \frac{N\varepsilon}{\sqrt{2\pi}} e^{-2c_3^2}\right) \geq 1 - 2e^{-d}. \tag{12}$$

Second term: $\frac{1}{2}\sum_{S,T\in\mathcal{T},S\neq T} \mathcal{G}(\mathcal{B}_S \cap \mathcal{B}_T)$. Use Lemma 7.4 in a union bound applied to all pairs of d-subsets in $[n]$ whose intersections are no larger than $(1-\frac{c_1}{2})d$. We upper bound the number of such pairs by $\binom{n}{d}^2 \leq c_9^d$ for some $c_9 > 1$. For $t > 2\pi$, we have that

$$\frac{1}{2}\sum_{S,T\in\mathcal{T},S\neq T} \mathcal{G}(\mathcal{B}_S \cap \mathcal{B}_T) \leq \binom{N}{2}\frac{\varepsilon^2 t}{2\pi}$$

holds with probability at least

$$1 - 3c_9^d\left(\frac{\pi^{3/2}}{\sqrt{2t}}\right)^{d-(1-\frac{c_1}{2})d-2} = 1 - \frac{6t}{\pi^3}\left(c_9\left(\frac{\pi^{3/2}}{\sqrt{2t}}\right)^{c_1/2}\right)^d.$$

Choose $t = c_4 := \frac{1}{2}\pi^3(ec_9)^{4/c_1}$ to get

$$\mathbb{P}\left(\frac{1}{2}\sum_{S,T\in\mathcal{T},S\neq T} \mathcal{G}(\mathcal{B}_S \cap \mathcal{B}_T) \leq \binom{N}{2}\frac{\varepsilon^2 c_4}{2\pi}\right) \geq 1 - \frac{6c_4}{\pi^3}e^{-d}. \tag{13}$$

Third term: $\mathcal{G}(P_n \setminus (P_n)_{-\varepsilon})$. From Lemma 2.7, we know $\mathcal{G}(P_n \setminus (P_n)_{-\varepsilon}) \leq c_5\varepsilon d^{1/4}$ for some absolute constant c_5. Combining (11), (12) and (13) we conclude, with probability $1 - o(1)$ as $d \to \infty$, that

$$\begin{aligned} V_n &\geq \sum_{S\in\mathcal{T}} \mathcal{G}(B_S) - \frac{1}{2}\sum_{S,T\in\mathcal{T},S\neq T} \mathcal{G}(B_S \cap B_T) - \mathcal{G}\big(P_n \setminus (P_n)_{-\varepsilon}\big) \\ &\geq \frac{N\varepsilon}{\sqrt{2\pi}}e^{-2c_3^2} - \binom{N}{2}\frac{\varepsilon^2 c_4}{2\pi} - c_5\varepsilon d^{1/4} \\ &\geq \frac{N\varepsilon}{\sqrt{2\pi}}\left(e^{-2c_3^2} - \frac{N\varepsilon c_4}{2\sqrt{2\pi}} - \frac{\sqrt{2\pi}c_5 d^{1/4}}{N}\right). \end{aligned}$$

Note that $\sqrt{2\pi}c_5 d^{1/4}/N$ decays exponentially in d. Therefore, when $\varepsilon \leq 1/e^{2c_3^2}c_4 N$,

$$\lim_{n\to\infty} \mathbb{P}\left(V_n \geq \frac{N\varepsilon}{2\sqrt{2\pi}e^{2c_3^2}}\right) = 1.$$

The proof is finished by setting $c_7 = 2\sqrt{2\pi}e^{2c_3^2}$ and $c_8 = e^{2c_3^2}c_4$. ∎

We are ready now to restate and prove the main result of the section.

Theorem 1.3. *Let $\delta \in (0, 1)$. Suppose $A = \{A_1, \ldots, A_{n+1}\}$ is a set of iid. standard Gaussian random vectors in $\mathbb{R}^d$ and $d = \lfloor \delta n \rfloor$. Let $P_{n+1} = \operatorname{conv}(A_1, \ldots, A_{n+1})$. Then*

$$\mathbb{P}\big(\operatorname{diam}(P_{n+1}) \geq \sqrt{d}\big) \geq 1 - e^{-\frac{nd}{32}},$$

and there exist constants $0 < c < 1$ and $0 < c' < 1$ (that depend only on δ) such that,

$$\lim_{n\to\infty} \mathbb{P}\big(\operatorname{vf}(P_{n+1}) \leq c^d\big) \geq c'.$$

Hence, the measure of conditioning $\kappa = \frac{\operatorname{vf}(P_{n+1})}{\operatorname{diam}(P_{n+1})}$ of A is exponentially small in d with constant probability.

Proof. For $\operatorname{diam}(P_{n+1})$, by Lemma 2.3 we have

$$\begin{aligned}
&\mathbb{P}\big(\operatorname{diam}(P_{n+1})^2 \leq 2d - 4\sqrt{dt}\big) \\
&\quad = \mathbb{P}\big(\|A_i - A_j\|^2 \leq 2d - 4\sqrt{dt},\ \forall i \neq j \in [n+1]\big) \\
&\quad \leq \mathbb{P}\left(\bigcap_{i=1}^{\lfloor (n+1)/2\rfloor} \|A_{2i-1} - A_{2i}\|^2 \leq 2d - 4\sqrt{dt}\right) \\
&\quad \leq (e^{-t})^{n/2}.
\end{aligned}$$

We get the claimed bound by setting $t = d/16$.

Apply Lemma 7.5 to $P_n = \operatorname{conv}(A_1, \ldots, A_n)$ with $\varepsilon = 1/(c_8 c_2^d)$, we have

$$\lim_{n\to\infty} \mathbb{P}\left(V_n \geq \frac{1}{c_7 c_8}\right) = 1.$$

Since $\operatorname{vf}(P_{n+1}) \leq \varepsilon$ when $A_{n+1} \in V_n$, then

$$\lim_{n\to\infty} \mathbb{P}\big(\operatorname{vf}(P_{n+1}) \leq 1/(c_8 c_2^d)\big) \geq \frac{1}{c_7 c_8}.$$

The claim follows by picking $c = 1/c_2$ and $c' = 1/c_7 c_8$. ∎

8. Discussion and open problems

In Section 4 we showed that, for $c > 1$, a d-by-n random Gaussian matrix with $n \geq cd$ has a d-by-d submatrix with minimum singular value that is exponentially small with high probability. Does this need to be a probabilistic statement or is there a comparable version that holds for *all* matrices? Say, is it true that for any $c > 1$ and any d-by-n matrix with $n \geq cd$ and unit columns one can find a d-by-d submatrix whose smallest singular value is at most $e^{-\Omega(d)}$? For concreteness one can take $n = 2d$ and restate the question geometrically using Lemma 2.6: Is it true that for any set of $2d$ unit vectors in $\mathbb{R}^d$ there is at least one vector that is at distance at most $e^{-\Omega(d)}$ to the span of some $d - 1$ other vectors from the set?

Acknowledgments. We would like to thank Nina Amenta, Jesús De Loera, Miles Lopes, Javier Peña, Thomas Strohmer, Roman Vershynin and Van Vu for helpful discussions. We also thank anonymous referees for helpful comments.

Funding. This material is based upon work supported by the National Science Foundation under Grants CCF-1657939, CCF-1422830, CCF-2006994 and CCF-1934568.

References

[1] F. Affentranger and J. A. Wieacker, On the convex hull of uniform random points in a simple d-polytope. *Discrete Comput. Geom.* **6** (1991), no. 4, 291–305 Zbl 0725.52004 MR 1098810

[2] N. Alon and V. H. Vũ, Anti-Hadamard matrices, coin weighing, threshold gates, and indecomposable hypergraphs. *J. Combin. Theory Ser. A* **79** (1997), no. 1, 133–160 Zbl 0890.05011 MR 1449753

[3] T. W. Anderson, The integral of a symmetric unimodal function over a symmetric convex set and some probability inequalities. *Proc. Amer. Math. Soc.* **6** (1955), 170–176 Zbl 0066.37402 MR 69229

[4] K. Ball, The reverse isoperimetric problem for Gaussian measure. *Discrete Comput. Geom.* **10** (1993), no. 4, 411–420 Zbl 0788.52010 MR 1243336

[5] F. Barthe, O. Guédon, S. Mendelson, and A. Naor, A probabilistic approach to the geometry of the l_p^n-ball. *Ann. Probab.* **33** (2005), no. 2, 480–513 Zbl 1071.60010 MR 2123199

[6] A. Bhaskara, M. Charikar, A. Moitra, and A. Vijayaraghavan, Smoothed analysis of tensor decompositions. 2013, arXiv:1311.3651

[7] A. Bhaskara, M. Charikar, and A. Vijayaraghavan, Uniqueness of tensor decompositions with applications to polynomial identifiability. In *COLT 2014 – 27th Annual Conference on Learning Theory*, pp. 742–778, Proceedings of Machine Learning Research 35, PMLR, 2014

[8] A. Beck and S. Shtern, Linearly convergent away-step conditional gradient for non-strongly convex functions. *Math. Program.* **164** (2017), no. 1–2, Ser. A, 1–27 Zbl 1370.90010 MR 3661022

[9] R. Beier and B. Vöcking, Typical properties of winners and losers in discrete optimization. In *STOC'04 – Proceedings of the 36th Annual ACM Symposium on Theory of Computing*, pp. 343–352, ACM, New York, 2004

[10] R. Beier and B. Vöcking, Typical properties of winners and losers in discrete optimization. *SIAM J. Comput.* **35** (2006), no. 4, 855–881 Zbl 1096.68066 MR 2203730

[11] K. J. Böröczky, G. Lugosi, and M. Reitzner, Facets of high-dimensional Gaussian polytopes. 2018, arXiv:1808.01431

[12] T. Brunsch, A. Großwendt, and H. Röglin, Solving totally unimodular LPs with the shadow vertex algorithm. In *32nd International Symposium on Theoretical Aspects of Computer Science*, pp. 171–183, LIPIcs. Leibniz Int. Proc. Inform. 30, Schloss Dagstuhl. Leibniz-Zent. Inform., Wadern, 2015 Zbl 1355.68114 MR 3356411

[13] T. Brunsch and H. Röglin, Finding short paths on polytopes by the shadow vertex algorithm. In *Automata, languages, and programming. Part I*, pp. 279–290, Lecture Notes in Comput. Sci. 7965, Springer, Heidelberg, 2013 Zbl 1336.68260 MR 3109078

[14] T. T. Cai, T. Jiang, and X. Li, Asymptotic analysis for extreme eigenvalues of principal minors of random matrices. *Ann. Appl. Probab.* **31** (2021), no. 6, 2953–2990 Zbl 1486.60010 MR 4350979

[15] E. J. Candes and T. Tao, Decoding by linear programming. *IEEE Trans. Inform. Theory* **51** (2005), no. 12, 4203–4215 Zbl 1264.94121 MR 2243152

[16] E. J. Candes and T. Tao, Near-optimal signal recovery from random projections: Universal encoding strategies? *IEEE Trans. Inform. Theory* **52** (2006), no. 12, 5406–5425 Zbl 1309.94033 MR 2300700

[17] V. Chernozhukov, D. Chetverikov, and K. Kato, Central limit theorems and bootstrap in high dimensions. *Ann. Probab.* **45** (2017), no. 4, 2309–2352 Zbl 1377.60040 MR 3693963

[18] D. Dadush and N. Hähnle, On the shadow simplex method for curved polyhedra. *Discrete Comput. Geom.* **56** (2016), no. 4, 882–909 Zbl 1377.90055 MR 3561794

[19] J. A. De Loera, J. Haddock, and L. Rademacher, The minimum Euclidean-norm point in a convex polytope: Wolfe's combinatorial algorithm is exponential. *SIAM J. Comput.* **49** (2020), no. 1, 138–169 Zbl 1453.90113 MR 4065201

[20] D. L. Donoho and J. Tanner, Neighborliness of randomly projected simplices in high dimensions. *Proc. Natl. Acad. Sci. USA* **102** (2005), no. 27, 9452–9457 Zbl 1135.60300 MR 2168716

[21] A. Edelman, Eigenvalues and condition numbers of random matrices. *SIAM J. Matrix Anal. Appl.* **9** (1988), no. 4, 543–560 Zbl 0678.15019 MR 964668

[22] F. Eisenbrand and S. Vempala, Geometric random edge. *Math. Program.* **164** (2017), no. 1-2, Ser. A, 325–339 Zbl 1373.90071 MR 3661034

[23] M. Frank and P. Wolfe, An algorithm for quadratic programming. *Naval Res. Logist. Quart.* **3** (1956), 95–110 MR 89102

[24] D. Garber and E. Hazan, A polynomial time conditional gradient algorithm with applications to online and stochastic optimization. 2013, arXiv:1301.4666

[25] J. Guélat and P. Marcotte, Some comments on Wolfe's "away step". *Math. Programming* **35** (1986), no. 1, 110–119 Zbl 0592.90074 MR 842638

[26] D. Hug, G. O. Munsonius, and M. Reitzner, Asymptotic mean values of Gaussian polytopes. *Beiträge Algebra Geom.* **45** (2004), no. 2, 531–548 Zbl 1082.52003 MR 2093024

[27] D. Hug and M. Reitzner, Gaussian polytopes: Variances and limit theorems. *Adv. in Appl. Probab.* **37** (2005), no. 2, 297–320 Zbl 1089.52003 MR 2144555

[28] S. Jukna, *Extremal combinatorics*. Second edn., Texts Theor. Comput. Sci., EATCS Ser., Springer, Heidelberg, 2011 Zbl 1239.05001 MR 2865719

[29] J. B. Kruskal, Three-way arrays: Rank and uniqueness of trilinear decompositions, with application to arithmetic complexity and statistics. *Linear Algebra Appl.* **18** (1977), no. 2, 95–138 Zbl 0364.15021 MR 444690

[30] S. Lacoste-Julien and M. Jaggi, An affine invariant linear convergence analysis for Frank–Wolfe algorithms. 2013, arXiv:1312.7864

[31] S. Lacoste-Julien and M. Jaggi, On the global linear convergence of Frank–Wolfe optimization variants. In *NeurIPS 2015 – Advances in Neural Information Processing Systems 28*, pp. 496–504, Curran Associates, Inc., 2015

[32] B. Laurent and P. Massart, Adaptive estimation of a quadratic functional by model selection. *Ann. Statist.* **28** (2000), no. 5, 1302–1338 Zbl 1105.62328 MR 1805785

[33] F. Nazarov, On the maximal perimeter of a convex set in $\mathbb{R}^n$ with respect to a Gaussian measure. In *Geometric aspects of functional analysis*, pp. 169–187, Lecture Notes in Math. 1807, Springer, Berlin, 2003 Zbl 1036.52014 MR 2083397

[34] F. Pedregosa, G. Negiar, A. Askari, and M. Jaggi, Linearly convergent Frank–Wolfe with backtracking line-search. In *23rd International Conference on Artificial Intelligence and Statistics*, pp. 1–10, Proceedings of Machine Learning Research 108, PMLR, 2020

[35] J. Peña and D. Rodríguez, Polytope conditioning and linear convergence of the Frank–Wolfe algorithm. *Math. Oper. Res.* **44** (2019), no. 1, 1–18 Zbl 1440.90048 MR 3920711

[36] J. Peña, D. Rodríguez, and N. Soheili, On the von Neumann and Frank–Wolfe algorithms with away steps. *SIAM J. Optim.* **26** (2016), no. 1, 499–512 Zbl 1382.90071 MR 3461322

[37] H. Raynaud, Sur l'enveloppe convexe des nuages de points aléatoires dans R^n. I. *J. Appl. Probability* **7** (1970), 35–48 Zbl 0192.53602 MR 258089

[38] A. A. Razborov, Bounded-depth formulae over $\{\&, \oplus\}$ and some combinatorial problems. *Problems of Cybernetics* (1988), no. 134, 149–166 Zbl 0668.94017 MR 944292

[39] H. Röglin and B. Vöcking, Smoothed analysis of integer programming. In *Integer programming and combinatorial optimization*, pp. 276–290, Lecture Notes in Comput. Sci. 3509, Springer, Berlin, 2005 Zbl 1119.90332 MR 2210028

[40] H. Röglin and B. Vöcking, Smoothed analysis of integer programming. *Math. Program.* **110** (2007), no. 1, Ser. B, 21–56 Zbl 1111.90077 MR 2306129

[41] M. Rudelson and R. Vershynin, Non-asymptotic theory of random matrices: Extreme singular values. In *Proceedings of the International Congress of Mathematicians. Volume III*, pp. 1576–1602, Hindustan Book Agency, New Delhi, 2010 Zbl 1227.60011 MR 2827856

[42] D. Spielman and S.-H. Teng, Smoothed analysis of algorithms: Why the simplex algorithm usually takes polynomial time. In *STOC'01 – Proceedings of the 36th Annual ACM Symposium on Theory of Computing*, pp. 296–305, ACM, New York, 2001 Zbl 1323.68636 MR 2120328

[43] D. Spielman and S.-H. Teng, Smoothed analysis: An attempt to explain the behavior of algorithms in practice. *Commun. ACM* **52** (2009), no. 10, 76–84

[44] A. M. Vershik and P. V. Sporyshev, Asymptotic behavior of the number of faces of random polyhedra and the neighborliness problem. *Selecta Math. Soviet* **11** (1992), no. 2, 181–201 Zbl 0791.52011 MR 1166627

[45] R. Vershynin, *High-dimensional probability: An introduction with applications in data science*. Camb. Ser. Stat. Probab. Math. 47, Cambridge Univ. Press, Cambridge, 2018 Zbl 1430.60005 MR 3837109

[46] P. Wolfe, Finding the nearest point in a polytope. *Math. Programming* **11** (1976), no. 2, 128–149 Zbl 0352.90046 MR 452683

[47] G. M. Ziegler, *Lectures on polytopes*. Grad. Texts in Math. 152, Springer, New York, 1995 Zbl 0823.52002 MR 1311028

[48] G. M. Ziegler, Lectures on 0/1-polytopes. In *Polytopes – combinatorics and computation (Oberwolfach, 1997)*, pp. 1–41, DMV Sem. 29, Birkhäuser, Basel, 2000 Zbl 0966.52012 MR 1785291

Received 9 March 2021; revised 7 July 2022.

Luis Rademacher
Department of Mathematics, University of California, Davis, One Shields Avenue, Davis, CA 95616, USA; lrademac@ucdavis.edu

Chang Shu
Department of Mathematics, University of California, Davis, One Shields Avenue, Davis, CA 95616, USA; ccshu@ucdavis.edu