The Mixing Time of the Dikin Walk in a Polytope — A Simple Proof

Sushant Sachdeva[a,*], Nisheeth K. Vishnoi[b]

[a] *Yale University, New Haven, CT, USA.*
[b] *École Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland.*

Abstract

We study the mixing time of the Dikin walk in a polytope — a random walk based on the log-barrier from the interior point method literature. This walk, and a close variant, were studied by Narayanan (2016) and Kannan-Narayanan (2012). Bounds on its mixing time are important for algorithms for sampling and optimization over polytopes. Here, we provide a simple proof of their result that this random walk mixes in time $O(mn)$ for an n-dimensional polytope described using m inequalities.

Keywords: Polytopes, Sampling, Volume computation, Random walks, Interior point methods.

1. Introduction

Sampling a point from the uniform distribution on a polytope $K \subseteq \mathbb{R}^n$ is an extensively-studied problem and is a crucial ingredient in several computational tasks involving convex bodies. Towards this, typically, one sets up an ergodic and reversible random walk inside K whose stationary distribution is uniform over K. The mixing time of such a walk determines its efficacy, and, in turn, depends on the isoperimetric constant of K with respect to the transition function of the walk. Starting with the influential work of Dyer *et al.* [3], there has been a long line of work on faster and faster algorithms for generating an approximately uniform point from a convex body. Moreover, since convex bodies show up in a variety of areas, there is a wide body of work connecting random walks and isoperimetry in convex bodies to several areas in mathematics and optimization.

One such important connection to the interior point method literature was presented in the works of Kannan and Narayanan [6] and Narayanan [9] who proposed the *Dikin walk* in a polytope. Roughly, the uniform version of the Dikin walk, considered by [6], when at a point $x \in K$, computes the *Dikin ellipsoid* at x, and moves to a random point in it after a suitable Metropolis filter. The Metropolis step ensures that the walk is ergodic and reversible. The Gaussian version of the Dikin walk, considered by [9], picks the new point from a Gaussian distribution centered at x with its covariance given by the Dikin ellipsoid at x, and applies a suitable Metropolis filter. The Dikin ellipsoid at a point x is the ellipsoid described by the Hessian of the log-barrier function at x. It was introduced by Dikin in the first interior point method for linear programming [2].

Several virtues of the Dikin ellipsoid (see [10, 11, 7]) were used by [6, 9] to prove that the mixing time of the Dikin walk is $O(mn)$ starting from a warm start, when K is described by m inequality constraints. Recall that a distribution over K is said to be a warm start if its density is bounded from above by a constant relative to the uniform distribution on K. Roughly, the proof (for either walk) consists of two parts: (1) an isoperimetric inequality, proved by Lovász [8], for convex bodies in terms of a distance introduced by Hilbert, and (2) a bound on the changes in the sampling distributions of the Dikin walk in terms of the Hilbert distance. The bound in (2) was the key technical contribution of [6, 9] towards establishing the mixing time of the Dikin walk. We present a simple proof of this bound for the Gaussian Dikin walk implying that it mixes in time $O(mn)$. Our proof uses well-known facts about Gaussians, and concentration of Gaussian polynomials.

1.1. Dikin walk on Polytopes

Suppose $K \subseteq \mathbb{R}^n$ is a bounded polytope with a non-empty interior, described by m inequalities, $a_i^\top x \geq b_i$, for $i \in [m]$. We use the notation $x \in K$ to denote that x is in the interior of K. The log-barrier function for K at $x \in K$ is $F(x) := -\sum_{i\in[m]} \log(a_i^\top x - b_i)$. Let $H(x)$ denote the Hessian of F at x, *i.e.*, $H(x) := \sum_{i\in[m]} \frac{1}{(a_i^\top x - b_i)^2} a_i a_i^\top$. For all $x \in K$, $H(x)$ is a positive definite matrix, and defines the *local norm* at x, denoted $\|\cdot\|_x$, as $\|v\|_x^2 := v^\top H(x) v$. The ellipsoid $\{z : \|z - x\|_x \leq 1\}$ is known as the Dikin ellipsoid at x.

From a point $x \in K$, the next point z in the Dikin walk is sampled from the Dikin ellipsoid at x. The uni-

*Corresponding author
Email addresses: sachdeva@cs.yale.edu (Sushant Sachdeva), nisheeth.vishnoi@epfl.ch (Nisheeth K. Vishnoi)

form Dikin walk, considered by [6], sampled the new point z from the uniform distribution in this ellipsoid. In the Gaussian Dikin walk, considered by [9], z is sampled from g_x, a multivariate Gaussian distribution centered at x with covariance matrix $\frac{r^2}{n}H(x)^{-1}$, where r is a constant. Thus, the density of the distribution is given by

$$g_x(z) = \sqrt{\det H(x)}\left(\frac{n}{2\pi r^2}\right)^{n/2} \cdot \exp\left(-\frac{n}{2r^2}\cdot \|z-x\|_x^2\right).$$

Equivalently, the next point z is given by

$$z = x + \frac{r}{\sqrt{n}}\left(H(x)\right)^{-1/2} g,$$

where g is an n-dimensional vector with each coordinate of g sampled as an independent standard gaussian $\mathcal{N}(0,1)$.

In order to convert this into a random walk that stays inside K, with its stationary distribution as the uniform distribution on K, we apply the Metropolis filter to obtain the transition probability density p_x of the Gaussian Dikin walk: $\forall z \neq x$, if $z \in K$, $p_x(z) = \min\{g_x(z), g_z(x)\}$ (the walk stays at x with the remaining probability).

1.2. Hilbert Metric, Isoperimetry, and Mixing Time

We introduce the distance function which plays an important role in establishing the mixing time of the Dikin walk. Given two points $x, y \in K$, let p, q be the end points of the chord in K passing through x, y, such that the points lie in the order p, x, y, q. We define $\sigma(x,y) := \frac{|\overline{xy}||\overline{pq}|}{|\overline{px}||\overline{qy}|}$, where $|\overline{xy}|$ denotes the length of the line segment xy. $\log(1+\sigma(x,y))$ is a metric on K, known as Hilbert metric.

Lovász proved the following theorem for any random walk on K: Suppose for any two initial points $x, y \in K$ that are close in σ distance, the statistical distance of the distributions after one step of the walk each from x and y, is bounded away from 1. Then, the lazy version of the random walk (where we stay at the current point with probability 1/2 at each step) mixes rapidly.

Theorem 1 (Lovász [8]). *Consider a reversible random walk in K with its stationary distribution being uniform on K. Suppose $\exists \Delta > 0$ such that for all $x, y \in K$ with $\sigma(x,y) \leq \Delta$, we have $\|p_x - p_y\|_1 \leq 1 - \Omega(1)$, where p_x denotes the distribution after one step of the random walk from x. Then, after $O(\Delta^{-2})$ steps, the lazy version of the walk from a warm start is within 1/4 total variation distance from the uniform distribution on K.*

Kannan and Narayanan proved that the transition function of the uniform Dikin walk, p_x, for $x \in \text{int}(K)$, satisfies the hypothesis of the theorem above with $\Delta = \Omega\left(\frac{1}{\sqrt{mn}}\right)$, thus implying that it mixes in $O(mn)$ steps from a warm start. An analogous result for the Gaussian Dikin walk is implicit in the work of Narayanan. Our main contribution is an alternative and simple proof of their main technical contributions. In particular, we prove the following theorem.

Theorem 2. *Let $\varepsilon \in (0, 1/2]$. For the Gaussian Dikin walk on K with $r \leq \frac{\varepsilon}{400}(\log\frac{200}{\varepsilon})^{-3/2}$, for any two points $x, y \in K$ such that $\|x-y\|_x \leq \frac{r}{\sqrt{n}}$, we have $\|p_x - p_y\|_1 \leq \varepsilon$.*

In order to use this theorem along with Theorem 1 to obtain the claimed mixing time bound, one needs a simple fact that, for any x, y in a polytope K, which is described using m inequalities, $\sigma(x,y) \geq \frac{1}{\sqrt{m}}\|x-y\|_x$. A proof of this fact is given in the appendix; see Lemma 9.

The following two lemmas are the main ingredients in the proof of Theorem 2: (1) If two points x, y are close in the local norm, *i.e*, $\|x-y\|_x \leq \frac{r}{\sqrt{n}}$, then the two Gaussian distributions g_x and g_y are close in statistical distance. (2) If r is small enough (as a function of ε), then for all x, p_x and g_x are ε-close in statistical distance.

Lemma 3. *Let $r \leq 1$, and $c \geq 0$ be such that $c \leq \min\{r, 1/3\}$. Let $x, y \in K$. If $\|x-y\|_x \leq \frac{c}{\sqrt{n}}$, then $\|g_x - g_y\|_1 \leq 3c$.*

This lemma relies on a well-known fact about the Kullback-Leibler divergence between two multivariate Gaussian distributions, and Pinsker's inequality that bounds the statistical distance between two distributions in terms of their divergence.

Lemma 4. *Given $\varepsilon \in [0, 1/2]$, for $r \leq \frac{\varepsilon}{100}(\log\frac{50}{\varepsilon})^{-3/2}$, we have $\|p_x - g_x\|_1 \leq \varepsilon$.*

This lemma, which shows that the Metropolis filter does not change the distribution much, relies on a result on the concentration of Gaussian polynomials, proved using hypercontractivity. Given the above lemmas, Theorem 2 follows by applying triangle inequality.

2. Statistical distance between Gaussians and the local norm

In this section, we present a proof of Lemma 3 that bounds the statistical distance between g_x and g_y for two points x, y that are close in the local norm. We need the following well-known fact about the Kullback-Leibler divergence between two multivariate Gaussian distributions.

Fact 5. *Let $G_1 = \mathcal{N}(\mu_1, \Sigma_1)$ and $G_2 = \mathcal{N}(\mu_2, \Sigma_2)$ be two n-dimensional Gaussian distributions. Then,*

$$D_{KL}(G_2||G_1) = \frac{1}{2}\left(Tr\left(\Sigma_1^{-1}\Sigma_2\right) - n + \log\frac{\det\Sigma_1}{\det\Sigma_2} + (\mu_1-\mu_2)^\top\Sigma_1^{-1}(\mu_1-\mu_2)\right),$$

where D_{KL} denotes the Kullback-Leibler divergence

$$D_{KL}(P||Q) = \int \log\frac{P(x)}{Q(x)}\,\mathrm{d}P(x).$$

In order to use this theorem, we have to bound the eigenvalues of $H(x)H(y)^{-1}$. For x, y that are close in the local norm, this follows since $H(x) \approx H(y)$.

Proof Lemma 3: From the assumption, we have,

$$\frac{c^2}{n} \geq \|x-y\|_x^2 = \sum_{i\in[m]} \frac{(a_i^\top(x-y))^2}{(a_i^\top x - b_i)^2} \geq \max_{i\in[m]} \frac{(a_i^\top(x-y))^2}{(a_i^\top x - b_i)^2}.$$

Thus, for all $i \in [m]$, we have

$$\left(1 - \frac{c}{\sqrt{n}}\right)(a_i^\top x - b_i) \leq (a_i^\top y - b_i) \leq \left(1 + \frac{c}{\sqrt{n}}\right)(a_i^\top x - b_i).$$

By the definition of H, we get,

$$\left(1 - \frac{c}{\sqrt{n}}\right)^2 H(y) \preceq H(x) \preceq \left(1 + \frac{c}{\sqrt{n}}\right)^2 H(y).$$

Thus, all eigenvalue $\lambda_1, \ldots, \lambda_n > 0$ of $H(x)H(y)^{-1}$, satisfy

$$\left(1 - \frac{c}{\sqrt{n}}\right)^2 \leq \lambda_i \leq \left(1 + \frac{c}{\sqrt{n}}\right)^2.$$

We can now bound the statistical distance between $g_x = \mathcal{N}\left(x, \frac{r^2}{n}H(x)^{-1}\right)$ and $g_y = \mathcal{N}\left(y, \frac{r^2}{n}H(y)^{-1}\right)$ by using Pinsker's inequality [1, p. 44], which gives that $\|g_x - g_y\|_1^2 \leq 2 \cdot \mathrm{D}_{\mathrm{KL}}(g_y \| g_x)$. Letting Σ_1, Σ_2 denote the covariance matrices of g_1, g_2, we can write $\mathrm{Tr}\left(\Sigma_1^{-1}\Sigma_2\right) = \sum_{i=1}^n \lambda_i$, and $\log \frac{\det \Sigma_1}{\det \Sigma_2} = \log \frac{1}{\det \Sigma_1^{-1}\Sigma_2} = \sum_{i=1}^n \log \frac{1}{\lambda_i}$.

$$\|g_x - g_y\|_1^2 \leq \sum_{i=1}^n \left(\lambda_i - 1 + \log\frac{1}{\lambda_i}\right) + \frac{n}{r^2}\|x-y\|_x^2$$

(Using Fact 5)

$$\leq \sum_{i=1}^n \left(\lambda_i + \frac{1}{\lambda_i} - 2\right) + \frac{n}{r^2}\|x-y\|_x^2$$

$\left(\text{Using } \log\frac{1}{\lambda} \leq \frac{1}{\lambda} - 1\right)$

$$\leq n \cdot \max\left\{\frac{(2c/\sqrt{n} - c^2/n)^2}{(1 - c/\sqrt{n})^2}, \frac{(2c/\sqrt{n} + c^2/n)^2}{(1 + c/\sqrt{n})^2}\right\} + \frac{n}{r^2}\cdot\frac{c^2}{n}.$$

$\left(\text{Using the convexity of } \lambda + \frac{1}{\lambda} - 2\right)$

$$\leq n \cdot \frac{(2c/\sqrt{n} - c^2/n)^2}{(1 - c/\sqrt{n})^2} + \frac{c^2}{r^2}$$
$$= c^2 \cdot \frac{(2 - c/\sqrt{n})^2}{(1 - c/\sqrt{n})^2} + \frac{c^2}{r^2} \leq \frac{25}{4}c^2 + c^2 \leq 9c^2,$$

where the last line uses $c \leq 1/3, r \leq 1$ and $n \geq 1$. □

3. The effect of the Metropolis filter

In this section, we prove Lemma 4 that shows that for any $x \in K$, the statistical distance between the Gaussian distribution g_x and the random walk distribution p_x, obtained by applying the Metropolis filter to g_x, is small. We have,

$$\left\|p_x(z) - g_x(z)\right\|_1 = 1 - \mathop{\mathbf{E}}_{z\sim g_x} \min\left\{1, \frac{g_z(x)}{g_x(z)}\right\}. \quad (1)$$

Given $\varepsilon \in (0, 1/2]$, we show that for an appropriate choice of r, the above statistical distance is bounded by ε.

The ratio of g_z and g_x has two terms: one involving the ratio of $\det H(x)$ and $\det H(z)$, and one involving the difference in local norms $\|z-x\|_z^2 - \|z-x\|_x^2$. Proposition 6 bounds the first by controlling the norm of $\nabla \log\det H(x)$. Proposition 7 bounds the second term by using concentration of Gaussian polynomials.

Proof of Lemma 4: We have,

$$\frac{g_z(x)}{g_x(z)} = \exp\left(-\frac{n}{2r^2}\left(\|z-x\|_z^2 - \|z-x\|_x^2\right) + \frac{1}{2}\left(\log\det H(z) - \log\det H(x)\right)\right).$$

From Proposition 6, for $r \leq \frac{\varepsilon}{4}(2\log 4/\varepsilon)^{-1/2}$, we have

$$\mathbf{Pr}[\log\det H(z) - \log\det H(x) \geq -\varepsilon/2] \geq 1 - \varepsilon/4.$$

Also, from Proposition 7, for $r \leq \frac{\varepsilon}{100}(\log 50/\varepsilon)^{-3/2}$, we have,

$$\mathbf{Pr}\left[\|z-x\|_z^2 - \|z-x\|_x^2 \leq \frac{\varepsilon}{2}\cdot\frac{r^2}{n}\right] \geq 1 - \varepsilon/4.$$

Combining the two using a union bound, we get that except with probability $\varepsilon/2$, we have, $\frac{g_z(x)}{g_x(z)} \geq e^{-\varepsilon/2} \geq 1 - \varepsilon/2$. Thus,

$$\mathop{\mathbf{E}}_{z\sim g_x} \min\left\{1, \frac{g_z(x)}{g_x(z)}\right\} \geq \left(1 - \frac{\varepsilon}{2}\right) \mathop{\mathbf{Pr}}_{z\sim g_x}\left[\frac{g_z(x)}{g_x(z)} \geq 1 - \frac{\varepsilon}{2}\right] \geq \left(1 - \frac{\varepsilon}{2}\right)^2 \geq 1 - \varepsilon.$$

The claim now follows from (1). □

Proposition 6. *Given $\varepsilon \in (0, 1/2]$, for $r \leq \frac{\varepsilon}{\sqrt{2\log 1/\varepsilon}}$, and $z \sim g_x$ we have*

$$\mathbf{Pr}[\log\det H(z) - \log\det H(x) \geq -2\varepsilon] \geq 1 - \varepsilon.$$

Proof of Proposition 6: Let $V(x) := \frac{1}{2}\log\det H(x)$. From the work of Vaidya [12], we know that $V(x)$ is a convex function. Thus, $V(z) - V(x) \geq (z-x)^\top \nabla V(x)$. We know that $z = x + \frac{r}{\sqrt{n}}\left(H(x)\right)^{-1/2} g$, where $g \sim \mathcal{N}(0, \mathbb{I}_n)$. Thus,

$$V(z) - V(x) \geq \frac{r}{\sqrt{n}} g^\top \left(H(x)\right)^{-1/2} \nabla V(x).$$

$g^\top \left(H(x)\right)^{-1/2} \nabla V(x)$ is a Gaussian with mean 0 and variance $\left\|\left(H(x)\right)^{-1/2} \nabla V(x)\right\|_2^2$. From Lemma 4.3 in the work of Vaidya and Atkinson [13], it follows that

$$\left\|\left(H(x)\right)^{-1/2} \nabla V(x)\right\|_2^2 \leq n.$$

Using standard tail bounds, we get that for all $\lambda > 0$,

$$\mathbf{Pr}[g^\top \left(H(x)\right)^{-1/2} \nabla V(x) \geq -\lambda\sqrt{n}] \geq 1 - \exp(-\lambda^2/2).$$

Picking $\lambda = \sqrt{2\log 1/\varepsilon}$, and combining, we get, $\mathbf{Pr}[V(z) - V(x) \geq -r\sqrt{2\log 1/\varepsilon}] \geq 1 - \varepsilon$. For $r \leq \frac{\varepsilon}{\sqrt{2\log 1/\varepsilon}}$, we have $r\sqrt{2\log 1/\varepsilon} \leq \varepsilon$, which gives the claim. □

Proposition 7. *Given $\varepsilon \in (0, 1/2]$, for $r \leq \frac{\varepsilon}{20}(\log 11/\varepsilon)^{-\frac{3}{2}}$, and $z \sim g_x$, we have,*

$$\mathbf{Pr}\left[\|z-x\|_z^2 - \|z-x\|_x^2 \leq 2\varepsilon \cdot \frac{r^2}{n}\right] \geq 1 - \varepsilon.$$

Proof Proposition 7: We have $z = x + \frac{r}{\sqrt{n}}\left(H(x)\right)^{-\frac{1}{2}} g$, where $g \sim \mathcal{N}(0, \mathbb{I}_n)$. If we let $\hat{a}_i = \frac{1}{a_i^\top x - b_i}\left(H(x)\right)^{-\frac{1}{2}} a_i$, we get $a_i^\top(z-x) = \frac{r}{\sqrt{n}}(a_i^\top x - b_i)\cdot \hat{a}_i^\top g$, and $\sum_{i=1}^m \hat{a}_i\hat{a}_i^\top = \mathbb{I}_n$.

$$\begin{aligned}
&\|z-x\|_z^2 - \|z-x\|_x^2 \\
&= \sum_{i=1}^m (a_i^\top(z-x))^2 \left(\frac{1}{(a_i^\top z - b_i)^2} - \frac{1}{(a_i^\top x - b_i)^2}\right) \\
&= \frac{r^2}{n}\sum_{i=1}^m (\hat{a}_i^\top g)^2 \left(\frac{1}{(1+\frac{r}{\sqrt{n}}\hat{a}_i^\top g)^2} - 1\right) \\
&= \frac{r^4}{n^2}\sum_{i=1}^m (\hat{a}_i^\top g)^4 \left(\frac{2}{(1+\frac{r}{\sqrt{n}}\hat{a}_i^\top g)} + \frac{1}{(1+\frac{r}{\sqrt{n}}\hat{a}_i^\top g)^2}\right) \\
&\qquad - \frac{2r^3}{n^{3/2}}\sum_{i=1}^m (\hat{a}_i^\top g)^3. \qquad (2)
\end{aligned}$$

We now use concentration of Gaussian polynomials (see Theorem 8) to bound the two terms above. Let $P_1(g) := \sum_{i=1}^m (\hat{a}_i^\top g)^3$. From Fact 10, we know $\mathbf{E}_g P_1(g)^2 \leq 15n$. Thus, using Theorem 8, we know that for any $\lambda_1 \geq (\sqrt{2e})^3$,

$$\mathop{\mathbf{Pr}}_g\left[|P_1(g)| \geq \lambda_1\sqrt{15n}\right] \leq \exp\left(-\frac{3}{2e}\lambda_1^{2/3}\right).$$

Picking $\lambda_1 = \left(\max\left\{2e, \frac{2e}{3}\log\frac{2}{\varepsilon}\right\}\right)^{3/2}$, and $r \leq \frac{\varepsilon}{2\sqrt{15}\lambda_1}$, we obtain, $\mathbf{Pr}\left[|P_1(g)| \geq \frac{\varepsilon}{2r}\sqrt{n}\right] \leq \frac{\varepsilon}{2}$. Thus, with probability at least $1 - \frac{\varepsilon}{2}$,

$$-\frac{2r^3}{n^{3/2}}\sum_{i=1}^m (\hat{a}_i^\top g)^3 \leq \frac{2r^3}{n^{3/2}}\cdot\frac{\varepsilon}{2r}\sqrt{n} = \varepsilon\cdot\frac{r^2}{n}. \qquad (3)$$

Now, we let $P_2(g) := \sum_{i=1}^m (\hat{a}_i^\top g)^4$. Again, from Fact 10, we know that $\mathbf{E}_g P_2(g)^2 \leq 105n^2$, and applying Theorem 8, we obtain that for $\lambda_2 = \left(\max\left\{2e, \frac{2e}{4}\log\frac{2}{\varepsilon}\right\}\right)^2$ and $r \leq \frac{\sqrt{\varepsilon}}{\sqrt{8\lambda_2\sqrt{105}}}$, we obtain, $\mathbf{Pr}\left[|P_2(g)| \geq \frac{\varepsilon}{8r^2}n\right] \leq \frac{\varepsilon}{2}$. Thus, with probability at least $1 - \frac{\varepsilon}{2}$.

$$\frac{r^4}{n^2}\sum_{i=1}^m (\hat{a}_i^\top g)^4 \leq \frac{r^4}{n^2}\cdot\frac{\varepsilon}{8r^2}n = \frac{\varepsilon}{8}\cdot\frac{r^2}{n}.$$

Note that this also implies that for all i, $\frac{r}{\sqrt{n}}|\hat{a}_i^\top g| \leq \left(\frac{\varepsilon r^2}{8n}\right)^{1/4} \leq \frac{1}{2}$, where the last inequality holds for all $r \leq 1$. Thus, with probability at least $1 - \frac{\varepsilon}{2}$, we have

$$\begin{aligned}
&\frac{r^4}{n^2}\sum_{i=1}^m (\hat{a}_i^\top g)^4 \left(\frac{2}{(1+\frac{r}{\sqrt{n}}\hat{a}_i^\top g)} + \frac{1}{(1+\frac{r}{\sqrt{n}}\hat{a}_i^\top g)^2}\right) \\
&\qquad \leq 8\frac{r^4}{n^2}\sum_{i=1}^m (\hat{a}_i^\top g)^4 \leq \varepsilon\cdot\frac{r^2}{n}.
\end{aligned}$$

Combining this with Equations (2) and (3), and applying a union bound, we get that with probability at least $1-\varepsilon$

$$\|z-x\|_z^2 - \|z-x\|_x^2 \leq 2\varepsilon\cdot\frac{r^2}{n}.$$

Finally, we verify that for $\varepsilon \in (0, 1/2]$, any $r \leq \frac{\varepsilon}{20}(\log 11/\varepsilon)^{-\frac{3}{2}}$ satisfies the conditions

$$r \leq \min\left\{1, \frac{\varepsilon}{2\sqrt{15}\lambda_1}, \frac{\sqrt{\varepsilon}}{\sqrt{8\lambda_2\sqrt{105}}}\right\}.$$

□

Theorem 8. ***(see Janson [5, Thm 6.7])*** *Let $P(g)$ be a degree q polynomial, where $g \in \mathbb{R}^n$ such that $g \sim \mathcal{N}(0, \mathbb{I}_n)$. Then, for any $t \geq \sqrt{2e}^q$, we have,*

$$\mathop{\mathbf{Pr}}_g\left[|P(g)| \geq t\left(\mathbf{E}\,P(g)^2\right)^{1/2}\right] \leq \exp\left(-\frac{q}{2e}t^{2/q}\right).$$

Acknowledgements

The work of the first author was supported by a Simons Investigator Award to Daniel Spielman.

References

References

[1] Csiszár, I., Körner, J., 2011. Information Theory: Coding Theorems for Discrete Memoryless Systems. Cambridge University Press.
URL https://books.google.com/books?id=2gsLkQlb8JAC

[2] Dikin, I. I., 1967. Iterative solution to problems of linear and quadratic programming. Doklady Akademii Nauk SSSR 174 (4), 747.

[3] Dyer, M., Frieze, A., Kannan, R., Jan. 1991. A random polynomial-time algorithm for approximating the volume of convex bodies. J. ACM 38 (1), 1–17.
URL http://doi.acm.org/10.1145/102782.102783
[4] Isserlis, L., 1918. On a formula for the product-moment coefficient of any order of a normal frequency distribution in any number of variables. Biometrika 12 (1/2), 134–139.
URL http://www.jstor.org/stable/2331932
[5] Janson, S., 1997. Gaussian Hilbert Spaces. Cambridge University Press.
URL http://dx.doi.org/10.1017/CBO9780511526169
[6] Kannan, R., Narayanan, H., 2012. Random walks on polytopes and an affine interior point method for linear programming. Mathematics of Operations Research 37 (1), 1–20.
URL http://dx.doi.org/10.1287/moor.1110.0519
[7] Karmarkar, N., 1984. A new polynomial-time algorithm for linear programming. Combinatorica 4 (4), 373–395.
URL http://dx.doi.org/10.1007/BF02579150
[8] Lovász, L., 1999. Hit-and-run mixes fast. Mathematical Programming 86 (3), 443–461.
URL http://dx.doi.org/10.1007/s101070050099
[9] Narayanan, H., 02 2016. Randomized interior point methods for sampling and optimization. Ann. Appl. Probab. 26 (1), 597–641.
URL http://dx.doi.org/10.1214/15-AAP1104
[10] Nesterov, Y., Nemirovskii, A., 1994. Interior-point polynomial algorithms in convex programming. Vol. 13. SIAM.
[11] Renegar, J., 2001. A Mathematical View of Interior-Point Methods in Convex Optimization. Society for Industrial and Applied Mathematics.
URL http://epubs.siam.org/doi/abs/10.1137/1.9780898718812
[12] Vaidya, P. M., 1996. A new algorithm for minimizing convex functions over convex sets. Mathematical Programming 73 (3), 291–341.
URL http://dx.doi.org/10.1007/BF02592216
[13] Vaidya, P. M., Atkinson, D. S., 1993. A technique for bounding the number of iterations in path following algorithms. Complexity in Numerical Optimization, 462–489.

Appendix A. Relating the local metric to the Hilbert metric

Lemma 9 ([6]). *For any $x, y \in K$, we have $\sigma(x,y) \geq \frac{1}{\sqrt{m}} \|x-y\|_x$.*

Proof Lemma 9: Let p, x, y, q be the points in order on the chord of K that passes through x, y with p, q being the end-points of the chord. Thus,

$$\begin{aligned}\sigma(x,y) = \frac{|x-y||p-q|}{|p-x||q-y|} &\geq \max\left\{\frac{|x-y|}{|p-x|}, \frac{|x-y|}{|x-q|}\right\} \\ &= \max_{i\in[m]} \frac{|a_i^\top (x-y)|}{(a_i^\top x - b_i)} \\ &\geq \frac{1}{\sqrt{m}} \left(\sum_{i\in[m]} \frac{(a_i^\top (x-y))^2}{(a_i^\top x - b_i)^2}\right)^{1/2} \\ &= \frac{1}{\sqrt{m}} \|x-y\|_x .\end{aligned}$$

□

Appendix B. Moments of Gaussian Polynomials

Fact 10. *Suppose $g \in \mathbb{R}^n$ is distributed according to $\mathcal{N}(0, \mathbb{I}_n)$, and $\sum_{i=1}^m b_i b_i^\top = \mathbb{I}_n$. Then, we have,*

$$\mathbf{E}\left(\sum_{i=1}^m (b_i^\top g)^3\right)^2 \leq 15n, \quad \textit{and} \quad \mathbf{E}\left(\sum_{i=1}^m (b_i^\top g)^4\right)^2 \leq 105n^2.$$

Proof of Fact 10: We first consider the first part of the fact. From Fact 11, we know that for all i, j,

$$\mathop{\mathbf{E}}_g (b_i^\top g)^3 (b_j^\top g)^3 = 9 \|b_i\|^2 \left\|b_j\right\|^2 (b_i^\top b_j) + 6(b_i^\top b_j)^3.$$

Summing over all i, j, we get,

$$\begin{aligned}\mathbf{E}\left(\sum_{i=1}^m (b_i^\top g)^3\right)^2 &= \sum_{i,j=1}^m \mathbf{E}(b_i^\top g)^3 (b_j^\top g)^3 \\ &= 9\sum_{i,j=1}^m \|b_i\|_2^2 \left\|b_j\right\|_2^2 (b_i^\top b_j) + 6\sum_{i,j=1}^m (b_i^\top b_j)^3. \quad \text{(B.1)}\end{aligned}$$

This equality can also be derived using Isserlis' theorem ([4]). If we let B be the $m \times n$ matrix with its i^{th} row being $b_i^\top$, and $w \in \mathbb{R}^m$ be such that $w_i = \|b_i\|_2^2$, we can simplify the first term in the above sum as follows.

$$\sum_{i,j=1}^m \|b_i\|_2^2 \left\|b_j\right\|_2^2 (b_i^\top b_j) = \left\|\sum_{i=1}^m \|b_i\|^2 b_i\right\|_2^2 = \left\|B^\top w\right\|_2^2.$$

Using $\sum_{i=1}^m b_i b_i^\top = \mathbb{I}_n$, we get $B^\top B = \mathbb{I}_n$. Thus, the $m \times m$ matrix $\Pi := BB^\top$ satisfies $\Pi^2 = \Pi$. Since Π is also symmetric, it is an orthogonal projection. Thus, we have $\|\Pi w\|_2 \leq \|w\|_2$. We obtain,

$$\begin{aligned}\left\|B^\top w\right\|_2^2 = w^\top BB^\top w &= w^\top \Pi w \\ &= w^\top \Pi^2 w \\ &= \|\Pi w\|_2^2 \leq \|w\|_2^2 = \sum_{i=1}^m \|b_i\|^4 .\end{aligned}$$

Since $\sum_{i=1}^m b_i b_i^\top = \mathbb{I}_n$, we get that for all i, $\|b_i\| \leq 1$. Moreover, taking trace, we obtain $\sum_{i=1}^m \|b_i\|^2 = n$. Thus,

$$\sum_{i=1}^m \|b_i\|^4 \leq \sum_{i=1}^m \|b_i\|^2 = n.$$

Thus, we can bound the first term in Equation (B.1) by n.

For the second term in Equation (B.1), using $\|b_i\|_2 \leq 1$ for all i, and Cauchy-Schwarz, we get $|b_i^\top b_j| \leq 1$ for all i, j. Using $\sum_{i=1}^m b_i b_i^\top = \mathbb{I}_n$, we also know that for all j, $\sum_{i=1}^m (b_i^\top b_j)^2 = \left\|b_j\right\|_2^2$. Thus, we get,

$$\sum_{i,j=1}^m (b_i^\top b_j)^3 \leq \sum_{i,j=1}^m (b_i^\top b_j)^2 = \sum_{j=1}^m \left\|b_j\right\|_2^2 = n.$$

Combining the bounds for the two terms in Equation (B.1), we get the first part of the fact.

For the second part of the fact, we use Cauchy-Schwarz inequality,

$$\mathbf{E}\left(\sum_{i=1}^{m}(b_i^\top g)^4\right)^2 = \sum_{i,j=1}^{m}\mathbf{E}(b_i^\top g)^4(b_j^\top g)^4 \leq \sum_{i,j=1}^{m}\left(\mathbf{E}(b_i^\top g)^8\right)^{1/2}\left(\mathbf{E}(b_j^\top g)^8\right)^{1/2}.$$

We have that $b_i^\top g$ is distributed as a Gaussian with mean 0 and variance $\|b_i\|_2^2$. Thus, $\mathbf{E}(b_i^\top g)^8 = 105\,\|b_i\|_2^8$. Hence, we get,

$$\mathbf{E}\left(\sum_{i=1}^{m}(b_i^\top g)^4\right)^2 \leq 105\sum_{i,j=1}^{m}\|b_i\|_2^4\,\|b_j\|_2^4 = 105\left(\sum_{i=1}^{m}\|b_i\|^4\right)^2 \leq 105n^2,$$

proving the second part of the fact. □

Note: The bounds given by the above fact are tight for the case where b_i form an orthonormal basis.

Fact 11 (Isserlis [4]). *Suppose $g \in \mathbb{R}^n$ is distributed according to $\mathcal{N}(0, \mathbb{I}_n)$, and b_1, b_2 are any two vectors in $\mathbb{R}^n$,*

$$\mathop{\mathbf{E}}_{g}(b_1^\top g)^3(b_2^\top g)^3 = 9\,\|b_1\|^2\,\|b_2\|^2\,(b_1^\top b_2) + 6(b_1^\top b_2)^3.$$

Proof of Fact 11: We define $\hat{b}_i := \frac{1}{\|b_i\|}\cdot b_i$ to be the corresponding unit vectors. Thus,

$$\mathbf{E}(b_1^\top g)^3(b_2^\top g)^3 = \|b_1\|^3\,\|b_2\|^3\,\mathbf{E}(\hat{b}_1^\top g)^3(\hat{b}_2^\top g)^3. \qquad \text{(B.2)}$$

We let $e_1, \ldots, e_n$ denote the standard basis vectors for $\mathbb{R}^n$, *i.e.*, e_i is 1 in the i^{th} coordinate and 0 elsewhere. Since the distribution of g is rotationally symmetric, we can assume that $\hat{b}_1 = e_1$, and $\hat{b}_2 = \cos\theta\cdot e_1 + \sin\theta\cdot e_2$, where θ is such that $\cos\theta = \hat{b}_1^\top\hat{b}_2$. Thus,

$$\begin{aligned}\mathbf{E}(\hat{b}_1^\top g)^3(\hat{b}_2^\top g)^3 &= \mathbf{E}\,g_1^3\,(\cos\theta\cdot g_1 + \sin\theta\cdot g_2)^3\\ &= \cos^3\theta\,\mathbf{E}\,g_1^6 + 0 + 3\cos\theta\sin^2\theta\,\mathbf{E}\,g_1^4\,\mathbf{E}\,g_2^2 + 0\\ &= 15\cos^3\theta + 9\cos\theta\sin^2\theta = 9\cos\theta + 6\cos^3\theta\\ &= 9\left(\hat{b}_1^\top\hat{b}_2\right) + 6\left(\hat{b}_1^\top\hat{b}_2\right)^3.\end{aligned}$$

Combining with Equation (B.2), we obtain the fact. □