# A fast sampling algorithm based on volumetric-logarithmic barrier

Yuansi Chen[†], Raaz Dwivedi[⋆], Martin J. Wainwright[†,⋆], Bin Yu[†,⋆]

Department of Statistics[†], and
Department of Electrical Engineering and Computer Sciences[⋆]
UC Berkeley, Berkeley, CA 94720

July 9, 2017

**Abstract**

We propose a new random walk known as Vaidya walk, to generate samples from the uniform distribution on the interior of a polytope. The random walk is an instance of sampling algorithm derived from an interior point method and is based on the volumetric-logarithmic barrier introduced by Vaidya. We show that Vaidya walk mixes in significantly fewer steps compared to Dikin walk, a random walk introduced by Kannan and Narayanan which was based on the logarithmic barrier method. In particular, we prove that for a polytope in $\mathbb{R}^n$ defined by $m$ constraints, the new random walk mixes in $\mathcal{O}\left(\sqrt{m/n}\right)$ fewer steps compared to Dikin walk. The per iteration cost for our method is at most twice that of Dikin walk, and hence the speed up is significant for polytopes with $m \gg n$. Furthermore, the algorithm is also faster than Ball walk and Hit-and-Run for a large family of polytopes. We illustrate the speed-up compared to Dikin walk via several numerical examples and discuss possible new and faster algorithms for sampling from polytopes.

## 1 Introduction

Sampling from distributions is ubiquitous in statistics for the tasks of estimation, prediction, and inference. Several integration problems are reduced to estimating expectations or probability of some event, under a particular probability distribution. Random samples drawn from that distribution are then used to estimate these quantities. Markov Chain Monte Carlo (MCMC) Algorithms form a broad class of algorithms that are employed for the task of sampling from distributions. An advantage of these algorithms is that they require the knowledge of the density of the target distribution only up to proportionality. It is well known that computing normalization constant is often a non-trivial task, e.g., for Bayesian posteriors, and hence MCMC methods come in handy in such scenarios. The downside of these algorithms is the poor convergence rate observed in practice. In fact, the theoretical understanding of approximate sampling algorithms is far from satisfactory. Nevertheless, continuous study of these algorithms has resulted in the development of many mathematical and algorithmic insights useful in practice.

Sampling algorithms have been used to solve another fundamental problem in computer science—estimation of the volume of a convex set. While it is NP-hard for any deterministic algorithm to estimate the volume of a convex set even up to exponential in dimension factors [Ele86, BF87], several works [LS90, LS92, LS93, LV06c, CV14] have led to fast polynomial-time randomized algorithms for this task. Several random walks, e.g., Ball Walk, Hit-and-Run and Dikin walk, have been invented over the years that can generates samples from log-concave distributions on a convex set. We discuss these random walks in detail in Section 2.

Further, sampling algorithms have been used to derive fast optimization algorithms in the past [DFK91, KV06, LV06a, NR10]. More recently, use of sampling algorithms for nonconvex optimization, especially to escape saddle-points, is increasingly becoming popular [RRT17, ZLC17]. Recently there has been work exploring the flip side of the relation between sampling and optimization, i.e., theoretical understanding of sampling algorithms that have been developed by drawing inspiration from iterative optimization algorithms. The classical example in this category is Langevin Monte Carlo (LMC) method invented for spinodal simulations in the field of chemical physics [MMPS83]. In the earlier theoretical works [RT96, RR01], this method was analyzed as a discretization of a stochastic differential equation. More recently [Dal16, DM16] it has become the classical algorithm that can be seen as a noisy optimization algorithm. Stochastic Gradient MCMC [WT11, ABW12] and projected LMC [BEL15] are among a few other instances of the sampling algorithms that can be seen as randomized versions of *first-order* optimization algorithms. Our work focuses on random walks based on interior point methods. The first such algorithm, Dikin walk [KN12] was designed to sample uniformly from a bounded convex set defined by linear constraints in polynomial time. In particular, for a polytope in $\mathbb{R}^n$ defined by $m$ linear constraints, the walk was shown to have a convergence rate of $\mathcal{O}\left(mn\right)$ from a "good start". Later [Nar16], the walk was extended with modifications to have polynomial time guarantees for general convex sets equipped with self-concordant barriers.

In this work, we add one more bond between sampling and optimization. We show that a faster interior point method can be tweaked to obtain a faster sampling algorithm. While Dikin walk was based on the classical log-barrier method, we introduce and analyze a new random walk—*Vaidya walk*—based on the optimization algorithm introduced by Vaidya [Vai89]. We show that for a polytope defined by $m$-constraints, our walk has a convergence rate of $\mathcal{O}\left(m^{1/2}n^{3/2}\right)$, i.e., $\sqrt{m/n}$ faster than Dikin walk. We also show that per iteration complexity of the two random walks differs only by constant (independent of $m, n$) factors. Thus, in the regime $m \gg n$ the overall complexity of generating an approximately uniform sample can be much smaller. We remark that the connection between sampling and optimization algorithms is still in nascent stage and it is neither straightforward to provide a recipe to convert an arbitrary optimization algorithm to a sampling algorithm, nor to argue if a faster optimization algorithm can be tweaked to obtain a faster sampling algorithm.

The remainder of the paper is organized as follows. In Section 2, we discuss some polynomial-time random walks on convex sets and polytopes, and motivate the starting point for our random walk. In Section 3, we formally introduce Vaidya walk and provide the rate of convergence for the walk. We demonstrate the contrast between Dikin and Vaidya walk for some illustrative examples in Section 3.3. We provide the details of the analysis for Vaidya walk and differences with Dikin walk in Section 4 and defer some more technical lemmas to the appendices. We conclude with discussion about applications of sampling from polytopes and possible extensions of our work in Section 5.

**Notation:** For two sequences $a_n$ and $b_n$, we say that $a_n = \mathcal{O}\left(b_n\right)$ if there exists a universal constant $C > 0$ such that $a_n \leq Cb_n$. For a set $\mathcal{K} \subset \mathbb{R}^n$, the sets $\text{int}\left(\mathcal{K}\right)$ and $\mathcal{K}^c$ denote the interior and complement of $\mathcal{K}$ respectively. Also $\gamma_{\mathcal{K}}$ denotes the condition number of the set $\mathcal{K}$. We denote the boundary of the set $\mathcal{K}$ by $\partial\mathcal{K}$. In particular, if the set $\mathcal{K}$ is contained in a ball of radius $R_{\min}$ and is contained in a ball of radius $R_{\max}$, then $\gamma_{\mathcal{K}} \leq R_{\max}/R_{\min}$. The Euclidean norm for any vector $x \in \mathbb{R}^n$ is denoted by $\left\|x\right\|_2$. For any square matrix $M$, we use $\det(M)$ and $\text{trace}(M)$ to denote the determinant and the trace of $M$ respectively. For two distributions $\mathcal{P}_1$ and $\mathcal{P}_2$ defined on the same probability space $(\mathcal{X}, \mathbb{B}(\mathcal{X}))$, we denote the total-variation (TV) distance between the two by $\|\mathcal{P}_1 - \mathcal{P}_2\|_{\mathrm{TV}}$. Furthermore, $\mathrm{KL}(\mathcal{P}_1 \| \mathcal{P}_2)$ denotes the Kullback-Leibler (KL) Divergence between the two distributions.

## 2 Background and problem setting

In this section, we provide a description of general MCMC algorithms and an overview of the rates of convergence of existing random walks on convex sets. We end the section by pointing the starting point for Vaidya walk.

### 2.1 Metropolis-Hastings Algorithms

The walk is an instance of Hastings and Metropolis algorithms, which were first introduced by Metropolis [MRR+53] in 1953 and Hastings [Has70] in 1970 for speeding up simulations in Chemical Physics. Since then, the class of these algorithms has developed to a great extent. For a more detailed introduction and discussion to these algorithms, we refer the readers to the books [Rob04, BGJM11] and the references therein. Here, we briefly describe standard notation for Markov chain and the standard construction of a Markov chain based on an algorithm from this family.

For a Markov chain $\{X_0, X_1, \ldots,\}$ on $\mathcal{X}$ with transition kernel $\mathcal{Q}$ and initial distribution $\mu_0$, we use $\mu_0 \mathcal{Q}^k$ to denote the probability distribution of its $k$-th iterate. Further, if the chain has a unique invariant distribution $\Pi$, then we define its $\delta$-mixing time as

$$k_{\mathrm{mix}}(\delta) := \min \left\{ k \middle| \left\| \mu_0 \mathcal{Q}^k - \Pi \right\|_{\mathrm{TV}} \leq \delta \right\}. \tag{1}$$

A distribution $\mathcal{P}_1$ (with density $p_1$) is said to be $M$-warm with respect to distribution $\mathcal{P}_2$ (with density $p_2$), if

$$\sup_{A \in \mathbb{B}(\mathcal{X})} \left( \frac{\mathcal{P}_1(A)}{\mathcal{P}_2(A)} \right) = \sup_{x \in \mathcal{X}} \left( \frac{p_1(x)}{p_2(x)} \right) \leq M.$$

In a typical MCMC algorithm, first we use a transition kernel with density $p(x, z), x, z \in \mathcal{X}$ to generate candidate proposals for a discrete-time Markov Chain with state space $\mathcal{X}$. The candidate proposals $z$ at $x$ are generated using the density $p(x, \cdot)$. To ensure, the stationary distribution of the chain is the "desired" target distribution $\Pi$ with density $\pi(\cdot)$, a proposal $z$ is then accepted with probability

$$\alpha(x, z) = \min \left\{ 1, \frac{\pi(z) p(z, x)}{\pi(x) p(x, z)} \right\},$$

and with probability $1 - \alpha(x, z)$ the chain stays at $x$. Thus, the actual transition kernel for the Markov chain, that we denote by $\mathcal{Q}$, is defined by a density given by

$$q(x, z) = p(x, z) \alpha(x, z) \text{ for } z \neq x,$$

and a probability mass at $x$, given by

$$\mathcal{Q}(x, \{x\}) = 1 - \int_{\mathcal{X}} p(x, z) \alpha(x, z) dz.$$

It is easy to verify that $\Pi$ distribution satisfies the detailed balanced condition, i.e.,

$$\pi(x)q(x,y) = \pi(y)q(y,x) \quad \text{for all } x, y \in \mathcal{X},$$

and hence $\Pi$ is a valid stationary distribution. In general, the existence and uniqueness of the stationary distribution of a general state space Markov chain is a technical issue. We use a *lazy* random walk as a simple guarantee for unique stationary distribution. For any random walk, its lazy version can be defined as follows: when at state $x$ with probability $1/2$ the walk stays at $x$ and with probability $1/2$ it makes a transition as per the original random walk. Having set up the chain in this fashion, the rate of convergence are conventionally reported in terms of $t_{\text{mix}}(\delta)$. One can also study convergence in $L_2$ distance or Wasserstein distance of order $p \geq 1$. We say that the Markov chain has a warm start, if $\mu_0$ is $M$-warm with respect to $\Pi$ for some $M < \infty$.

## 2.2 Sampling from convex sets

We now describe some random walks tailored to generating samples from approximate uniform distribution on bounded convex sets. We say that Markov chain mixes in $\mathcal{O}\left(f(\delta)\right)$ steps to mean that for any $\delta \in (0,1)$, we have $k_{\text{mix}}(\delta) = \mathcal{O}\left(f(\delta)\right)$. The general problem of sampling from convex body given by a membership oracle has seen significant progress [LS90, LS93, Lov99, LV06b, LV06a, LV06c, LV07]. In such a problem, for each point $x \in \mathbb{R}^n$, one can query an oracle that answers Yes/No depending on whether $x \in \mathcal{K}$ or not. The complexity of these algorithms is measured in terms of the number of oracle calls to obtain an approximate sample from the target distribution on $\mathcal{K}$. In Ball walk [LS90], when at point $x$ one generates a uniform point $u$ from a ball of radius $r$ centered at $x$, where $r$ denotes the step size of the algorithm. If $u \in \mathcal{K}$, the walk moves to $u$ else remains at $x$. This walk mixes in $\mathcal{O}\left(M^2 n \gamma_{\mathcal{K}}^2 \log(M/\delta)/(\delta^2 r^2)\right)$ steps from an $M$-warm start, provided $r < 1/\sqrt{n}$. In Hit-and-Run [Lov99], when at point $x$, we draw a uniform line $\ell$ and sample a point uniformly from the intersection $\ell \cap \mathcal{K}$. From an $M$-warm start, Hit-and-Run mixes in $\mathcal{O}\left(n^2 \gamma_{\mathcal{K}}^2 \log^3(M/\epsilon)\right)$ steps. Using the standard accept-reject step described in Section 2.1, these algorithms can be adapted for more general continuous distributions which admit a density, logarithm of which is a concave function. Such distributions are more commonly known as log-concave distributions. The algorithms described so far are *zeroth-order* in nature, i.e., they query only the value of the (unnormalized) density at an arbitrary point and if the point belongs to the set $\mathcal{K}$. For uniform sampling on $n$-dimensional polytopes defined by $m$-linear constraints, ball walk simply requires to query if the new point is inside $\mathcal{K}$ and hence the per iteration complexity is equivalent to a matrix-vector product, i.e., $\mathcal{O}\left(mn\right)$. Besides, Hit-and-run needs to also find a radius such that the chord $\ell$ is completely inside $\mathcal{K}$ and using binary search one can show that the per iteration complexity is $\mathcal{O}\left(mn \log \gamma_{\mathcal{K}}\right)$.

We now describe a *first-order* sampling method for bounded convex sets—Projected Langevin Monte Carlo [BEL15]—that queries first order information, namely the value and the gradient of the density at an arbitrary point. For a log-concave distribution, for a given state of the random walk the next state is obtained in three steps: (1) First, take a gradient ascent step on the log-likelihood at the current state, (2) then add a scaled isotropic Gaussian noise, and (3) finally compute the orthogonal projection onto $\mathcal{K}$ of the resultant vector after step (2). The variance of the noise added along each direction is of the same order of the step size of the gradient ascent to ensure that the walk can explore the space. Although, for any fixed step size the random walk converges to a slightly biased distribution, still by suitably choosing the step size the random walk can be made to generate samples approximately from uniform distribution in $\mathcal{O}\left(\gamma_{\mathcal{K}}^{6} n^{7}/\delta^{8}\right)$. The per iteration complexity for the method is equal to the sum of the complexity of performing the gradient ascent step and finding the orthogonal projection of a point on a convex set. The later step is equivalent to solving a constrained least-squares problem, and hence the efficient polynomial time ellipsoidal algorithms can be used.

Our work focuses on sampling uniformly from polytopes. Dikin walk is one such algorithm designed to sample uniformly from polytopes [KN12]. Later on Narayanan [Nar16] defined a random walk on Riemannian manifold for the case when the state space is a convex set with a self-concordant barrier [Nar16]. These random walks are instances of randomized interior point methods. Dikin walk [KN12] proceeds by proposing a uniform point in a suitable *state-dependent ellipsoid* followed by an accept-reject step. This algorithm is similar to ball walk except that state dependent ellipsoids are used in place of fixed Euclidean balls to generate proposals. Dikin walk successfully adapts to the boundary and the ellipsoid of proposal remains inside $\mathcal{K}$ for all $x \in \mathcal{K}$, unlike ball walk where we need to reduce the radius of the proposal ball to ensure we do not have have exponential number of rejections near the boundary. Further, the walk is affine invariant and consequently does not depend on the condition number of the set $\mathcal{K}$. In the equivalent version [Nar16], uniform proposals in the ellipsoid are replaced by Gaussian proposals with *suitable covariance* such that the high probability proposal ellipsoid stays inside the polytope for each $x \in \mathcal{K}$. For polytopes $\mathcal{K} = \{x \in \mathbb{R}^n \mid Ax \leq b\}$ where the matrix $A \in \mathbb{R}^{m \times n}$ and the vector $b \in \mathbb{R}^n$ are known, the mixing time of Dikin walk was proven to be $\mathcal{O}(mn \log(M/\delta))$ from $M$-warm start. Note that the set $\mathcal{K}$ is bounded and hence the matrix $A$ has full-rank $n \leq m$. Strictly speaking, though this algorithm can not be categorized as either zeroth-order or first-order sampling algorithm. The per iteration computation complexity for this random walk is equivalent to solving an $m \times n$ linear system of equations, which is $\mathcal{O}\left(mn^{\omega-1}\right)$ where $\omega < 2.38$ for the state-of-the-art method [DDH07]. To motivate the covariance used in Dikin walk, we outline the problem of optimizing a convex function on a polytope. Let $a_i^\top$ denote the $i$-th row vector of matrix $A$. Consider the *logarithmic-barrier* for the polytope $\mathcal{K}$ given by

$$F_x := F(x) = -\sum_{i=1}^{m} \log(b_i - a_i^T x).$$

In an interior point method, the inverse of the Hessian of the log barrier is used at each step to move to a new point in the polytope, where the Hessian is given by

$$\nabla^2 F_x := \nabla^2 F(x) \ = \sum_{i=1}^{m} \frac{a_i a_i^\top}{(b_i - a_i^\top x)^2}.$$

(See, e.g., Chapter 11 of the book [BV04] for introduction to interior point methods and the book [NN94] for a detailed discussion.) Narayanan et al. [KN12] designed the random walk with proposal ellipsoid at point $x$ (or scaled inverse covariance in the paper [Nar16]) given by $D_x := \nabla^2 F_x$.

In this work, we use a different interior point algorithm to design the new random walk. In Vaidya walk, the Gaussian proposals at $x$ are generated using the following inverse covariance matrix (up to scaling)

$$V_x := \sum_{i=1}^{m} (\sigma_{x,i} + \beta) \frac{a_i a_i^\top}{(b_i - a_i^\top x)^2}, \tag{2}$$

where $\beta := n/m$ and $\sigma_x$ denotes the leverage scores assciated with the matrix $\nabla^2 F_x$ and is defined as

$$\sigma_x := \left( \frac{a_1^\top (\nabla^2 F_x)^{-1} a_1}{(b_1 - a_1^\top x)^2}, \ldots, \frac{a_m^\top (\nabla^2 F_x)^{-1} a_m}{(b_m - a_m^\top x)^2} \right)^\top \qquad \text{for } x \in \text{int}\,(\mathcal{K})\,. \tag{3}$$

The covariance matrix $V_x$ is related to the Hessian of the following combination of *volumetric barrier* and the logarithmic barrier

$$\mathfrak{F}_x := \log\det \nabla^2 F_x + \beta F_x. \tag{4}$$

In particular, the quadratic forms $v^\top V_x v$ and $v^\top \nabla^2 \mathfrak{F}_x v$ satisfy the condition

$$\forall v \in \mathbb{R}^n, \quad 5 v^\top \nabla^2 \mathfrak{F}_x v \geq v^\top V_x v \geq v^\top \nabla^2 \mathfrak{F}_x v.$$

The barrier (4) was introduced by Vaidya et al. [Vai89, VA93] and they proved superior convergence rates for interior point methods defined with this new barrier compared to methods defined with the log-barrier.

More recently, a random walk on Riemannian manifolds—geodesic walk [LV16]—was introduced to sample from uniform distribution on polytopes. The geodesic paths bend away from the boundary which allows the walk to take large steps while still staying inside the polytope. From a warm start, geodesic walk has an $\mathcal{O}\left(mn^{3/4}\right)$ mixing time, thereby breaking the quadratic barrier on mixing times.

We summarize the rates and per step complexity for different random walks in Table 1. Except for ball walk and Hit-and-run, all the random walks have a per iteration complexity of the order of linear-system solver. Often the guarantees for ball-walk and Hit-and-run are presented for the case when the set is isotropic in which case $\gamma_{\mathcal{K}} = \mathcal{O}\left(\sqrt{n}\right)$. In the rightmost column of the Table 1, we mention the overall complexity of different random walks for sampling from isotropic convex sets. We remark that making a set isotropic is closely interconnected with sampling, as uniform samples from the set are used to bring it in an isotropic position. However, Lovász and Vempala [LV07] provide an $\mathcal{O}\left(\text{poly-log}(n) n^5\right)$ algorithm that brings an arbitrary convex set to an approximately isotropic position using a membership oracle.

# 3 Vaidya walk and convergence

In this section, we provide the details of Vaidya walk, provide its rate of convergence and illustrate its performance via several simulated examples.

## 3.1 Vaidya walk

With the basic background in place, we now formally describe the random walk with the new covariance structure. Vaidya walk with parameter $r$ (VW($r$)) is defined with a proposal distribution at $x$ given by

$$\mathcal{P}_x := \mathcal{P}_x^r = \mathcal{N}\left(x, \frac{r^2}{\sqrt{mn}} {V_x}^{-1}\right)$$

Hence the proposal density is given by

$$p_x(z) := p(x,z) = \sqrt{\det V_x} \left(\frac{mn}{2\pi r^2}\right)^{n/2} \exp\left(-\frac{\sqrt{mn}}{2r^2}\,(z-x)^\top V_x (z-x)\right). \tag{5}$$

| **Random walk** | $\mathbf{k_{mix}}(\delta)$ | **Iteration cost** | **Complexity for** $\gamma_{\mathcal{K}} = \mathcal{O}\left(\sqrt{n}\right)$ |
|---|---|---|---|
| Ball walk | $\mathcal{O}\left(n^2\gamma_{\mathcal{K}}^2 \frac{M^2}{\delta^2} \log \frac{M}{\delta}\right)$ | $\mathcal{O}\left(mn\right)$ | $\mathcal{O}\left(mn^4 \cdot \frac{M^2}{\delta^2} \log \frac{M}{\delta}\right)$ |
| Hit-and-Run | $\mathcal{O}\left(n^2\gamma_{\mathcal{K}}^2 \log^3 \frac{M}{\delta}\right)$ | $\mathcal{O}\left(mn \log \gamma_{\mathcal{K}}\right)$ | $\mathcal{O}\left(mn^4 \log n \cdot \log \frac{M}{\delta}\right)$ |
| Dikin walk | $\mathcal{O}\left(mn \log \frac{M}{\delta}\right)$ | $\mathcal{O}\left(mn^{\omega-1}\right)$ | $\mathcal{O}\left(m^2 n^{\omega} \cdot \log \frac{M}{\delta}\right)$ |
| Geodesic walk | $\mathcal{O}\left(mn^{3/4} \log \frac{M}{\delta}\right)$ | $\mathcal{O}\left(mn^{\omega-1}\right)$ | $\mathcal{O}\left(m^2 n^{\omega-1/4} \cdot \log \frac{M}{\delta}\right)$ |
| Vaidya walk | $\mathcal{O}\left(m^{1/2} n^{3/2} \log \frac{M}{\delta}\right)$ | $\mathcal{O}\left(mn^{\omega-1}\right)$ | $\mathcal{O}\left(m^{3/2} n^{\omega+1/2} \cdot \log \frac{M}{\delta}\right)$ |

**Table 1.** Computational complexity of random walks from $M$-warm start on polytope $\mathcal{K} = \{x \in \mathbb{R}^n | Ax \leq b, A \in \mathbb{R}^{m\times n}, b \in \mathbb{R}^m\}$. Note that $mn^{\omega-1}$ is the complexity of linear system solver where $\omega < 2.38$.

As the target distribution for our walk is the uniform distribution on $\mathcal{K}$, the overall transition kernel for the walk, denoted by $\mathcal{Q}_x$, is defined by a density given by

$$q_x(z) := q(x, z) = \begin{cases} \min\left\{p_x(z), p_z(x)\right\}, & z \in \mathcal{K} \text{ and } z \neq x, \\ 0, & z \notin \mathcal{K}, \end{cases}$$

and the probability mass at $x$

$$\mathcal{Q}_x(\{x\}) := \mathcal{Q}(x, \{x\}) = 1 - \int\limits_{z\in\mathcal{K}} \min\left\{p_x(z), p_z(x)\right\} dz.$$

In Algorithm 1, we summarize the different steps of the Vaidya walk.

**Remark:** Note that Dikin walk with Gaussian proposals [Nar16] differs with Vaidya walk in the proposal step only. For Dikin walk, the proposal distribution at point $x$ is given by $\mathcal{N}\left(x, \frac{r'^2}{n} \nabla^2 F_x^{-1}\right)$ for a suitable constant $r'$.

### 3.2 Bound on mixing time

Now we state our main theorem which provides a bound for the mixing time of the walk.

**Theorem 1.** *Let $\mu_0$ be $M$-warm with respect to the distribution $\Pi$. Then for any $\delta \in (0, 1]$, the Vaidya walk with paramter $r = 1/9000$ (Algorithm 1) satisfies*

$$\left\|\mu_0 \mathcal{Q}^k - \Pi\right\|_{TV} \leq \delta, \quad \text{for all } k \geq C m^{1/2} n^{3/2} \log\left(\frac{\sqrt{M}}{\delta}\right),$$

*where $C > 0$ is a universal constant.*

**Algorithm 1:** Vaidya Walk with parameter $r$ (VW($r$))

```
Input: Parameter $r$ and $x_0 \in \text{int}(\mathcal{K})$
Output: Sequence $x_1, x_2, \ldots$
for $i = 0, 1, \ldots$ do
    $C_i \sim$ Fair Coin
    if $C_i = Heads$ then $x_{i+1} \leftarrow x_i$ // lazy step
    else
        $\xi_{i+1} \sim \mathcal{N}(0, \mathbb{I}_n)$
        $z_{i+1} = x_{i+1} + \frac{r}{(mn)^{1/4}} V_{x_i}^{-1/2} \xi_{i+1}$ // propose a new state
        if $z_{i+1} \notin \mathcal{K}$ then $x_{i+1} \leftarrow x_i$ // reject an infeasible proposal
        else
            $\alpha_{i+1} = \min\left\{1, \frac{p_{z_{i+1}}(x_{i+1})}{p_{x_{i+1}}(z_{i+1})}\right\}$
            $U_{i+1} \sim U[0,1]$
            if $U_{i+1} \geq \alpha_{i+1}$ then $x_{i+1} \leftarrow x_i$ // reject even a valid proposal
            else $x_{i+1} \leftarrow z_{i+1}$     // accept the proposal
        end
    end
end
```

The proof is provided in Section 4.4. The $\delta$-mixing time bound for Dikin walk from an $M$-warm start is $\mathcal{O}(mn \log(M/\delta))$. To contrast the overall computation time, we now discuss the per iteration cost of Vaidya walk. The proposal step of Vaidya walk requires computation of (a) the matrix $\Sigma$, and (b) the vector $V_x^{-1/2}\xi$. For part (a), we need to perform matrix inversion and matrix multiplication. Part (b) can be done by doing eigenvalue decomposition. The accept-reject step requires computation of determinants of $n \times n$ matrices besides a few matrix inverses and matrix-vector products. The complexity of all aforementioned operations is $\mathcal{O}\left(mn^{\omega-1}\right)$ (see the paper [DDH07]). Thus per iteration computational complexity for Vaidya walk is $\mathcal{O}\left(mn^{\omega-1}\right)$. Narayanan [Nar16] showed a similar complexity for Dikin walk. In fact, we can argue that Vaidya walk takes exactly twice the computation time of Dikin walk.

Vaidya walk assumes access to a point in the interior of the polytope $\mathcal{K}$. Our mixing time guarantee assumes access to an $M$-warm distribution. Instead, we can also have a deterministic start for the walk from a point $x_0 \in \text{int}(\mathcal{K})$ that is not too close to the boundary $\partial\mathcal{K}$. Such a point can be found using standard optimization methods, e.g., using a Phase-I method for Newton's algorithm. (See Section 11.5.4 in the book [BV04] for more discussion on different types of Phase-I method.) As expected, the mixing times now depend on the distance of the starting point from the boundary. A point $x \in \text{int}(\mathcal{K})$ is called $s$-central if for any chord $\overline{ef}$ passing through $x$ such that $e, f \in \partial\mathcal{K}$, we have $\left\|e - x\right\|_2 / \left\|f - x\right\|_2 \leq s$. For a start at an $s$-central point $x_0$, Dikin walk with proposals uniformly generated from the Dikin ellipsoid of radius $3/40$ (defined at $x$ by $\nabla^2 F_x$) has a polynomial bound on mixing time. (See Algorithm 1 in the paper [KN12].) The authors showed that when the walk moves to a new state for the first time, the distribution of the iterate is $\left(\sqrt{2ms}/r'\right)^n$-warm with respect to the distribution $\Pi$. Furthermore, it was also shown that the number of steps needed to make a non-trivial move follows a geometric distribution with a bounded mean. This motivates us to define the following hybrid walk for an $s$-central start.

Given an $s$-central point $x_0$, simulate Dikin walk till we observe a new state. Let $k_1$ denote the (random) number of steps taken to make the first non-trivial move. After $k_1$ steps, we run the walk $\mathrm{VW}(r)$ with $x_{k_1}$ as the initial point. We call such a walk as "$s$-central Dikin-start-Vaidya-walk". The role of Dikin walk for first few steps is to simply provide a warm start to Vaidya walk. Let $\mathcal{Q}_{\text{Dikin}}$ denote the transition kernel of the Dikin walk stated above. Then, we have the following mixing time bound for this hybrid walk.

**Corollary 1.** *Any $s$-central Dikin-start-Vaidya-walk with parameter $r = 1/9000$ satisfies*

$$\left\|\delta_{x_0}\mathcal{Q}_{Dikin}^{k_1}\mathcal{Q}^k - \Pi\right\|_{TV} \leq \delta, \quad \textit{for all } k \geq Cm^{1/2}n^{5/2}\log\left(\frac{ms}{\delta}\right),$$

*where $k_1$ is a geometric random variable with $\mathbb{E}[k_1] \leq C'$, and $C, C' > 0$ are universal constants.*

The proof follows immediately from Theorem 1 by Kannan et al. [KN12] and our Theorem 1 and is thereby omitted. Once again we observe that the mixing time bounds are improved by a factor of $\mathcal{O}\left(\sqrt{m/n}\right)$ when compared to Dikin walk from an $s$-central start [KN12, Nar16]. We now present the performance of Vaidya walk for some simulated examples.

### 3.3 Numerical experiments

In this section, we demonstrate the speed-up gained by Vaidya walk over Dikin walk for a warm start on different polytopes. We also provide an efficient implementation of both random walks on *github*. In particular, we simulate the random walks in $\mathbb{R}^2$ with initial distribution $\mu_0 = \mathcal{N}(0, 0.04\,\mathbb{I}_2)$, on the following three different types of polytopes:

1. The set $[-1, 1]^2$,
2. symmetric polytopes with $m$-constraints generated randomly, and
3. the interior of regular $m$-polygons on the unit circle.

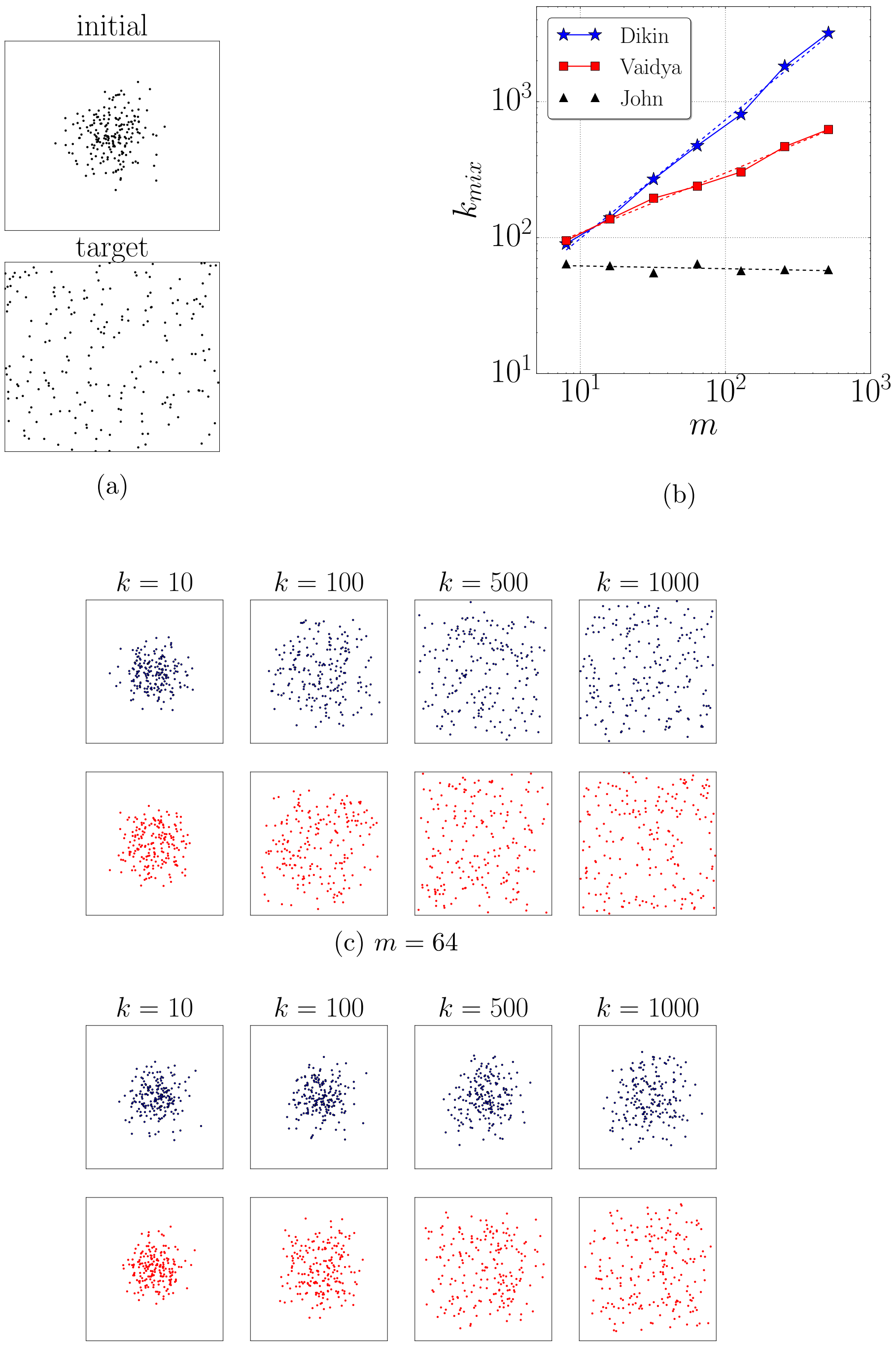
We remark that the warmness-$M$ in all cases is bounded by 500.

For Case 1, we can represent the set exactly by 4 linear constraints. Repeating the constraints increases $m$ for the matrix $A$ associated with $\mathcal{K}$ and hence affects the mixing times of the two random walks. While the mixing time for Dikin walk is affected linearly in number of constraints, Vaidya walk slows down sub-linearly with $m$. We plot the empirical distribution for the iterates from the two random walks for $m = 64$ and 512 in Figure 1 from which we observe significant difference in the effect of $m$ on the rates of two walks. Note that the warmness $M \leq 8$ for this case. Further, we also plot the approximate mixing time for the set $\mathcal{S} = ([-1, -1/2] \cup [1/2, 1]) \times [-1, 1]$. Let $\mu_0\hat{\mathcal{Q}}^k$ denote the empirical measure after $k$-iterations across 1000 experiments. In Figure 1b, we plot $k_{\text{mix}}$ as a function of $m$ where

$$k_{\text{mix}} := k_{\text{mix}}\left(\mathcal{S}, \frac{1}{4}\right) := \min\left\{k \middle| \frac{\Pi(\mathcal{S}) - \mu_0\hat{\mathcal{Q}}^k(\mathcal{S})}{\Pi(\mathcal{S})} \leq \frac{1}{4}\right\}, \tag{6}$$

and observe that the slope of the best fit lines in the log-log plot is approximately 0.9 and 0.45 for Dikin and Vaidya walk, which is in good accordance with the theoretical guarantees of the two walks.

In Case 2, for each constraint $i$, we fix $b_i = 1$. To generate $a_i$, first we draw two uniform random variables and then flip the sign of both of them with probability 1/2. The set $\mathcal{K} = \left\{x | a_i^\top x \leq b_i, i = 1, \ldots, m\right\}$ is closed under such a procedure with high probability. From Figure 2a-2b we observe that the effect of $m$ on the mixing time of the two walks is different for this case, and Vaidya walk seems to mix faster than Dikin walk. A similar observation can be made even from Figure 2c-2d for Case 3.



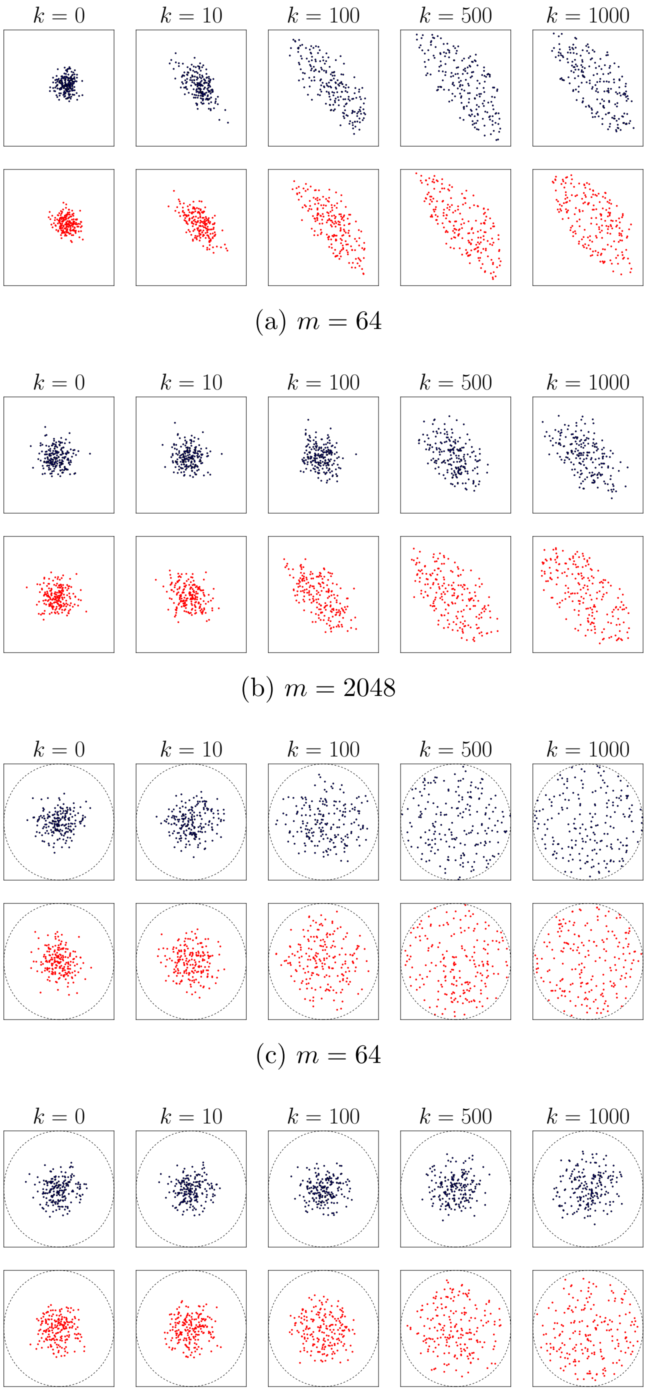
(a)

(b)

(c) $m = 64$

(d) $m = 2048$

**Figure 1.** Comparison of Dikin and Vaidya walks (500 runs) on the polytope $\mathcal{K} = [-1, 1]^2$. (a) Samples from the initial distribution $\mu_0 = \mathcal{N}(0, 0.04\,\mathbb{I}_2)$ and the uniform distribution ($\Pi$) on $[-1, 1]^2$. (b) Plot of $k_{\text{mix}}$ (6) versus the number of constraints ($m$) for $\mathcal{S} = ([-1, -1/2] \cup [1/2, 1]) \times [-1, 1]$. Dotted lines show the best-fit lines which have slopes 0.88 and 0.45 for Dikin and Vaidya walks respectively. (c, d) Empirical distribution of samples of Dikin (blue/top rows) and Vaidya (red/bottom rows) walks for different values of $m$ at iteration $k = 10, 100, 500$ and 1000. From the figures we can observe: (1) As $m$ increases, both walks take more number of iterations to mix. (2) The effect of increasing $m$ on mixing time of Vaidya walk is significantly lesser compared to that on the mixing time of Dikin walk.



(a) $m = 64$

(b) $m = 2048$

(c) $m = 64$

(d) $m = 2048$

**Figure 2.** Empirical distribution of samples from 500 runs of Dikin (blue/top rows) and Vaidya (red/bottom rows) walks on different polytopes. $k$ denotes the iteration number. (a, b) We simulate the two random walks on random polytopes with 64 and 2048 constraints respectively (for details refer to Section 3.3). (c, d) We simulate the walks on regular $m$-polygons inscribed in the unit circle, for $m = 64$ and 2048. For both cases, we observe that higher $m$ slows down the walks, with visibly more effect on Dikin walk than Vaidya walk.

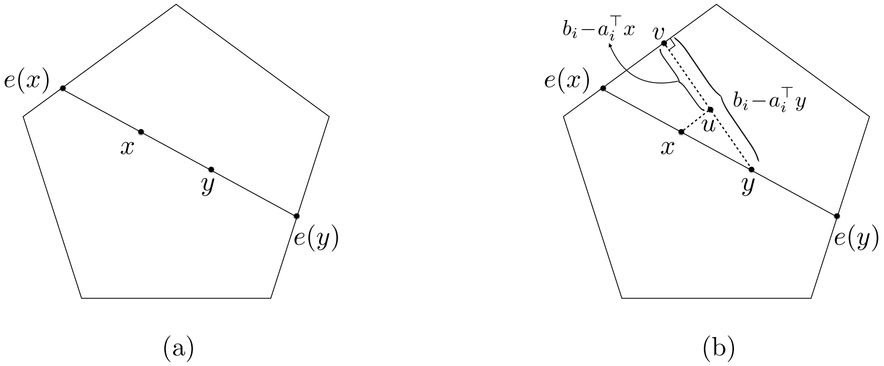


**Figure 3.** Polytope $\mathcal{K} = \{x \in \mathbb{R}^n | Ax \leq b\}$. (a) The points $e(x)$ and $e(y)$ denote the intersection points of the chord joining $x$ and $y$ with $\mathcal{K}$ such that $e(x), x, y, e(y)$ are in order. (b) A geometric illustration of the argument (11). It is straightforward to observe that $\|x-y\|_2/\|e(x)-x\|_2 = \|u-y\|_2/\|u-v\|_2 = \left|a_i^\top(y-x)\right| / \left(b_i - a_i^\top x\right)$.

# 4 Analysis of Vaidya walk

In this section, we first provide an outline of a general method of bounding the rate of convergence of geometric random walks on convex sets in Section 4.1, followed by some auxiliary results in Section 4.2 that are used to invoke the general method in our case. Using these auxiliary results, we first contrast the differences between Dikin walk and Vaidya walk in Section 4.3, and then prove our main result in Section 4.4. We provide the proofs of the auxiliary results in Sections 4.5, 4.6 and 4.7, and defer some technical results to appendices.

## 4.1 A general method to bound mixing time

For a discrete-space Markov chain, a bound on mixing time is obtained via bounds on the *spectral gap* of the transition matrix associated with the chain. Often, an indirect bound on the spectral gap is obtained via Cheeger's inequality that bounds the spectral gap in terms of the *conductance* of the chain. Lovász and Simonovits [LS93] proved a similar connection between conductance and convergence rate for continuous-space Markov chains. Thus proving an upper bound for the mixing time of a geometric random walk on convex sets often boils down to showing a good lower bound on the conductance of the chain—these arguments have been used for ball walk [LS90], Hit-and-run [Lov99, LV06b] and Dikin walk [Nar16, KN12, SV16] on convex sets. When the underlying space is convex and bounded, using some isoperimetric inequalities, Lovasz showed that to get a lower bound on the conductance of the Markov chain with stationary distribution uniform, it suffices to establish that the chain satisfies the following good-neighborhood-property: "if two points are close, then their one-step transition distribution are also close." The mixing time of the chain roughly scales with the square of the inverse of "how close the two-points need to be" in order for their one-step transition distributions to be close. We show that for Vaidya walk when compared to Dikin walk, the points can be much far apart, with their one-step transition distributions still being close. Consequently, our walk mixes faster than Dikin walk. Much of our work focuses on quantifying the last claim formally. We restrict our attention to formalizing the "good-neighborhood-property", and refer the reader to the survey by Vempala [Vem05] for a more thorough and formal discussion about the usual techniques and the proofs of other arguments that have been summarized above.

We now present a formal statement of the good-neighborhood-property for a random walk on a bounded convex set. For quantifying closeness in the property, the distributions are contrasted with the total-variation distance, while for distance between points we use the cross ratio, that we define next. For a given pair of points $x, y \in \mathcal{K}$, let $e(x), e(y) \in \partial\mathcal{K}$ denote the intersection of the chord joining $x$ and $y$ with $\mathcal{K}$ such that $e(x), x, y, e(y)$ are in order (see Figure 3a). The cross-ratio $d_{\mathcal{K}}(x, y)$ is given by

$$d_{\mathcal{K}}(x, y) = \frac{\left\|e(x) - e(y)\right\|_2 \left\|x - y\right\|_2}{\left\|e(x) - x\right\|_2 \left\|e(y) - y\right\|_2}. \tag{7}$$

The ratio $d_{\mathcal{K}}(x, y)$ is related to the Hilbert metric on $\mathcal{K}$, which is given by $\log\left(1 + d_{\mathcal{K}}(x, y)\right)$ (see the paper by Bushell [Bus73] for more details).

Let $X_0, X_1, \ldots$ denote a *lazy* reversible random walk on a bounded convex set $\mathcal{K}$ with transition kernel $\widetilde{\mathcal{T}} : x \mapsto \delta_x(\cdot)/2 + \mathcal{T}_x(\cdot)/2$ and stationary with respect to the uniform distribution on $\mathcal{K}$ (denoted by $\Pi$). The following lemma gives a bound on the mixing-time of the chain $\{X_t, t \geq 0\}$.

**Lemma 1.** *Suppose that $\left\|\mathcal{T}_x - \mathcal{T}_y\right\|_{TV} \leq 1 - \rho$, for all $x, y \in int(\mathcal{K})$ such that $d_{\mathcal{K}}(x, y) \ < \Delta$, for some $\rho, \Delta \in (0, 1)$. Then for every distribution $\mu_0$ that is $M$-warm with respect to $\Pi$, we have*

$$\left\|\mu_0 \widetilde{\mathcal{T}}^k - \Pi\right\|_{TV} \leq \sqrt{M} \exp\left(-k\,\frac{\Delta^2 \rho^2}{512}\right).$$

The proof of the lemma follows from an isoperimetry inequality involving cross-ratio [Lov99] and the classical mixing time bound in terms of conductance [LS93]. Similar results have also been used in the proofs of Hit-and-run and Dikin walk. We provide the proof of the lemma in Section 4.5. To prove Theorem 1, we show that the random walk VW(1/9000) satisfies the assumptions of Lemma 1 with suitable $\Delta$ and $\rho$ which yields the claimed mixing time bound. Besides Lemma 1, our proof techniques are inspired by the proofs of convergence rate of Dikin walk on polytopes presented by Kannan et al. [KN12] and the simple proof of Dikin walk provided by Sachdeva et al. [SV16].

### 4.2 Some auxiliary results

We now introduce some notation and auxiliary results that will be useful for the proof. For all $x \in \mathcal{K}$ let $s_x := (b_1 - a_1^\top x, \ldots, b_m - a_m^\top x)^\top$ denote the "slackness at $x$". For all $x \in \text{int}\left(\mathcal{K}\right)$, define the "local norm at $x$" as

$$\left\|.\right\|_x : v \mapsto \left\|V_x^{1/2} y\right\|_2 = \sqrt{\sum_{i=1}^m (\sigma_{x,i} + \beta) \frac{(a_i^\top v)^2}{s_{x,i}^2}}, \tag{8a}$$

and the "slack sensitivity at $x$" as

$$\theta_x := \left(\left\|\frac{a_1}{s_{x,1}}\right\|_x^2, \ldots, \left\|\frac{a_m}{s_{x,m}}\right\|_x^2\right)^\top = \left(\frac{a_1^\top V_x^{-1} a_1}{s_{x,1}^2}, \ldots, \frac{a_m^\top V_x^{-1} a_m}{s_{x,m}^2}\right)^\top. \tag{8b}$$

The following lemma provides useful properties of the leverage scores $\sigma_x$ (3) and the slack sensitivity $\theta_x$ for all $x \in \text{int}\left(\mathcal{K}\right)$. The importance of these properties is highlighted in the discussion that follows in the next subsection.

**Lemma 2.** *For any $x \in \mathit{int}(\mathcal{K})$, the following properties hold:*

*(a)* $\sigma_{x,i} \in [0,1]$ *for all* $i \in [m]$,

*(b)* $\theta_{x,i} \in \left[0, \sqrt{m/n}\right]$ *for all* $i \in [m]$, *and*

*(c)* $\sum_{i=1}^{m} \sigma_{x,i} = n$.

The proof of this lemma is provided in Section 4.6. Next, we state a lemma that shows that if two points $x, y \in \operatorname{int}(\mathcal{K})$ are close in local norm, then the one-step transition probability distributions $\mathcal{Q}_x$ and $\mathcal{Q}_y$ of the random walk VW$(r)$ for suitable choice of the radius $r$, are also close in TV-distance.

**Lemma 3.** *There exists a continuous non-decreasing function* $f : [0, 1/12] \to \mathbb{R}_+$ *with* $f(1/12) \geq 1/9000$, *such that for any* $\epsilon \in (0, 1/4]$, *the random walk VW(r) with* $r \in [0, f(\epsilon)]$ *satisfies*

$$\left\|\mathcal{P}_x - \mathcal{P}_y\right\|_{TV} \leq \epsilon, \quad \forall x, y \in \mathit{int}(\mathcal{K}) \text{ such that } \left\|x - y\right\|_x \leq \frac{\epsilon r}{2(mn)^{1/4}}, \quad \text{and} \tag{9a}$$

$$\left\|\mathcal{Q}_x - \mathcal{P}_x\right\|_{TV} \leq 5\epsilon, \ \ \forall x \in \mathit{int}(\mathcal{K}). \tag{9b}$$

We provide the proof of the lemma in Section 4.7. We now provide an outline of the proof of Theorem 1 with some remarks that contrast Vaidya and Dikin walks.

### 4.3 Vaidya walk vs Dikin walk

With the help of Lemma 1, our proof of Theorem 1 proceeds by answering the following three questions:

(a) How does the local norm for the random walk relate to the cross-ratio?

(b) How small does the distance $\left\|x - y\right\|_x$ need to be for the proposal distributions $\mathcal{P}_x$ and $\mathcal{P}_y$ to be close?

(c) What should be the value of $r$ such that for any $x \in \operatorname{int}(\mathcal{K})$, the Metropolis accept-reject step does not reject the proposal made by $\mathcal{P}_x$ with high probability?

The first question is dealt directly in the proof of Theorem 1 while the other two questions are dealt by Lemma 3. We now provide a high-level discussion of these questions with remarks that contrast Vaidya walk with Dikin walk. We contrast the two walks in terms of their *respective* local-norms which are defined with respect to the matrix $V_x$ (8a) for the Vaidya walk and with respect to the matrix $\nabla^2 F_x$ for the Dikin walk.

**Part (a):** While the cross-ratio is bounded below by a factor of $1/\sqrt{n}$ (see bound (10)) to the local-norm for Vaidya walk, the same factor is $1/\sqrt{m}$ for Dikin walk (see, e.g., Lemma 9 in the paper [SV16]). Since the mixing time is affected inversely by this factor, we see that Vaidya walk has an $\mathcal{O}\left(\sqrt{m/n}\right)$ advantage over Dikin walk for this part.

**Part (b):** To have a fast mixing chain, the distributions $\mathcal{P}_x$ and $\mathcal{P}_y$ need to be similar even when the points $x$ and $y$ are quite far apart. If we apply Pinsker's inequality, it suffices to show that the covariance $V_y V_x^{-1} \approx \mathbb{I}_n$, when $x$ and $y$ are close in local-norm. As we elaborate later, it suffices to control the eigenvalues of the matrix $V_y V_x^{-1}$ as a function of the local-norm of $x - y$. We show that the local-norm associated with Vaidya walk is *weaker* than the local-norm associated with Dikin walk, i.e., a small perturbation in local-norm changes the covariance matrix much more for Vaidya walk compared to Dikin walk. As a result, $x$ and $y$ need to be much closer in local-norm for Vaidya walk ($\mathcal{O}\left(1/(mn)^{1/4}\right)$), compared to Dikin walk ($\mathcal{O}\left(1/\sqrt{n}\right)$) for $\mathcal{P}_x$ and $\mathcal{P}_y$ to be close. Thus, on this front Vaidya walk is worse by a factor of $\mathcal{O}\left((m/n)^{1/4}\right)$ to Dikin walk.

**Part (c):** Note that the scale parameter in the proposal step of the walk VW($r$) is $r/(mn)^{1/4}$ and for Dikin walk the scale parameter is $r'/\sqrt{n}$, where $r$ and $r'$ are universal constants. As a result, the scaling factor for Vaidya walk is worse by $\mathcal{O}\left((m/n)^{1/4}\right)$ to Dikin walk. We now elaborate the reason for such a scaling at a high level. The scaling factor is chosen such that the proposals are accepted with at least a constant probability for any $x \in \text{int}(\mathcal{K})$. In order to ensure this "good proposal property", we have to show that for the given scalings the proposal $z \sim \mathcal{P}_x$ satisfies the following two properties with high probability—$(i)$ the proposal $z$ remains inside $\mathcal{K}$, and $(ii)$ the ratio $p_z(x)/p_x(z)$ is bounded away from zero. It is easy to show part $(i)$ using concentration of $n$-dimensional standard Gaussian vector on the unit sphere in $\mathbb{R}^n$, and the bound on the slack sensitivity $\theta_x$. However, proving part $(ii)$ needs some work that we summarize now. To control the ratio $p_z(x)/p_x(z)$, we need to carefully bound the difference in the volume of unit-ellipsoids defined by $V_z$ and $V_x$, and the difference in local-norms at $z$ and $x$ scaled by $\sqrt{mn}$. It turns out that controlling the eigenvalues of the matrix $V_z V_x^{-1}$ leads to a non-useful bound and hence we make use of Taylor expansion. For both the parts $(i)$ and $(ii)$, it turns out that if the slack sensitivity $\theta_x$ is smaller, we can use a larger scaling factor. Since the local sensitivity induced by Vaidya walk is larger than Dikin walk, we need to use a scaling of $r/\left(mn\right)^{1/4}$ for the former compared to $r'/\sqrt{n}$ for the later.

**Who wins the race?** Putting the three parts together, we see that we gain a factor of $\mathcal{O}\left(\sqrt{m/n}\right)$ in part(a) and lose a factor of $\mathcal{O}\left((m/n)^{1/4}\right)$ both in part (b) and part (c). We prove later that the order of the ratio of mixing times of the two random walks is affected quadratically with respect to the gain from part (a) times maximum of loss from parts (b) and (c). Thus we see an overall gain of $\mathcal{O}\left(\sqrt{m/n}\right)$ in mixing time for Vaidya walk over Dikin walk.

We are now ready to provide a formal proof of Theorem 1.

### 4.4 Proof of Theorem 1

We prove Theorem 1 using Lemmas 1, 2 and 3. To invoke Lemma 1 for VW(1/9000), we need to show that for any two points $x, y \in \text{int}(\mathcal{K})$ such that $d_{\mathcal{K}}(x, y)$ is small, we have that $\|\mathcal{Q}_x - \mathcal{Q}_y\|_{\text{TV}}$ is small. Along the outline discussed in previous subsection, we break our analysis in two steps—(A) We first relate the cross-ratio $d_{\mathcal{K}}(x, y)$ to the local norm (8a) at $x$, and (B) then use Lemma 3 to show that if $x, y \in \text{int}(\mathcal{K})$ are close in local-norm, then the transition kernels $\mathcal{Q}_x$ and $\mathcal{Q}_y$ are close in TV-distance.

**Step (A):** We claim that for all $x, y \in \text{int}(\mathcal{K})$, the cross-ratio can be lower bounded as

$$d_{\mathcal{K}}(x, y) \geq \frac{1}{\sqrt{2n}} \left\| x - y \right\|_x . \tag{10}$$

Note that we have

$$\begin{aligned} d_{\mathcal{K}}(x, y) = \frac{\left\| e(x) - e(y) \right\|_2 \left\| x - y \right\|_2}{\left\| e(x) - x \right\|_2 \left\| e(y) - y \right\|_2} &\overset{(i)}{\geq} \max \left\{ \frac{\left\| x - y \right\|_2}{\left\| e(x) - x \right\|_2}, \frac{\left\| x - y \right\|_2}{\left\| e(y) - y \right\|_2} \right\} \\ &\overset{(ii)}{\geq} \max \left\{ \frac{\left\| x - y \right\|_2}{\left\| e(x) - x \right\|_2}, \frac{\left\| x - y \right\|_2}{\left\| e(y) - x \right\|_2} \right\} \end{aligned}$$

where step $(i)$ follows from the inequality $\left\| e(x) - e(y) \right\|_2 \geq \max \left\{ \left\| e(y) - y \right\|_2, \left\| e(x) - x \right\|_2 \right\}$ and step $(ii)$ from the inequality $\left\| e(x) - x \right\|_2 \leq \left\| e(y) - x \right\|_2$. Furthermore, from Figure 3b, we observe that

$$\max \left\{ \frac{\left\| x - y \right\|_2}{\left\| e(x) - x \right\|_2}, \frac{\left\| x - y \right\|_2}{\left\| e(y) - x \right\|_2} \right\} = \max_{i \in [m]} \left| \frac{a_i^\top (x - y)}{s_{x,i}} \right| . \tag{11}$$

Note that maximum of a set of non-negative numbers is greater than the mean of the numbers. Combining this fact with properties (a) and (c) from Lemma 2,

$$d_{\mathcal{K}}(x, y) \geq \sqrt{\frac{1}{\sum_{i=1}^m (\sigma_{x,i} + \beta)} \sum_{i=1}^m (\sigma_{x,i} + \beta) \frac{(a_i^\top (x - y))^2}{s_{x,i}^2}} = \frac{\left\| x - y \right\|_x}{\sqrt{2n}},$$

thereby proving the claim (10).

**Step (B):** By the triangle inequality, we have

$$\left\| \mathcal{Q}_x - \mathcal{Q}_y \right\|_{\text{TV}} \leq \left\| \mathcal{Q}_x - \mathcal{P}_x \right\|_{\text{TV}} + \left\| \mathcal{P}_x - \mathcal{P}_y \right\|_{\text{TV}} + \left\| \mathcal{P}_y - \mathcal{Q}_y \right\|_{\text{TV}} .$$

Thus, for any $(r, \epsilon)$ such that $\epsilon \in [0, 1/4]$ and $r \leq f(\epsilon)$, Lemma 3 implies that

$$\left\| \mathcal{Q}_x - \mathcal{Q}_y \right\|_{\text{TV}} \leq 11\epsilon, \quad \forall x, y \in \text{int}(\mathcal{K}) \text{ such that } \left\| x - y \right\|_x \leq \frac{r\epsilon}{2(mn)^{1/4}}.$$

Consequently, the walk $\text{VW}(r)$ satisfies the assumptions of Lemma 1 with

$$\Delta := \frac{1}{\sqrt{2n}} \cdot \frac{r\epsilon}{2(mn)^{1/4}} \quad \text{and} \quad \rho := 1 - 11\epsilon.$$

Since $f(1/12) \geq 1/9000$, we can set $\epsilon = 1/12$ and $r = 1/9000$, whence

$$\Delta^2 \rho^2 = \frac{(1 - 11\epsilon)^2 \epsilon^2 r^2}{8n\sqrt{mn}} = \frac{1}{12^4} \frac{1}{9000^2} \cdot \frac{1}{n\sqrt{mn}} \geq \frac{1}{2} \cdot 10^{-12} \frac{1}{n\sqrt{mn}}.$$

Observing that $\Delta < 1$ yields the claimed upper bound for the mixing time of Vaidya Walk.

### 4.5 Proof of Lemma 1

We begin by formally defining the conductance $(\Phi)$ of a Markov chain on $(\mathcal{K}, \mathbb{B}(\mathcal{K}))$ with arbitrary transition kernel $\mathcal{Q}_x$ and stationary distribution $\Pi$

$$\Phi := \inf_{\substack{\mathcal{S}\in\mathbb{B}(\mathcal{K})\\ \Pi(\mathcal{S})\in(0,1/2)}} \frac{\Phi(\mathcal{S})}{\Pi(\mathcal{S})} \quad \text{where} \quad \Phi(\mathcal{S}) := \int_{\mathcal{S}} \mathcal{Q}_u(\mathcal{K}\cap\mathcal{S}^c)d\Pi(u) \quad \text{for any } \mathcal{S}\subseteq\mathcal{K}.$$

The conductance denotes the measure of the flow from a set to its complement relative to its own measure, when initialized in the stationary distribution. If the conductance is high, the following celebrated result shows that the Markov chain mixes fast.

**Lemma 4.** *For any $M$-warm start $\mu_0$, the mixing time of the Markov chain with conductance $\Phi$ is bounded as*

$$\left\|\mu_0\mathcal{Q}^k - \Pi\right\|_{TV} \leq \sqrt{M}\left(1-\frac{\Phi^2}{2}\right)^k \leq \sqrt{M}\exp\left(-k\frac{\Phi^2}{2}\right).$$

Note that this result holds for a general distribution $\Pi$ although we apply for uniform $\Pi$. The result can be derived from Cheeger's inequality for continuous-space discrete-time Markov chain and elementary results in Calculus. See, e.g., Theorem 1.4 and Corollary 1.5 in the paper [LS93] for a proof. For ease in notation define $\mathcal{K}\backslash\mathcal{S} := \mathcal{K}\cap\mathcal{S}^c$. We now state a key isoperimetric inequality.

**Lemma 5** (Theorem 6 [Lov99])**.** *For any measurable sets $\mathcal{S}_1, \mathcal{S}_2 \subseteq \mathcal{K}$, we have*

$$\text{vol}(\mathcal{K}\backslash\mathcal{S}_1\backslash\mathcal{S}_2)\cdot\text{vol}(\mathcal{K}) \geq d_{\mathcal{K}}(\mathcal{S}_1,\mathcal{S}_2)\cdot\min\left\{\text{vol}(\mathcal{S}_1),\text{vol}(\mathcal{S}_2)\right\},$$

*where $d_{\mathcal{K}}(\mathcal{S}_1,\mathcal{S}_2) := \inf_{x\in\mathcal{S}_1,y\in\mathcal{S}_2} d_{\mathcal{K}}(x,y)$.*

(There is a typo on the RHS in the statement of the theorem in the paper.) Since $\Pi$ is the uniform measure on $\mathcal{K}$, this lemma implies that

$$\Pi(\mathcal{K}\backslash\mathcal{S}_1\backslash\mathcal{S}_2) \geq d_{\mathcal{K}}(\mathcal{S}_1,\mathcal{S}_2)\min\left\{\Pi(\mathcal{S}_1),\Pi(\mathcal{S}_2)\right\}. \tag{12}$$

In fact, such an inequality holds for an arbitrary log-concave distribution [LV03]. In words, the inequality says that for a bounded convex set any two subsets which are far apart, can not have a large volume. Taking these lemmas as given, we now complete the proof.

*Proof of Lemma 1.* We first bound the conductance of the Markov chain using the assumptions of the lemma. From Lemma 4, we see that the Markov chain mixes fast if all the sets $\mathcal{S}$ have a high conductance $\Phi(\mathcal{S})$. We claim that

$$\Phi \geq \frac{\rho\Delta}{16}, \tag{13}$$

from which the proof follows by applying Lemma 4. We now prove the claim (13) along the lines of Theorem 11 in the paper [Lov99]. In particular, we show that under the assumptions in the lemma, the sets with bad conductance are far apart and thereby have a small measure under $\Pi$, whence the ratio $\Phi(\mathcal{S})/\Pi(\mathcal{S})$ is not arbitrarily small. Consider a partition $\mathcal{S}_1, \mathcal{S}_2$ of the set $\mathcal{K}$ such that $\mathcal{S}_1$ and $\mathcal{S}_2$ are measurable. Since $\Pi$ is the uniform measure, it suffices to show that

$$\int_{\mathcal{S}_1} \tilde{\mathcal{T}}_u(\mathcal{S}_2)du \geq \frac{\rho\Delta}{16} \cdot \min\left\{\mathrm{vol}(\mathcal{S}_1), \mathrm{vol}(\mathcal{S}_2)\right\}.$$

Consider the sets

$$\mathcal{S}_1' := \left\{u \in \mathcal{S}_1 \middle| \mathcal{T}_u(\mathcal{S}_2) < \frac{\rho}{2}\right\}, \quad \mathcal{S}_2' \quad := \left\{v \in \mathcal{S}_2 \middle| \mathcal{T}_v(\mathcal{S}_1) < \frac{\rho}{2}\right\}, \quad \text{and} \quad \mathcal{S}_3' := \mathcal{K}\backslash\mathcal{S}_1'\backslash\mathcal{S}_2'. \tag{14}$$

If we have $\mathrm{vol}(\mathcal{S}_1') \leq \mathrm{vol}(\mathcal{S}_1)/2$ and consequently $\mathrm{vol}(\mathcal{K}\backslash\mathcal{S}_1') \geq \mathrm{vol}(\mathcal{S}_1)/2$, then

$$\int_{\mathcal{S}_1} \tilde{\mathcal{T}}_u(\mathcal{S}_2)du \overset{(i)}{\geq} \frac{1}{2}\int_{\mathcal{K}\backslash\mathcal{S}_1'} \mathcal{T}_u(\mathcal{S}_2)du \overset{(ii)}{\geq} \frac{\rho}{4}\mathrm{vol}(\mathcal{S}_1) \overset{(iii)}{\geq} \frac{\rho\Delta}{16} \cdot \min\left\{\mathrm{vol}(\mathcal{S}_1), \mathrm{vol}(\mathcal{S}_2)\right\},$$

and we are done. In the above sequence of inequalities, step $(i)$ follows from the definition of the kernel $\tilde{\mathcal{T}}$, step $(ii)$ follows from the definition of the set $\mathcal{S}_1'$ (14) and step $(iii)$ from the fact that $\Delta < 1$.

Hence, without loss of generality we can assume $\mathrm{vol}(\mathcal{S}_i') \geq \mathrm{vol}(\mathcal{S}_i)/2$ for each $i \in \{1, 2\}$. Now for any $u \in \mathcal{S}_1'$ and $v \in \mathcal{S}_2'$ we have

$$\|\mathcal{T}_u - \mathcal{T}_v\|_{\mathrm{TV}} \geq \mathcal{T}_u(\mathcal{S}_1) - \mathcal{T}_v(\mathcal{S}_1) = 1 - \mathcal{T}_u(\mathcal{S}_2) - \mathcal{T}_v(\mathcal{S}_1) > 1 - \rho,$$

and hence by assumption we have $d_{\mathcal{K}}(\mathcal{S}_1', \mathcal{S}_2') \geq \Delta$. Applying Lemma 5 and the definition of $\mathcal{S}_3'$ (14) we find that

$$\mathrm{vol}(\mathcal{S}_3') \geq \Delta \min\left\{\mathrm{vol}(\mathcal{S}_1'), \mathrm{vol}(\mathcal{S}_2')\right\} \geq \frac{\Delta}{2}\min\left\{\mathrm{vol}(\mathcal{S}_1), \mathrm{vol}(\mathcal{S}_2)\right\}. \tag{15}$$

We now show that

$$\int_{\mathcal{S}_1} \tilde{\mathcal{T}}_u(\mathcal{S}_2)du = \int_{\mathcal{S}_2} \tilde{\mathcal{T}}_v(\mathcal{S}_1)dv. \tag{16}$$

Noting that $\mathcal{S}_1 = \mathcal{K}\backslash\mathcal{S}_2$, we have

$$\frac{1}{\mathrm{vol}(\mathcal{K})}\int_{\mathcal{S}_1} \tilde{\mathcal{T}}_u(\mathcal{S}_2)du = \int_{\mathcal{S}_1} \tilde{\mathcal{T}}_u(\mathcal{S}_2)\pi(u)du = \Phi(\mathcal{S}_1) \overset{(i)}{=} \Phi(\mathcal{K}\backslash\mathcal{S}_1) = \frac{1}{\mathrm{vol}(\mathcal{K})}\int_{\mathcal{S}_2} \tilde{\mathcal{T}}_v(\mathcal{S}_1)dv,$$

where in step $(i)$ we have used the following claim: For any measurable $\mathcal{S} \subseteq \mathcal{K}$, we have

$$\Phi(\mathcal{S}) = \Phi(\mathcal{K}\backslash\mathcal{S}). \tag{17}$$

We now prove the claim (17). Noting that $\int_{\mathcal{K}} \tilde{\mathcal{T}}_u(\mathcal{S})d\Pi(u) = \Pi(\mathcal{S})$, we have

$$\Phi(\mathcal{K}\backslash\mathcal{S}) = \int_{\mathcal{K}\backslash\mathcal{S}} \tilde{\mathcal{T}}_u(\mathcal{S})d\Pi(u) = \int_{\mathcal{K}} \tilde{\mathcal{T}}_u(\mathcal{S})d\Pi(u) - \int_{\mathcal{S}} \tilde{\mathcal{T}}_u(\mathcal{S})d\Pi(u) = \Pi(\mathcal{S}) - \int_{\mathcal{S}} \tilde{\mathcal{T}}_u(\mathcal{S})d\Pi(u).$$

Using the fact that $1 - \tilde{\mathcal{T}}_u(\mathcal{S}) = \tilde{\mathcal{T}}_u(\mathcal{K}\backslash\mathcal{S})$, we obtain

$$\Pi(\mathcal{S}) - \int_{\mathcal{S}} \tilde{\mathcal{T}}_u(\mathcal{S})d\Pi(u) = \int_{\mathcal{S}} d\Pi(u) - \int_{\mathcal{S}} \tilde{\mathcal{T}}_u(\mathcal{S})d\Pi(u) = \int_{\mathcal{S}} \tilde{\mathcal{T}}_u(\mathcal{K}\backslash\mathcal{S})d\Pi(u) = \Phi(\mathcal{S}),$$

thereby yielding the claim.

Using the equation (16), we find that

$$\int_{\mathcal{S}_1} \widetilde{\mathcal{T}}_u(\mathcal{S}_2)du = \frac{1}{2}\left(\int_{\mathcal{S}_1} \widetilde{\mathcal{T}}_u(\mathcal{S}_2)du + \int_{\mathcal{S}_2} \widetilde{\mathcal{T}}_v(\mathcal{S}_1)dv\right) \overset{(i)}{\geq} \frac{1}{2}\left(\frac{1}{2}\int_{\mathcal{K}\backslash\mathcal{S}_1'} \mathcal{T}_u(\mathcal{S}_2)du + \frac{1}{2}\int_{\mathcal{K}\backslash\mathcal{S}_2'} \mathcal{T}_v(\mathcal{S}_2)dv\right)$$
$$\overset{(ii)}{\geq} \frac{\rho}{8}\operatorname{vol}(\mathcal{S}_3')$$
$$\overset{(iii)}{\geq} \frac{\rho\Delta}{16}\min\left\{\operatorname{vol}(\mathcal{S}_1), \operatorname{vol}(\mathcal{S}_2)\right\},$$

where step $(i)$ follows from the definition of the kernel $\widetilde{\mathcal{T}}$, step $(ii)$ follows from the definition of the set $\mathcal{S}_3'$ (14) and step $(iii)$ follows from the inequality (15). Putting together the pieces yields the claim (13). □

### 4.6 Proof of Lemma 2

To prove part (a) observe that the Hessian $\nabla^2 F_x = \sum_{i=1}^m a_i a_i^\top / s_{x,i}^2$ is a sum of rank one positive semidefinite (PSD) matrices. Also, we can write $\nabla^2 F_x = A_x^\top A_x$ where

$$A_x := \begin{bmatrix} a_1^\top / s_{x,1} \\ \vdots \\ a_m^\top / s_{x,m} \end{bmatrix} \quad \text{for all } x \in \operatorname{int}(\mathcal{K}).$$

Since $\operatorname{rank}(A_x) = n$, we conclude that the matrix $\nabla^2 F_x$ is invertible and thus, both the matrices $\nabla^2 F_x$ and $\left(\nabla^2 F_x\right)^{-1}$ are PSD. Since $\sigma_{x,i} = a_i^\top \left(\nabla^2 F_x\right)^{-1} a_i / s_{x,i}^2$, we have $\sigma_{x,i} \geq 0$. Further, the fact that $a_i a_i^\top / s_{x,i}^2 \preceq \nabla^2 F_x$ implies that $\sigma_{x,i} \leq 1$.

For part (c), using the equality $\operatorname{trace}(AB) = \operatorname{trace}(BA)$, we obtain

$$\sum_{i=1}^m \sigma_{x,i} = \operatorname{trace}\left(\sum_{i=1}^m \frac{a_i^\top \left(\nabla^2 F_x\right)^{-1} a_i}{s_{x,i}^2}\right) = \operatorname{trace}\left(\left(\nabla^2 F_x\right)^{-1} \sum_{i=1}^m \frac{a_i a_i^\top}{s_{x,i}^2}\right) = \operatorname{trace}(\mathbb{I}_n) = n.$$

Now we prove part (b). Using the fact that $\sigma_{x,i} \geq 0$, and an argument similar to part (a) we find that that the matrices $V_x$ and $V_x^{-1}$ are PSD. Since $\theta_{x,i} = a_i^\top V_x^{-1} a_i / s_{x,i}^2$, we have $\theta_{x,i} \geq 0$. It is straightforward to see that $\beta \nabla^2 F_x \preceq V_x$ which implies that $\theta_{x,i} \leq \sigma_{x,i}/\beta$. Further, we also have $(\sigma_{x,i} + \beta) \frac{a_i a_i^\top}{s_{x,i}^2} \preceq V_x$ and whence $\theta_{x,i} \leq 1/\left(\sigma_{x,i} + \beta\right)$. Combining the two inequalities yields the claim.

### 4.7 Proof of Lemma 3

We prove the lemma for the following function

$$f(\epsilon) := \min\left\{\frac{1}{40\sqrt{\log\left(4/\epsilon\right)}}, \frac{\sqrt{\epsilon}}{30\sqrt{\log\left(4/\epsilon\right)}}, \frac{\epsilon}{84 \cdot \left(\log(4/\epsilon)\right)^{\frac{3}{2}}}, \frac{\sqrt{\epsilon}}{100 \cdot \log(4/\epsilon)}\right\}. \tag{18a}$$

Some straightforward algebra shows that $f(1/12) \geq 1/9000$.

### 4.7.1 Proof of claim (9a)

In order to bound the total variation distance $\|\mathcal{P}_x - \mathcal{P}_y\|_{\text{TV}}$, we apply Pinsker's inequality, which provides an upper bound on the TV-distance in terms of the KL Divergence:

$$\|\mathcal{P}_x - \mathcal{P}_y\|_{\text{TV}} \le \sqrt{2\text{KL}(\mathcal{P}_x\|\mathcal{P}_y)}.$$

For Gaussian distributions, the KL Divergence has a closed form expression. In particular, for two normal-distributions $\mathcal{G}_1 = \mathcal{N}(\mu_1, \Sigma_1)$ and $\mathcal{G}_2 = \mathcal{N}(\mu_2, \Sigma_2)$, the Kullback-Leibler divergence between the two is given by

$$\text{KL}(\mathcal{G}_1\|\mathcal{G}_2) = \frac{1}{2}\Big(\text{trace}\left(\Sigma_1^{-1}\Sigma_2\right) - n + \log\det\left(\Sigma_1\Sigma_2^{-1}\right) + \left(\mu_1 - \mu_2\right)^\top \Sigma_1^{-1}\left(\mu_1 - \mu_2\right)\Big).$$

Substituting $\mathcal{G}_1 = \mathcal{P}_x$ and $\mathcal{G}_2 = \mathcal{P}_y$ into the above expression and applying Pinsker's inequality, we find that

$$\begin{aligned}\|\mathcal{P}_x - \mathcal{P}_y\|_{\text{TV}}^2 \le 2\text{KL}(\mathcal{P}_y\|\mathcal{P}_x) &= \text{trace}(V_yV_x^{-1}) - n + \log\left(\frac{\det V_x}{\det V_y}\right) + \frac{\sqrt{mn}}{r^2}\|x-y\|_x \\ &= \sum_{i=1}^n\left(\lambda_i - 1 + \log\frac{1}{\lambda_i}\right) + \frac{\sqrt{mn}}{r^2}\|x-y\|_x\,, \qquad (19)\end{aligned}$$

where $\lambda_1, \ldots, \lambda_n > 0$ denote the eigenvalues of $V_yV_x^{-1}$ and we have used the facts that $\det\left(V_yV_x^{-1}\right) = \prod_{i=1}^n\lambda_i$ and $\text{trace}\left(V_yV_x^{-1}\right) = \sum_{i=1}^n\lambda_i$. The following lemma is useful to bound the expression (19).

**Lemma 6.** *For any scalar $t \in [0, \sqrt{n}/12]$ and any pair $x, y \in int(\mathcal{K})$ such that $\|x-y\|_x \le t/(mn)^{1/4}$ we have*

$$\left(1 - \frac{6t}{\sqrt{n}} + \frac{t^2}{n}\right)V_x \preceq V_y \preceq \left(1 + \frac{6t}{\sqrt{n}} + \frac{t^2}{n}\right)V_x.$$

See Appendix A.2 for its proof. For $\epsilon \in (0, 1/4]$ and $r = 1/9000$, we have $t = \epsilon r/2 \le 1/12$, whence the eigenvalues $\{\lambda_i, i \in [n]\}$ can be sandwiched as

$$1 - \frac{3\epsilon r}{\sqrt{n}} + \frac{\epsilon^2 r^2}{4n} \le \lambda_i \le 1 + \frac{3\epsilon r}{\sqrt{n}} + \frac{\epsilon^2 r^2}{4n} \quad \text{for all } i \in n. \qquad (20)$$

We are now ready to bound the TV distance between $\mathcal{P}_x$ and $\mathcal{P}_y$. Using the bound (19) and the inequality $\log\gamma \le \gamma - 1$ , valid for $\gamma > 0$, we obtain

$$\|\mathcal{P}_x - \mathcal{P}_y\|_{\text{TV}}^2 \le \sum_{i=1}^n\left(\lambda_i - 2 + \frac{1}{\lambda_i}\right) + \frac{\sqrt{mn}}{r^2}\|x-y\|_x\,.$$

Using the assumption that $\|x-y\|_x \le \epsilon r/\left(2(mn)^{1/4}\right)$, and plugging in the bounds (20) for the eigenvalues $\{\lambda_i, i \in [n]\}$, we find that

$$\sum_{i=1}^n\left(\lambda_i - 2 + \frac{1}{\lambda_i}\right) + \frac{\sqrt{mn}}{r^2}\|x-y\|_x \le \frac{141\epsilon^2 r^2}{4n} + \frac{\epsilon^2}{4}.$$

In asserting this inequality, we have used the facts that

$$\frac{1}{1 - 6\gamma + \gamma^2} \le 1 + 6\gamma + 70\gamma^2, \quad \text{and} \quad \frac{1}{1 + 6\gamma + \gamma^2} \le 1 - 6\gamma + 70\gamma^2 \quad \text{for all } \gamma \in \left[0, \tfrac{1}{12}\right].$$

Note that for any $r \in [0, 1/12]$ we have that $141r^2/(4n) \le 1/2$. Putting the pieces together yields $\|\mathcal{P}_x - \mathcal{P}_y\|_{\text{TV}} \le \epsilon$, as claimed.

### 4.7.2 Proof of claim (9b)

Note that

$$\mathcal{Q}_x(\{x\}) = \mathcal{P}_x(\mathcal{K}^c) + 1 - \int_{\mathcal{K}} \min\left\{1, \frac{p_z(x)}{p_x(z)}\right\} p_x(z)dz, \tag{21}$$

where $\mathcal{K}^c$ denotes the complement of $\mathcal{K}$. Consequently, we obtain that

$$\begin{aligned} \|\mathcal{P}_x - \mathcal{Q}_x\|_{\mathrm{TV}} &= \frac{1}{2}\left(\mathcal{Q}_x(\{x\}) + \int_{\mathbb{R}^n} p_x(z)dz - \int_{\mathcal{K}} \min\left\{1, \frac{p_z(x)}{p_x(z)}\right\} p_x(z)dz\right) \\ &= \frac{1}{2}\left(\mathcal{P}_x(\mathcal{K}^c) + 2 - 2\int_{\mathbb{R}^n} \min\left\{1, \frac{p_z(x)}{p_x(z)}\right\} p_x(z)dz + 2\int_{\mathcal{K}^c} \min\left\{1, \frac{p_z(x)}{p_x(z)}\right\} p_x(z)dz\right) \\ &\leq \underbrace{\frac{3}{2}\mathcal{P}_x(\mathcal{K}^c)}_{T_1:=} + \underbrace{1 - \mathbb{E}_{z\sim\mathcal{P}_x}\left[\min\left\{1, \frac{p_z(x)}{p_x(z)}\right\}\right]}_{T_2:=}, \end{aligned} \tag{22}$$

Thus it suffices to show that both $T_1$ and $T_2$ are small, where the probability is taken over the randomness in the proposal $z$. In particular, we show that $T_1 \leq \epsilon$ and $T_2 \leq 4\epsilon$.

**Bounding $T_1$:** Note that a random variable $z \sim \mathcal{P}_x$ can be written as

$$z \overset{d}{=} x + \frac{r}{(mn)^{1/4}} V_x^{-1/2}\xi, \tag{23}$$

where $\xi \sim \mathcal{N}(0, \mathbb{I}_n)$ and $\overset{d}{=}$ denotes equality in distribution. Using equation (23) and definition (8b) of $\theta_{x,i}$, we obtain that

$$\frac{\left(a_i^\top (z-x)\right)^2}{s_{x,i}^2} = \frac{r^2}{(mn)^{\frac{1}{2}}}\left[\frac{a_i^\top V_x^{-1/2}\xi}{s_{x,i}}\right]^2 \overset{(i)}{\leq} \frac{r^2}{(mn)^{\frac{1}{2}}}\theta_{x,i}\|\xi\|_2^2 \overset{(ii)}{\leq} \frac{r^2}{n}\|\xi\|_2^2, \tag{24}$$

where step $(i)$ follows from Cauchy-Schwarz inequality and step $(ii)$ from the bound on $\theta_{x,i}$ from Lemma 2(b). Define the events

$$\mathcal{E} := \left\{\frac{r^2}{n}\|\xi\|_2^2 < 1\right\} \quad \text{and} \quad \mathcal{E}' := \{z \in \mathrm{int}(\mathcal{K})\}.$$

Inequality (24) implies that $\mathcal{E} \subseteq \mathcal{E}'$ and hence $\mathbb{P}[\mathcal{E}'] \geq \mathbb{P}[\mathcal{E}]$. Using a standard Gaussian tail bound and noting that $r \leq \frac{1}{1+\sqrt{2/n\log(2/\epsilon)}}$, we obtain $\mathbb{P}[\mathcal{E}] \geq 1 - \epsilon/2$ and whence $\mathbb{P}[\mathcal{E}'] \geq 1 - \epsilon/2$. Thus, we have shown that $\mathbb{P}[z \notin \mathcal{K}] \leq \epsilon/2$ which implies that $T_1 \leq \epsilon$.

**Bounding $T_2$:** By Markov's inequality, we have

$$\mathbb{E}_{z\sim\mathcal{P}_x}\left[\min\left\{1, \frac{p_z(x)}{p_x(z)}\right\}\right] \geq \alpha\mathbb{P}\left[p_z(x) \geq \alpha p_x(z)\right] \quad \text{for all } \alpha \in [0, 1]. \tag{25}$$

By definition (5) of $p_x$, we obtain

$$\frac{p_z(x)}{p_x(z)} = \exp\left(-\frac{\sqrt{mn}}{2r^2}\left(\|z-x\|_z^2 - \|z-x\|_x^2\right) + \frac{1}{2}\left(\log\det V_z - \log\det V_x\right)\right).$$

The following lemma provides us with useful bounds on the two terms in this expression.

**Lemma 7.** *For any $\epsilon \in (0, \frac{1}{4}]$ and $x \in int(\mathcal{K})$, if $r \le f(\epsilon)$, we have*

$$\mathbb{P}_{z \sim \mathcal{P}_x}\left[\frac{1}{2}\log\det V_z - \frac{1}{2}\log\det V_x \ge -\epsilon\right] \ge 1 - \epsilon, \text{ and} \tag{26a}$$

$$\mathbb{P}_{z \sim \mathcal{P}_x}\left[\|z - x\|_z^2 - \|z - x\|_x^2 \le 2\epsilon\frac{r^2}{\sqrt{mn}}\right] \ge 1 - \epsilon. \tag{26b}$$

The proof of this lemma is provided in Appendix B. Using Lemma 7, we now complete the proof. For $r \le f(\epsilon)$, we obtain $p_z(x)/p_x(z) \ge \exp(-2\epsilon) \ge 1 - 2\epsilon$ with probability at least $1 - 2\epsilon$. Substituting $\alpha = 1 - 2\epsilon$ in inequality (25) yields the bound $T_2 \le 4\epsilon$.

## 5 Discussion

In this section, we discuss some applications of random walks on polytopes and possible extensions for Vaidya walk.

### 5.1 Applications

Access to a samples from a distribution have numerous applications in practice. In particular, uniform samples from a polytope can be used for approximating the integral of an arbitrary function. Besides, algorithms like multi-phase Monte Carlo for volume computation of convex set using sampling methods have been extensively analyzed in the literature [Law91, Kha93, LV06c, CV14]. Since, a faster sampling algorithm leads to a better volume algorithm and thus our method is a natural candidate for such a task for appropriate regime of $m$ and $n$. In principle, a sampling algorithm can be used to design a randomized algorithm for convex optimization. And for such problems, a faster sampling algorithm implies a faster optimization method. However, as such algorithms are often not useful in practice, we omit any further discussion. Dikin walk was shown to perform effectively on numerical experiments for approximately solving mixed integer programs [HM13]. We expect Vaidya walk to be more effective than Dikin walk under the same set-up owing to the faster convergence rates.

### 5.2 Extensions

In this work, we specialized our discussion to the polytopes. The volumetric-logarithmic barrier for polytopes has a smaller self-concordance parameter than the log-barrier [Vai89]. This fact was the primary source of the speed up in convergence rate of Vaidya walk over Dikin walk on a polytope. This nice self-concordance property and several other key properties exhibited by the combined barrier for polytopes were extended by Anstreicher [Ans00] to more general convex sets defined by semidefinite-constraints, namely, linear matrix inequality (LMI) constraints. Moreover, Narayanan [Nar16] showed that for a convex set defined by LMI constraints and equipped with the log-det barrier, the mixing time of Dikin walk from an $M$-warm start is $k_{\text{mix}}(\delta) = \mathcal{O}\left(mn^2\log(M/\delta)\right)$. It is possible that an appropriate Vaidya walk on such sets would have a speedup over Dikin walk. Also Narayanan et al. [NR10] use Dikin walk to generate samples from time varying log-concave distributions with appropriate scaling of the radius for difference class of distributions. It would be interesting to see if a suitable adaptation of Vaidya walk for such cases would provide a significant gain.

Another possible extension of our work can be a new random walk on Riemannian manifolds based on the matrix $V_x$, in contrast to the Geodesic walk [LV16] where the manifold is based on the Hessian $\nabla^2 F_x$. In contrast to Dikin walk's $\mathcal{O}(mn)$ mixing time, the Geodesic walk has an $\mathcal{O}\left(mn^{3/4}\right)$ dependence on mixing time. It would be interesting to see whether a geodesic version of Vaidya walk has a convergence rate of $\mathcal{O}\left(m^{1/2}n^{5/4}\right)$. Random walk based on Hessian sketch [PW15a, PW15b] is another possible line of work, that can reduce the per iteration complexity and there by speedup the random walk.

The results in the paper provide an upper bound for the mixing time. Also, we observe that the scaling of the mixing time matches the stated results for the hypercube. This observation hints at the possibility that hypercube is the worst case for ball-walk like sampling algorithms. A result providing a lower bound for a general class of convex sets can provide a more formal backing to this conjecture and furthermore provide more insight in designing better algorithms. We believe that our work will be seen as yet another step to unifying the optimization and sampling algorithms and providing theoretical insight to inform practice for MCMC algorithms.

# A Technical results and Proof of Lemma 6

In this appendix, we first summarize some auxiliary results involved in the proofs of Lemma 6 and 7. We begin with introducing some notation. Recall $A \in \mathbb{R}^{m\times n}$ is a matrix with $a_i^\top$ as its $i$-th row vector. For any positive integer $p$ and any vector $v = (v_1, \ldots, v_p)^\top$, $\text{diag}(v) = \text{diag}(v_1, \ldots, v_p)$ denotes a $p \times p$ diagonal matrix, with $i$-th entry on the diagonal equal to $v_i$. Let $S_x$ be defined as follows:

$$S_x = \text{diag}\left(s_{x,1}, \ldots, s_{x,m}\right) \text{ where } s_{x,i} = b_i - a_i^\top x \text{ for each } i \in [m]. \tag{27}$$

It is easy to see that $S_x$ is positive semidefinite for all $x \in \mathcal{K}$ and moreover that $S_x$ is strictly positive definite for all $x \in \text{int}\,(\mathcal{K})$. Furthermore, define $A_x = S_x^{-1}A$ for all $x \in \text{int}\,(\mathcal{K})$, and let $\Upsilon_x$ denote the projection matrix for the column space of $A_x$, i.e.,

$$\Upsilon_x := A_x(A_x^\top A_x)^{-1}A_x^\top = A_x \nabla^2 F_x^{-1} A_x^\top. \tag{28}$$

Note that for the scores $\sigma_x$ (3), we have $\sigma_{x,i} = (\Upsilon_x)_{ii}$ for each $i \in [m]$. Let $\Sigma_x$ be an $m \times m$ diagonal matrix defined as

$$\Sigma_x = \text{diag}\left(\sigma_{x,1}, \ldots, \sigma_{x,m}\right). \tag{29}$$

Let $\sigma_{x,i,j} := (\Upsilon_x)_{ij}$, and let $\Upsilon_x^{(2)}$ denote the Hadamard product of $\Upsilon_x$ with itself, i.e.,

$$(\Upsilon_x^{(2)})_{ij} = \sigma_{x,i,j}^2 = \frac{\left(a_i^\top \nabla^2 F_x^{-1} a_j\right)^2}{s_{x,i}^2 s_{x,j}^2} \quad \text{for all } i, j \in [m]. \tag{30}$$

Define the matrices $\Theta_x$ and $\Xi_x$

$$\Theta_x := \text{diag}\left(\theta_{x,1}, \ldots, \theta_{x,m}\right) \text{ where } \theta_{x,i} = \frac{a_i^\top V_x^{-1} a_i}{s_{x,i}^2} \quad \text{for } i \in [m], \text{ and}$$

$$\Xi_x := (\theta_{x,i,j}^2) \text{ where } \theta_{x,i,j}^2 = \frac{\left(a_i^\top V_x^{-1} a_j\right)^2}{s_{x,i}^2 s_{x,j}^2} \quad \text{for } i, j \in [m].$$

In our new notation, we can re-write $V_x = A_x^\top \left(\Sigma_x + \beta \mathbb{I}\right) A_x$.

### A.1 Basic Properties

First we summarize some key properties of various terms.

**Lemma 8.** *For any $x \in int(\mathcal{K})$, the following properties hold:*

*(a)* $\sigma_{x,i} = \sum_{j=1}^{m} \sigma_{x,i,j}^{2} = \sum_{j,k=1}^{m} \sigma_{x,i,j}\sigma_{x,j,k}\sigma_{x,k,i}$ *for each* $i \in [m]$*,*

*(b)* $\Sigma_x \succeq \Upsilon_x^{(2)}$*,*

*(c)* $\sum_{i=1}^{m} \theta_{x,i} \left(\sigma_{x,i} + \beta\right) = n$*,*

*(d)* $\forall i \in [m],\ \theta_{x,i} = \sum_{j=1}^{m} \left(\sigma_{x,j} + \beta\right) \theta_{x,i,j}^{2}$*, for each* $i \in [m]$*,*

*(e)* $\theta_x^{\top} \left(\Sigma_x + \beta \mathbb{I}\right) \theta_x = \sum_{i=1}^{m} \theta_{x,i}^{2} \left(\sigma_{x,i} + \beta\right) \le \sqrt{mn}$*, and*

*(f)* $\beta\, \nabla^2 F_x \preceq V_x \preceq \left(1 + \beta\right) \nabla^2 F_x$*.*

*Proof.* We prove each property separately.

**Part (a):** Using $\mathbb{I}_n = \nabla^2 F_x \left(\nabla^2 F_x\right)^{-1}$, we find that

$$\sigma_{x,i} = \frac{a_i^{\top} \left(\nabla^2 F_x\right)^{-1} \nabla^2 F_x \left(\nabla^2 F_x\right)^{-1} a_i}{s_{x,i}^{2}} = \frac{a_i^{\top} \left(\nabla^2 F_x\right)^{-1} \nabla^2 \sum_{j=1}^{m} \frac{a_j^{\top} a_j}{s_{x,j}^{2}} \left(\nabla^2 F_x\right)^{-1} a_i}{s_{x,i}^{2}} = \sum_{i,j=1}^{m} \sigma_{x,i,j}^{2}.$$

Applying a similar trick twice and performing some algebra, we obtain

$$\sigma_{x,i} = \frac{a_i^{\top} \left(\nabla^2 F_x\right)^{-1} \nabla^2 F_x \left(\nabla^2 F_x\right)^{-1} \nabla^2 F_x \left(\nabla^2 F_x\right)^{-1} a_i}{s_{x,i}^{2}} = \sum_{i,j,k=1}^{m} \sigma_{x,i,j}\sigma_{x,j,k}\sigma_{x,k,i}.$$

**Part (b):** From part (a), we have that $\Sigma_x - \Upsilon_x^{(2)}$ is a symmetric and diagonally dominant matrix with non-negative entries on the diagonal. Applying Gershgorin Disk Theorem we conclude that it is PSD (see, e.g., the books [Bha13, HJ12] for more details on the theorem).

**Part (c):** Since $\mathrm{trace}(AB) = \mathrm{trace}(BA)$, we have

$$\sum_{i=1}^{m} \theta_{x,i} \left(\sigma_{x,i} + \beta\right) = \mathrm{trace}\left( V_x^{-1} \sum_{i=1}^{m} \left(\sigma_{x,i} + \beta\right) \frac{a_i a_i^{\top}}{s_{x,i}^{2}} \right) = \mathrm{trace}\left(\mathbb{I}_n\right) = n.$$

**Part (d):** An argument similar to part (a) implies that

$$\theta_{x,i} = \frac{a_i^{\top} V_x^{-1} V_x V_x^{-1} a_i}{s_{x,i}^{2}} = \frac{a_i^{\top} V_x^{-1} \sum_{j=1}^{m} \left(\sigma_{x,i} + \beta\right) \frac{a_j^{\top} a_j}{s_{x,j}^{2}} V_x^{-1} a_i}{s_{x,i}^{2}} = \sum_{i,j=1}^{m} \left(\sigma_{x,i} + \beta\right) \theta_{x,i,j}^{2}.$$

**Part (e):** Using part (c) and Lemma 2(b) yields the claim.

**Part (f):** The left inequality is by the definition of $V_x$. The right inequality uses the fact that $\Sigma_x \preceq \mathbb{I}_n$. □

We now prove an important result that relates the *slackness* $s_x$ and $s_y$ at two points, in terms of $\|x-y\|_x$.

**Lemma 9.** *For all $x, y \in int(\mathcal{K})$, we have*

$$\left|1-\frac{s_{y,i}}{s_{x,i}}\right| \leq \left(\frac{m}{n}\right)^{\frac{1}{4}} \|x-y\|_x \quad \textit{for each } i \in [m].$$

*Proof.* Let $x, y \in \operatorname{int}(\mathcal{K})$ be such that $\|x-y\|_x \leq \delta$, i.e.,

$$(x-y)^\top V_x (x-y) \leq \delta^2.$$

Since $\operatorname{trace}(AB) = \operatorname{trace}(BA)$, we have

$$\operatorname{trace}\left(V_x^{\frac{1}{2}} (x-y)(x-y)^\top V_x^{\frac{1}{2}}\right) \leq \delta^2.$$

We observe that $V_x^{\frac{1}{2}} (x-y)(x-y)^\top V_x^{\frac{1}{2}}$ is a rank one matrix, whence

$$V_x^{\frac{1}{2}} (x-y)(x-y)^\top V_x^{\frac{1}{2}} \preceq \delta^2 \mathbb{I}.$$

Multiplying $V_x^{-\frac{1}{2}}$ on left and right (and thereby preserving the matrix order), we obtain

$$(x-y)(x-y)^\top \preceq \delta^2 V_x^{-1}.$$

In particular, we have

$$\left(a_i^\top (x-y)\right)^2 \leq \delta^2 a_i^\top V_x^{-1} a_i = \delta^2 \theta_{x,i} s_{x,i}^2 \leq \delta^2 \sqrt{\frac{m}{n}} s_{x,i}^2 \quad \text{for all } i \in [m],$$

where for the last inequality we have used the bound on $\theta_{x,i}$ from Lemma 2(b). Noting the fact that $a_i^\top (x-y) = s_{y,i} - s_{x,i}$, the claim follows after simple algebra. □

We are now ready to prove Lemma 6.

### A.2 Proof of Lemma 6

As a direct consequence of Lemma 9, we find that

$$\left|1-\frac{s_{y,i}}{s_{x,i}}\right| \leq \frac{t}{\sqrt{n}}, \quad \text{for any } x, y \in \operatorname{int}(\mathcal{K}) \text{ such that } \|x-y\|_x \leq \frac{t}{(mn)^{1/4}}.$$

The Hessian $\nabla^2 F_y$ is thus sandwiched as

$$\left(1-\frac{t}{\sqrt{n}}\right)^2 \nabla^2 F_x \preceq \nabla^2 F_y \preceq \left(1+\frac{t}{\sqrt{n}}\right)^2 \nabla^2 F_x.$$

By the definition of $\sigma_{x,i}$ and $\sigma_{y,i}$, we have

$$\frac{\left(1-\frac{t}{\sqrt{n}}\right)^2}{\left(1+\frac{t}{\sqrt{n}}\right)^2}\sigma_{x,i} \leq \sigma_{y,i} \leq \frac{\left(1+\frac{t}{\sqrt{n}}\right)^2}{\left(1-\frac{t}{\sqrt{n}}\right)^2}\sigma_{x,i} \quad \text{for all } i \in [m]. \tag{31}$$

Consequently, we find that

$$\frac{\left(1-\frac{t}{\sqrt{n}}\right)^2}{\left(1+\frac{t}{\sqrt{n}}\right)^4}V_x \preceq V_y \preceq \frac{\left(1+\frac{t}{\sqrt{n}}\right)^2}{\left(1-\frac{t}{\sqrt{n}}\right)^4}V_x.$$

Note that

$$\frac{(1-\gamma)^2}{(1+\gamma)^4} \geq 1-6\gamma+\gamma^2, \quad \text{and} \quad \frac{(1+\gamma)^2}{(1-\gamma)^4} \leq 1+6\gamma+\gamma^2 \quad \text{for any } \gamma \in \left[0, \tfrac{1}{12}\right].$$

Applying these inequalities with $\gamma = t/\sqrt{n}$ yields

$$\left(1-\frac{6t}{\sqrt{n}}+\frac{t^2}{n}\right)V_x \preceq V_y \preceq \left(1+\frac{6t}{\sqrt{n}}+\frac{t^2}{n}\right)V_x \quad \text{for any } \frac{t}{\sqrt{n}} \in \left[0, \frac{1}{12}\right],$$

which completes the proof.

## B Proof of Lemma 7

We begin by defining

$$\varphi_{x,i} := \frac{\sigma_{x,i}+\beta}{s_{x,i}^2} \text{ for } i \in [m], \quad \text{and} \quad \Psi_x := \frac{1}{2}\log\det V_x, \quad \text{for all } x \in \text{int}\left(\mathcal{K}\right). \tag{32}$$

Further, for any two points $x$ and $z$, let $\overline{xz}$ denote the set of points on the line segment joining $x$ and $z$. The proof of Lemma 7 relies on Taylor Series expansion which in turn requires careful handling of $\sigma, \varphi, \Psi$ and their derivatives. Such a technique was also used in the proofs of Dikin walk [KN12, Nar16, SV16]. The road-map of the proof is as follows: (1) expand the desired quantities around $x$ along $\overline{xz}$ using Taylor expansion where the resulting expressions depend on derivatives, (2) transfer the bounds of terms involving some arbitrary $y \in \overline{xz}$ to terms involving only $x$ and $z$ and then (3) use concentration of Gaussian polynomials to obtain high probability bounds.

We now discuss discuss the auxiliary results that we use for the different steps outlined in the road-map. The following lemma provides expressions for gradients of $\sigma, \varphi$ and $\Psi$ and bounds for directional Hessian of $\varphi$ and $\Psi$.

**Lemma 10.** *Let $e_i \in \mathbb{R}^n$ denote a vector with 1 in the i-th position and 0 otherwise. For any $h \in \mathbb{R}^n$ and $x \in int(\mathcal{K})$, let $\eta_{x,h,i} = \eta_{x,i} := a_i^\top h/s_{x,i}$, for each $i \in [m]$. Then the following statements are true:*

*(a) Gradient of $\sigma$: $\nabla\sigma_{x,i} = 2A_y^\top(\Sigma_x - P_x^{(2)})e_i$ for each $i \in [m]l$*

*(b) Gradient of $\varphi$: $\nabla\varphi_{x,i} = \dfrac{2}{s_{y,i}^2}A_x^\top\left[2\Sigma_x + \beta\,\mathbb{I} - P_x^{(2)}\right]e_i$ for each $i \in [m]$;*

(c) *Gradient of* $\Psi$*:* $\nabla\Psi_x = A_x^\top\left(2\,\Sigma_x + \beta\,\mathbb{I} - P_x^{(2)}\right)\theta_x$*;*

(d) *Bound on* $\nabla^2\varphi$*:* $s_{x,i}^2\left|\frac{1}{2}h^\top\nabla^2\varphi_{x,i}h\right| \le 14\,(\sigma_{x,i}+\beta)\,\eta_{x,i}^2 + 11\sum_{j=1}^m \sigma_{x,i,j}^2\eta_{x,j}^2$ *for* $i \in [m]$*;*

(e) *Bound on* $\nabla^2\Psi$*:* $\left|\frac{1}{2}h^\top\left(\nabla^2\Psi_x\right)h\right| \le 13\sum_{i=1}^m(\sigma_{x,i}+\beta)\,\theta_{x,i}\eta_{x,i}^2 + \frac{17}{2}\sum_{i,j=1}^m \sigma_{x,i,j}^2\theta_{x,i}\eta_{x,j}^2$*.*

The proof is deferred to Section B.4. In the discussion that follows we would always assume that the parameter $r$ is positive. The following lemma that shows that for a random variable $z \sim \mathcal{P}_x$, the slackness $s_{z,i}$ is close to $s_{x,i}$ with high probability.

**Lemma 11.** *For any* $\epsilon \in (0, 1/4], r \in (0, 1)$ *and* $x \in int(\mathcal{K})$*, we have*

$$\mathbb{P}_{z\sim\mathcal{P}_x}\left[\forall i \in [m], \forall v \in \overline{xz},\ \frac{s_{x,i}}{s_{v,i}} \in \left(1-\sqrt{1+\delta}\,r, 1+\sqrt{1+\delta}\,r\right)\right] \ge 1-\epsilon/4,$$

*where* $\delta = \sqrt{\frac{2}{n}\log\left(\frac{4}{\epsilon}\right)}$*. Thus for any* $n \ge 1$ *and* $r \le 1/(20(1+\sqrt{2}\log(4/\epsilon))^{1/2}$*, we have*

$$\mathbb{P}_{z\sim\mathcal{P}_x}\left[\forall i \in [m], \forall v \in \overline{xz},\ \frac{s_{x,i}}{s_{v,i}} \in (0.95, 1.05)\right] \ge 1-\epsilon/4.$$

This result comes in handy for transferring bounds for different expressions in Taylor expansion involving an arbitrary $y$ on $\overline{xz}$ to bounds on terms involving simply $x$. The proof follows from Lemma 9 and a simple application of the standard Gaussian tail bounds and is thereby omitted. For brevity, we define the shorthand

$$\hat{a}_i = \frac{1}{s_{x,i}}V_x^{-1/2}a_i \quad \text{for each } i \in [m], \tag{33}$$

where we have omitted the dependence of $\hat{a}_i$ on $x$. In the following lemma we state the tail bounds for useful Gaussian polynomials. The proof is deferred to Section B.3.

**Lemma 12.** *For any* $\epsilon \in (0, 1/4]$*, define* $\gamma_k = \max\left\{2e, 2e/k\cdot\log(4/\epsilon)\right\}^{k/2}$ *for* $k = 2, 3$ *and* 4*. Then for* $\xi \sim \mathcal{N}(0, \mathbb{I}_n)$ *and any* $x \in int(\mathcal{K})$ *the following high probability bounds hold:*

$$\mathbb{P}\left[\sum_{i=1}^m(\sigma_{x,i}+\beta)\left(\hat{a}_i^\top\xi\right)^2 \le \frac{\epsilon}{90}\frac{n}{r^2}\right] \ge 1-\frac{\epsilon}{4} \quad \textit{for any } r \le \sqrt{\frac{\epsilon}{90\sqrt{3}\gamma_2}}, \tag{34a}$$

$$\mathbb{P}\left[\left|\sum_{i=1}^m(\sigma_{x,i}+\beta)\left(\hat{a}_i^\top\xi\right)^3\right| \le \frac{\epsilon}{7}\frac{(mn)^{\frac{1}{4}}}{r}\right] \ge 1-\frac{\epsilon}{4} \quad \textit{for any } r \le \frac{\epsilon}{7\sqrt{15}\gamma_3}, \tag{34b}$$

$$\mathbb{P}\left[\left|\sum_{i,j=1}^m\sigma_{x,i,j}^2\left(\left(\frac{\hat{a}_i+\hat{a}_j}{2}\right)^\top\xi\right)^3\right| \le \frac{\epsilon}{4}\frac{(mn)^{\frac{1}{4}}}{r}\right] \ge 1-\frac{\epsilon}{4} \quad \textit{for any } r \le \frac{\epsilon}{4\sqrt{15}\gamma_3}, \textit{ and} \tag{34c}$$

$$\mathbb{P}\left[\sum_{i=1}^m(\sigma_{x,i}+\beta)\left(\hat{a}_i^\top\xi\right)^4 \le \frac{\epsilon}{100}\frac{\sqrt{mn}}{r^2}\right] \ge 1-\frac{\epsilon}{4} \quad \textit{for any } r \le \sqrt{\frac{\epsilon}{100\sqrt{105}\gamma_4}}. \tag{34d}$$

Now we summarize the final ingredients needed for our proofs. We recall equation (23) that relates the proposal $z$ and the current state $x$

$$z \overset{d}{=} x + \frac{r}{(mn)^{1/4}}V_x^{-1/2}\xi,$$

where $\xi \sim \mathcal{N}(0, \mathbb{I}_n)$. We also use the following fundamental inequalities:

$$\text{Cauchy-Schwarz inequality:} \qquad |u^\top v| \le \|u\|_2 \|v\|_2 \tag{C-S}$$

$$\text{Sum of squares inequality:} \qquad \|a+b\|_2^2 \le 2\left(\|a\|_2^2 + \|b\|_2^2\right), \tag{SSI}$$

$$\text{AM-GM inequality:} \qquad \omega\kappa \le \frac{1}{2}(\omega^2 + \kappa^2). \tag{AM-GM}$$

Note that SSI and AM-GM are equivalent. We are now ready for Lemma 7 which we now prove part-wise.

## B.1 Proof of claim (26a)

Using the second degree Taylor expansion, we have

$$\Psi_z - \Psi_x = (z-x)^\top \nabla\Psi_x + \frac{1}{2}(z-x)^\top \nabla^2\Psi_y (z-x), \quad \text{for some } y \in \overline{xz}.$$

We claim that for $r \le f(\epsilon)$, we have

$$\mathbb{P}\left[(z-x)^\top \nabla\Psi_x \ge -\epsilon/2\right] \ge 1 - \epsilon/2, \text{ and} \tag{35a}$$

$$\mathbb{P}\left[\frac{1}{2}(z-x)\nabla^2\Psi_y(z-x) \ge -\epsilon/2\right] \ge 1 - \epsilon/2. \tag{35b}$$

The result follows from these claims which we now prove.

### B.1.1 Proof of bound (35a)

Equation (23) implies that

$$(z-x)^\top \nabla\Psi_x \sim \mathcal{N}\left(0, \frac{r^2}{\sqrt{mn}} \nabla\Psi_x^\top V_x^{-1} \nabla\Psi_x\right).$$

We claim that

$$\nabla\Psi_x^\top V_x^{-1}\nabla\Psi_x \le 9\sqrt{mn} \quad \text{for all } x \in \text{int}(\mathcal{K}). \tag{36}$$

We prove this inequality at the end of this subsection. Taking it as given for now, let $\xi' \sim \mathcal{N}(0, 9r^2)$. Then using inequality (36) and a standard Gaussian tail bound, we find that

$$\mathbb{P}\left[(z-x)^\top \nabla\Psi_x \ge -\gamma\right] \ge \mathbb{P}\left[\xi' \ge -\gamma\right] \ge 1 - \exp(-\gamma^2/(18r^2)), \quad \text{valid for all } \gamma \ge 0.$$

Setting $\gamma = \epsilon/2$ and noting that $r \le \frac{\epsilon}{\sqrt{18\log(2/\epsilon)}}$ completes the claim.

### B.1.2 Proof of bound (35b)

Let $\eta_{x,i} = \frac{a_i^\top(z-x)}{s_{x,i}} = \frac{r}{(mn)^{\frac{1}{4}}}\hat{a}_i^\top \xi$. Using Lemma 10(e), we have

$$\begin{aligned}\left|\frac{1}{2}(z-x)\nabla^2\Psi_y(z-x)\right| &\le 13\sum_{i=1}^m (\sigma_{y,i}+\beta)\,\theta_{y,i}\frac{s_{x,i}^2}{s_{y,i}^2}\eta_{x,i}^2 + \frac{17}{2}\sum_{i,j=1}^m \sigma_{y,i,j}^2\theta_{y,i}\frac{s_{x,j}^2}{s_{y,j}^2}\eta_{x,j}^2 \\ &\le \frac{43}{2}\sqrt{\frac{m}{n}}\sum_{i=1}^m (\sigma_{x,i}+\beta)\frac{(\sigma_{y,i}+\beta)}{(\sigma_{x,i}+\beta)}\frac{s_{x,i}^2}{s_{y,i}^2}\eta_{x,i}^2. \end{aligned} \tag{37}$$

Now, let $\tau = 1.05$ and define the events $\mathcal{E}_1$ and $\mathcal{E}_2$ as follows:

$$\mathcal{E}_1 = \left\{ \forall i \in [m],\ \frac{s_{x,i}}{s_{y,i}} \in [2-\tau, \tau] \right\}, \quad \text{and} \tag{38a}$$

$$\mathcal{E}_2 = \left\{ \forall i \in [m],\ \frac{\sigma_{x,i}}{\sigma_{y,i}} \in \left[0, \frac{\tau^2}{(2-\tau)^2}\right] \right\}. \tag{38b}$$

It is straightforward to see that $\mathcal{E}_1 \subseteq \mathcal{E}_2$. Since $r \leq \frac{1}{20\sqrt{1+\sqrt{2}\log(4/\epsilon)}}$, Lemma 11 implies that $\mathbb{P}\left[\mathcal{E}_1\right] \geq 1 - \epsilon/4$ whence $\mathbb{P}\left[\mathcal{E}_2\right] \geq 1 - \epsilon/4$. Using these high probability bounds and $\tau = 1.05$, we obtain that with probability at least $1 - \epsilon/4$

$$\sqrt{\frac{m}{n}} \sum_{i=1}^{m} (\sigma_{x,i} + \beta) \frac{(\sigma_{y,i} + \beta)\, s_{x,i}^2}{(\sigma_{x,i} + \beta)\, s_{y,i}^2} \eta_{x,i}^2 \leq 2\sqrt{\frac{m}{n}} \sum_{i=1}^{m} (\sigma_{x,i} + \beta)\, \eta_{x,i}^2 = \frac{2r^2}{n} \sum_{i=1}^{m} (\sigma_{x,i} + \beta) \left(\hat{a}_i^\top \xi\right)^2. \tag{39}$$

Since $r$ satisfies the assumptions of Lemma 12, we can apply the high probability bound (34a). Combining it with the bounds (37) and (39) yields the claim.

### B.1.3 Proof of bound (36)

We now return to prove our earlier inequality (36). Using the expression for the gradient $\nabla\Psi_x$ from Lemma 10(c), we have that for any vector $u \in \mathbb{R}^n$

$$\begin{aligned}
u^\top \nabla\Psi_x \nabla\Psi_x^\top u &= \left\langle u, A_x^\top \left(2\Sigma_x - \Upsilon_x^{(2)} + \beta\mathbb{I}\right) \theta_x \right\rangle^2 \\
&= \left\langle A_x u, \left(2\Sigma_x - \Upsilon_x^{(2)} + \beta\mathbb{I}\right) \theta_x \right\rangle^2 \\
&= \left\langle \left(\Sigma_x + \beta\mathbb{I}\right)^{\frac{1}{2}} A_x u, \left(\Sigma_x + \beta\mathbb{I}\right)^{-1/2} \left(2\Sigma_x - \Upsilon_x^{(2)} + \beta\mathbb{I}\right) \theta_x \right\rangle^2 \\
&\leq u^\top V_x u \cdot \theta_x^\top \left(2\Sigma_x - \Upsilon_x^{(2)} + \beta\mathbb{I}\right) \left(\Sigma_x + \beta\mathbb{I}\right)^{-1} \left(2\Sigma_x - \Upsilon_x^{(2)} + \beta\mathbb{I}\right) \theta_x
\end{aligned} \tag{40}$$

where the last step follows from the Cauchy-Schwarz inequality. As a consequence of Lemma 8(b), the matrix $\Sigma_x + \Upsilon_x^{(2)}$ is a diagonally dominant matrix with positive entries on the diagonal, and hence it is a PSD matrix. Thus, we have

$$2\Sigma_x - \Upsilon_x^{(2)} + \beta\mathbb{I} \preceq 3\left(\Sigma_x + \beta\mathbb{I}\right).$$

Consequently, we find that

$$\underbrace{\left(3\Sigma_x + 3\beta\mathbb{I}\right)^{-1/2} \left(2\Sigma_x - \Upsilon_x^{(2)} + \beta\mathbb{I}\right) \left(3\Sigma_x + 3\beta\mathbb{I}\right)^{-1/2}}_{=:L} \preceq \mathbb{I}.$$

We deduce that all eigenvalues of the matrix $L$ are less than 1 and hence all the eigenvalues of the matrix $L^2$ belong to the interval $[0, 1]$. As a result, we have

$$\left(2\Sigma_x - \Upsilon_x^{(2)} + \beta\mathbb{I}\right) \left(3\Sigma_x + 3\beta\mathbb{I}\right)^{-1} \left(2\Sigma_x - \Upsilon_x^{(2)} + \beta\mathbb{I}\right) \preceq \left(3\Sigma_x + 3\beta\mathbb{I}\right).$$

Thus, we obtain

$$\theta_x^\top \left(2\Sigma_x - \Upsilon_x^{(2)} + \beta\mathbb{I}\right) \left(\Sigma_x + \beta\mathbb{I}\right)^{-1} \left(2\Sigma_x - \Upsilon_x^{(2)} + \beta\mathbb{I}\right) \theta_x \leq 9\theta_x^\top \left(\Sigma_x + \beta\mathbb{I}\right) \theta_x. \tag{41}$$

Finally, applying Lemma 8(e) and combining bounds (40) and (41) yields the claim.

### B.2 Proof of claim (26b)

The quantity of interest can be written as

$$\|z-x\|_z^2 - \|z-x\|_x^2 = \sum_{i=1}^{m} \left(a_i^\top (z-x)\right)^2 (\varphi_{z,i} - \varphi_{x,i}) .$$

We can write $z = x + \alpha u$, where $\alpha$ is a scalar and $u$ is a unit vector in $\mathbb{R}^n$. Then we have

$$\|z-x\|_z^2 - \|z-x\|_x^2 = \alpha^2 \sum_{i=1}^{m} \left(a_i^\top u\right)^2 (\varphi_{z,i} - \varphi_{x,i}) .$$

We use Taylor's series expansion for $\sum_{i=1}^{m} \left(a_i^\top u\right)^2 (\varphi_{z,i} - \varphi_{x,i})$ around the point $x$, along the line $u$. There exists a point $y \in \overline{xz}$ such that

$$\sum_{i=1}^{m} \left(a_i^\top u\right)^2 (\varphi_{z,i} - \varphi_{x,i}) = \sum_{i=1}^{m} \left(a_i^\top u\right)^2 \left( (z-x)^\top \nabla \varphi_{x,i} + \frac{1}{2} (z-x)^\top \nabla^2 \varphi_{y,i} (z-x) \right) .$$

Multiplying both sides by $\alpha^2$, and using the shorthand $\eta_{x,i} = \frac{a_i^\top (z-x)}{s_{x,i}}$, we obtain

$$\|z-x\|_z^2 - \|z-x\|_x^2 = \sum_{i=1}^{m} \eta_{x,i}^2 s_{x,i}^2 (z-x)^\top \nabla \varphi_{x,i} + \sum_{i=1}^{m} \eta_{x,i}^2 s_{x,i}^2 \frac{1}{2} (z-x)^\top \nabla^2 \varphi_{y,i} (z-x) . \quad (42)$$

Substituting the expression for $\nabla \varphi_{x,i}$ from Lemma 10(b) in equation (42) and performing some algebra, the first term on the RHS of equation (42) can be written as

$$\sum_{i=1}^{m} \eta_{x,i}^2 s_{x,i}^2 (z-x)^\top \nabla \varphi_{x,i} = 2 \sum_{i=1}^{m} \left( \frac{7}{3} \sigma_{x,i} + \beta \right) \eta_{x,i}^3 - \frac{1}{3} \sum_{i,j=1}^{m} \sigma_{x,i,j}^2 \left(\eta_{x,i} + \eta_{x,j}\right)^3 . \quad (43)$$

On the other hand, using Lemma 10(d), we have

$$\frac{1}{2} s_{x,i}^2 \left| (z-x)^\top \nabla^2 \varphi_{y,i} (z-x) \right| \leq \frac{s_{x,i}^2}{s_{y,i}^2} \left[ 14 (\sigma_{y,i} + \beta) \frac{s_{x,i}^2}{s_{y,i}^2} \eta_{x,i}^2 + 11 \left( \sum_{j=1}^{m} \sigma_{y,i,j}^2 \eta_{x,j}^2 \frac{s_{x,j}^2}{s_{y,j}^2} \right) \right] . \quad (44)$$

Now, we use a fourth degree Gaussian polynomials to bound both the terms on the RHS of inequality (44). To do so, we use high probability bound for $s_{x,i}/s_{y,i}$. In particular, we use the high probability bounds for the events $\mathcal{E}_1$ and $\mathcal{E}_2$ defined in equations (38a) and (38b). Multiplying both sides of inequality (44) by $\eta_{x,i}^2$ and summing over the index $i$, we obtain that with probability at least $1-\epsilon/4$, we have

$$\begin{aligned}
\sum_{i=1}^{m} \eta_{x,i}^2 s_{x,i}^2 \left| \frac{1}{2} (z-x)^\top \nabla^2 \varphi_{y,i} (z-x) \right| \quad &\leq \quad \left[ 14 \sum_{i=1}^{m} (\sigma_{y,i} + \beta) \frac{s_{x,i}^4}{s_{y,i}^4} \eta_{x,i}^4 + 11 \sum_{i,j=1}^{m} \sigma_{y,i,j}^2 \eta_{x,i}^2 \eta_{x,j}^2 \frac{s_{x,i}^2 s_{x,j}^2}{s_{y,i}^2 s_{y,j}^2} \right] \\
&\overset{\text{(hpb.(38a))}}{\leq} \tau^4 \left[ 14 \sum_{i=1}^{m} (\sigma_{y,i} + \beta) \eta_{x,i}^4 + 11 \sum_{i,j=1}^{m} \sigma_{y,i,j}^2 \eta_{x,i}^2 \eta_{x,j}^2 \right] \\
&\overset{\text{(AM}-\text{GM)}}{\leq} \tau^4 \left[ 14 \sum_{i=1}^{m} (\sigma_{y,i} + \beta) \eta_{x,i}^4 + \frac{11}{2} \sum_{i,j=1}^{m} \sigma_{y,i,j}^2 (\eta_{x,i}^4 + \eta_{x,j}^4) \right] \\
&\overset{\text{(Lem. 8(a))}}{\leq} 25 \tau^4 \sum_{i=1}^{m} (\sigma_{y,i} + \beta) \eta_{x,i}^4 \\
&\overset{\text{(hpb.(38b))}}{\leq} 50 \sum_{i=1}^{m} (\sigma_{x,i} + \beta) \eta_{x,i}^4, \qquad (45)
\end{aligned}$$

where "hpb" stands for high probability bound for events $\mathcal{E}_1$ and $\mathcal{E}_2$. In the last step, we have used the fact that $\tau^6/(2-\tau)^2 \leq 2$ for $\tau = 1.05$. Combining equations (42), (43) and (45) and noting that $\eta_{x,i} = r\hat{a}_i^\top \xi/(mn)^{1/4}$, we find that

$$\begin{aligned}
\left| \|z-x\|_z^2 - \|z-x\|_x^2 \right| &\leq \frac{14}{3} \left| \sum_{i=1}^{m} (\sigma_{x,i} + \beta) \eta_{x,i}^3 \right| + \frac{8}{3} \left| \sum_{i,j=1}^{m} \sigma_{x,i,j}^2 \left( (\eta_{x,i} + \eta_{x,j})/2 \right)^3 \right| + 38 \sum_{i=1}^{m} \sigma_{x,i} \eta_{x,i}^4 \\
&\leq \frac{14}{3} \frac{r^3}{(mn)^{3/4}} \left| \sum_{i=1}^{m} (\sigma_{x,i} + \beta) \left( \hat{a}_i^\top \xi \right)^3 \right| + \frac{8}{3} \frac{r^3}{(mn)^{3/4}} \left| \sum_{i,j=1}^{m} \sigma_{x,i,j}^2 \left( \frac{1}{2} (\hat{a}_i + \hat{a}_j)^\top \xi \right)^3 \right| \\
&\quad + 50 \frac{r^4}{mn} \sum_{i=1}^{m} (\sigma_{x,i} + \beta) (\hat{a}_i^\top \xi)^4, \qquad (46)
\end{aligned}$$

where the last step follows from the fact that $0 \leq \sigma_{x,i} \leq \sigma_{x,i} + \beta$. In order to show that $\left| \|z-x\|_z^2 - \|z-x\|_x^2 \right|$ is $\mathcal{O}(1/\sqrt{mn})$ with high probability, it suffices to show that with high probability, the third and fourth degree polynomials of $\hat{a}_i^\top \xi$, that appear in bound (46), are bounded by $\mathcal{O}\left((mn)^{1/4}\right)$ and $\mathcal{O}\left(\sqrt{mn}\right)$ respectively. Note that for any $(\epsilon, r)$ such that $\epsilon \in (0, \frac{1}{4}]$ and $r \leq f(\epsilon)$, the pair $(\epsilon, r)$ satisfies the assumptions for bounds (34b), (34c) and (34d) in Lemma 12 and hence we have

$$\|z-x\|_z^2 - \|z-x\|_x^2 \leq 2\epsilon \frac{r^2}{\sqrt{mn}},$$

with probability at least $1-\epsilon$.

### B.3 Proof of Lemma 12

The proof relies on the classical (although surprising) property of Gaussian polynomials: exponential tail decay independent of dimension. In the following lemma, we state the tail bounds for arbitrary Gaussian polynomials.

**Lemma 13.** *(Thm 6.7, Janson [Jan97]) For any $n \geq 1$, let $P : \mathbb{R}^n \to \mathbb{R}$ be a polynomial of degree $k$, and let $\xi \sim \mathcal{N}(0, \mathbb{I}_n)$. Then for any $t \geq (2e)^{k/2}$, we have*

$$\mathbb{P}\left[|P(\xi)| \geq t\left(\mathbb{E}P(\xi)^2\right)^{\frac{1}{2}}\right] \leq \exp\left(-\frac{k}{2e}t^{2/k}\right).$$

We outline the key steps of the proof and refer the readers to the book [Jan97] for details. The proof proceeds in two steps. First, it is established that any two Gaussian norms of an arbitrary polynomial of degree $k$ can be related as

$$[\mathbb{E}\,|P(\xi)|^p]^{1/p} \leq c(p,q)^k\,[\mathbb{E}\,|P(\xi)|^q]^{1/q}\,, \quad \text{for } \xi \sim \mathcal{N}(0, \mathbb{I}_n) \text{ and any } p, q > 0, \tag{47}$$

and the term $c(p,q)$ does not depend on $n$. Further Janson shows that $c(2,q) = (q-1)^{1/2}$ for $q \geq 2$. The proof then follows from Chebyshev's inqeuality and the relation (47) for $p = 2$ and a suitably chosen $q$.

Also, the following observations for $\hat{a}_i$ (33) will come in handy for the proofs that follow:

$$\|\hat{a}_i\|_2^2 = \theta_{x,i} \leq \sqrt{\frac{m}{n}} \quad \text{for all } i \in [m], \quad \text{and} \tag{48a}$$

$$(\hat{a}_i^\top \hat{a}_j)^2 = \theta_{x,i,j}^2 \quad \text{for all } i, j \in [m]. \tag{48b}$$

### B.3.1 Proof of bound (34a)

We have

$$\begin{aligned}
\mathbb{E}\left(\sum_{i=1}^m (\sigma_{x,i} + \beta)\left(\hat{a}_i^\top \xi\right)^2\right)^2 &= \sum_{i,j=1}^m (\sigma_{x,i} + \beta)(\sigma_{x,j} + \beta)\,\mathbb{E}\left(\hat{a}_i^\top \xi\right)^2\left(\hat{a}_j^\top \xi\right)^2 \\
&= \sum_{i,j=1}^m (\sigma_{x,i} + \beta)(\sigma_{x,j} + \beta)\left(\|\hat{a}_i\|_2^2\,\|\hat{a}_j\|_2^2 + 2\left(\hat{a}_i^\top \hat{a}_j\right)^2\right) \\
&= \sum_{i,j=1}^m (\sigma_{x,i} + \beta)(\sigma_{x,j} + \beta)\left(\theta_{x,i}\theta_{x,j} + 2\theta_{x,i,j}^2\right) \\
&\overset{(i)}{=} n^2 + 2n \\
&\leq 3n^2,
\end{aligned}$$

where step $(i)$ follows from properties (c) and (d) from Lemma 8. Applying Lemma 13 with $k = 2, t = \gamma_2$ and $r$ such that $\sqrt{3}\gamma_2 \leq \epsilon/90r^2$ the claim follows.

### B.3.2 Proof of bound (34b)

Using Isserlis' theorem [Iss18] for Gaussian moments, we obtain

$$\begin{aligned}\mathbb{E}\left(\sum_{i=1}^{m}\left(\sigma_{x,i}+\beta\right)\left(\hat{a}_i^\top\xi\right)^3\right)^2 &= \sum_{i,j=1}^{m}\left(\sigma_{x,i}+\beta\right)\left(\sigma_{x,i}+\beta\right)\mathbb{E}\left(\hat{a}_i^\top\xi\right)^3\left(\hat{a}_j^\top\xi\right)^3\\ &= 9\underbrace{\sum_{i,j=1}^{m}\left(\sigma_{x,i}+\beta\right)\left(\sigma_{x,j}+\beta\right)\|\hat{a}_i\|_2^2\,\|\hat{a}_j\|_2^2\left(\hat{a}_i^\top\hat{a}_j\right)}_{=:N_1}\\ &\quad+6\underbrace{\sum_{i,j=1}^{m}\left(\sigma_{x,i}+\beta\right)\left(\sigma_{x,j}+\beta\right)\left(\hat{a}_i^\top\hat{a}_j\right)^3}_{=:N_2}.\end{aligned}\tag{49}$$

We claim that $N_1 \leq \sqrt{mn}$ and $N_2 \leq \sqrt{mn}$. Assuming the claims as given, we now complete the proof. Plugging in the bounds for $N_1$ and $N_2$ in equation (49) we find that $\mathbb{E}\big(\sum_{i=1}^{m}\left(\sigma_{x,i}+\beta\right)\left(\hat{a}_i^\top\xi\right)^3\big)^2 \leq 15\sqrt{mn}$. Applying Lemma 13 with $k=3, t=\gamma_2$ and $r$ such that $\sqrt{15}\gamma_3 \leq \epsilon/7r$ yields the claim. We now turn to proving the bounds on $N_1$ and $N_2$.

**Bounding $N_1$:** Let $B$ be an $n\times n$ matrix with its $i$-th row given by $\sqrt{(\sigma_{x,i}+\beta)}\hat{a}_i^\top$. Observe that

$$\sum_{i=1}^{m}\left(\sigma_{x,i}+\beta\right)\hat{a}_i\hat{a}_i^\top = V_x^{-1/2}\left(\sum_{i=1}^{m}\left(\sigma_{x,i}+\beta\right)\frac{a_ia_i^\top}{s_{x,i}^2}\right)V_x^{-1/2} = V_x^{-1/2}V_xV_x^{-1/2} = \mathbb{I}_n.\tag{50}$$

Thus we have $B^\top B = \mathbb{I}_n$, which implies that $BB^\top$ is an orthogonal projection matrix. Let $v\in\mathbb{R}^m$ be a vector such that $v_i = \sqrt{(\sigma_{x,i}+\beta)}\,\|\hat{a}_i\|_2^2$.

$$\sum_{i,j=1}^{m}(\sigma_{x,i}+\beta)\|\hat{a}_i\|_2^2\,\hat{a}_i^\top\,(\sigma_{x,j}+\beta)\|\hat{a}_j\|_2^2\,\hat{a}_j = \left\|\sum_{i=1}^{m}(\sigma_{x,i}+\beta)\|\hat{a}_i\|_2^2\,\hat{a}_i\right\|_2^2 = \left\|B^\top v\right\|_2^2 \overset{(i)}{\leq} \|v\|_2^2\,,$$

where inequality $(i)$ follows from the fact that $v^\top Pv \leq \|v\|_2^2$ for any orthogonal projection matrix $P$. Equation (48a) implies that $v_i^2 = \left(\sigma_{x,i}+\beta\right)\theta_{x,i}^2$. Using Lemma 8(e), we find that

$$\|v\|_2^2 = \sum_{i=1}^{m}\left(\sigma_{x,i}+\beta\right)\theta_{x,i}^2 \leq \sqrt{mn}.$$

**Bounding $N_2$:** We see that

$$\begin{aligned}\sum_{i,j=1}^{m}\left(\sigma_{x,i}+\beta\right)\left(\sigma_{x,j}+\beta\right)\left(\hat{a}_i^\top\hat{a}_j\right)^3 &\overset{\text{(C−S)}}{\leq} \sum_{i,j=1}^{m}\left(\sigma_{x,i}+\beta\right)\left(\sigma_{x,j}+\beta\right)\left(\hat{a}_i^\top\hat{a}_j\right)^2\|\hat{a}_i\|_2\,\|\hat{a}_j\|_2\\ &\overset{\text{(eqns.(48a),(48b))}}{\leq} \sum_{i,j=1}^{m}\left(\sigma_{x,i}+\beta\right)\left(\sigma_{x,j}+\beta\right)\theta_{x,i,j}^2\sqrt{\theta_{x,i}\theta_{x,j}}\\ &\overset{\text{(Lem. 2(b))}}{\leq} \sqrt{\frac{m}{n}}\sum_{i,j=1}^{m}\left(\sigma_{x,i}+\beta\right)\left(\sigma_{x,j}+\beta\right)\theta_{x,i,j}^2.\end{aligned}$$

We now apply Lemma 8(d) followed by Lemma 8(c) to obtain the claimed bound on $N_2$.

### B.3.3 Proof of bound (34c)

Let $c_{i,j} = \dfrac{(\hat{a}_i + \hat{a}_j)}{2}$ for $i, j \in [m]$. Using Isserlis' theorem for Gaussian moments, we obtain

$$\begin{aligned}\mathbb{E}\left(\sum_{i,j=1}^{m} \sigma_{x,i,j}^2 \left(c_{i,j}^\top \xi\right)^3\right)^2 &= \sum_{i,j,k,l=1}^{m} \sigma_{x,i,j}^2 \sigma_{x,k,l}^2 \mathbb{E}\left(c_{i,j}^\top \xi\right)^3 \left(c_{k,l}^\top \xi\right)^3 \\ &= 9 \underbrace{\sum_{i,j,k,l=1}^{m} \sigma_{x,i,j}^2 \sigma_{x,k,l}^2 \|c_{i,j}\|_2^2 \, \|c_{k,l}\|_2^2 \left(c_{i,j}^\top c_{k,l}\right)}_{C_1:=} + 6 \underbrace{\sum_{i,j,k,l=1}^{m} \sigma_{x,i,j}^2 \sigma_{x,k,l}^2 \left(c_{i,j}^\top c_{k,l}\right)^3}_{C_2:=}\end{aligned}$$

We claim that $C_1 \leq \sqrt{mn}$ and $C_2 \leq \sqrt{mn}$. Assuming the claims as given, the result follows using similar arguments as in the previous part. We now bound $C_i, i = 1, 2$, using arguments similar to the ones used in Section B.3.2 to bound $N_i, i = 1, 2$, respectively. The following bounds on $\|c_{i,j}\|_2^2$ are used in the arguments that follow:

$$\|c_{i,j}\|_2^2 \overset{\text{SSI}}{\leq} \frac{1}{2}\left(\|\hat{a}_i\|_2^2 + \|\hat{a}_j\|_2^2\right) = \frac{1}{2}\left(\theta_{x,i} + \theta_{x,j}\right) \tag{51a}$$

$$\overset{\text{Lem. } 2(b)}{\leq} \sqrt{\frac{m}{n}}. \tag{51b}$$

**Bounding $C_1$:** Let $B$ be the same $m \times n$ matrix as in the proof of previous part with its $i$-th row given by $\sqrt{(\sigma_{x,i} + \beta)}\hat{a}_i^\top$. Define the vector $u \in \mathbb{R}^n$ with entries given by $u_i = \sum_{j=1}^m \sigma_{x,i,j}^2 \, \|c_{i,j}\|_2^2 / (\sigma_{x,i} + \beta)^{1/2}$. We have

$$\begin{aligned}\sum_{i,j,k,l=1}^{m} \sigma_{x,i,j}^2 \sigma_{x,k,l}^2 \, \|c_{i,j}\|_2^2 \, \|c_{k,l}\|_2^2 \left(c_{i,j}^\top c_{k,l}\right) &\leq \left\|\sum_{i,j=1}^{m} \sigma_{x,i,j}^2 \, \|c_{i,j}\|_2^2 \, c_{i,j}\right\|_2^2 \\ &\overset{\text{(SSI)}}{\leq} \frac{1}{2}\left(\left\|\sum_{i,j=1}^{m} \sigma_{x,i,j}^2 \, \|c_{i,j}\|_2^2 \, \hat{a}_i\right\|_2^2 + \left\|\sum_{i,j=1}^{m} \sigma_{x,i,j}^2 \, \|c_{i,j}\|_2^2 \, \hat{a}_j\right\|_2^2\right) \\ &= \left\|B^\top u\right\|_2^2 \\ &\overset{(i)}{\leq} \|u\|_2^2,\end{aligned}$$

where inequality $(i)$ follows from the fact that $v^\top P v \leq \|v\|_2^2$ for any orthogonal projection matrix $P$. It is left to bound the term $u_i^2$. We see that

$$\begin{aligned}u_i^2 = \frac{1}{\sigma_{x,i} + \beta} \sum_{j,k=1}^{m} \sigma_{x,i,j}^2 \sigma_{x,i,k}^2 \, \|c_{i,j}\|_2^2 \, \|c_{i,k}\|_2^2 &\overset{\text{(bnd. (51b))}}{\leq} \sqrt{\frac{m}{n}} \frac{1}{\sigma_{x,i} + \beta} \sum_{j,k=1}^{m} \sigma_{x,i,j}^2 \sigma_{x,i,k}^2 \, \|c_{i,j}\|_2^2 \\ &\overset{\text{(Lem. } 8(a))}{\leq} \sqrt{\frac{m}{n}} \frac{\sigma_{x,i}}{\sigma_{x,i} + \beta} \sum_{j=1}^{m} \sigma_{x,i,j}^2 \, \|c_{i,j}\|_2^2 \\ &\overset{\text{(bnd. (51a))}}{\leq} \sqrt{\frac{m}{n}} \sum_{j=1}^{m} \sigma_{x,i,j}^2 \frac{\theta_{x,i} + \theta_{x,j}}{2}.\end{aligned}$$

Now, summing over $i$ and using symmetry of indices $i, j$, we find that

$$\|u\|_2^2 \leq \sqrt{\frac{m}{n}} \sum_{i=1}^{m} \sum_{j=1}^{m} \sigma_{x,i,j}^2 \theta_{x,i} \overset{\text{(Lem. 8(a))}}{=} \sqrt{\frac{m}{n}} \sum_{i=1}^{m} \sigma_{x,i} \theta_{x,i} \overset{\text{(Lem. 8(c))}}{\leq} \sqrt{mn},$$

thereby implying that $C_1 \leq \sqrt{mn}$.

**Bounding $C_2$:** Using Cauchy-Schwarz inequality and bound (51b), we find that

$$\sum_{i,j,k,l=1}^{m} \sigma_{x,i,j}^2 \sigma_{x,k,l}^2 \left(c_{i,j}^\top c_{k,l}\right)^3 \leq \sum_{i,j,k,l=1}^{m} \sigma_{x,i,j}^2 \sigma_{x,k,l}^2 \left(c_{i,j}^\top c_{k,l}\right)^2 \|c_{i,j}\|_2 \|c_{k,l}\|_2 \leq \sqrt{\frac{m}{n}} \sum_{i,j,k,l=1}^{m} \sigma_{x,i,j}^2 \sigma_{x,k,l}^2 \left(c_{i,j}^\top c_{k,l}\right)^2.$$

Using SSI and the symmetry of pairs of indices $(i, j)$ and $(k, l)$, we obtain

$$\sum_{i,j,k,l=1}^{m} \sigma_{x,i,j}^2 \sigma_{x,k,l}^2 \left(c_{i,j}^\top c_{k,l}\right)^2 \leq \sum_{i,j,k,l=1}^{m} \sigma_{x,i,j}^2 \sigma_{x,k,l}^2 \left(\hat{a}_i^\top \hat{a}_k\right)^2 = \sum_{i,k=1}^{m} \sigma_{x,i} \sigma_{x,k} \left(\hat{a}_i^\top \hat{a}_k\right)^2.$$

The resulting expression can be bounded as follows:

$$\sum_{i,k=1}^{m} \sigma_{x,i} \sigma_{x,k} \left(\hat{a}_i^\top \hat{a}_k\right)^2 \overset{\text{(eqn.(48b))}}{=} \sum_{i,k=1}^{m} \sigma_{x,i} \sigma_{x,k} \theta_{x,i,k}^2 \overset{\text{(Lem. 8(d))}}{\leq} \sum_{i=1}^{m} \sigma_{x,i} \theta_{x,i} \overset{\text{(Lem. 8(c))}}{\leq} n.$$

Putting the pieces together yields the claimed bound on $C_2$.

### B.3.4 Proof of bound (34d)

Observe that $\hat{a}_i^\top \xi \sim \mathcal{N}(0, \theta_{x,i})$ and hence $\mathbb{E}\left(\hat{a}_i^\top \xi\right)^8 = 105\, \theta_{x,i}^4$. Thus we have

$$\begin{aligned}
\mathbb{E}\left(\sum_{i=1}^{m} \sigma_{x,i} \left(\hat{a}_i^\top \xi\right)^4\right)^2 &\overset{\text{C−S}}{\leq} \sum_{i,j=1}^{m} \sigma_{x,i} \sigma_{x,j} \left(\mathbb{E}\left(\hat{a}_i^\top \xi\right)^8\right)^{\frac{1}{2}} \left(\mathbb{E}\left(\hat{a}_j^\top \xi\right)^8\right)^{\frac{1}{2}} \\
&= 105 \sum_{i,j=1}^{m} \sigma_{x,i} \sigma_{x,j} \theta_{x,i}^2 \theta_{x,j}^2 \\
&= 105 \left(\sum_{i=1}^{m} \sigma_{x,i} \theta_{x,i}^2\right)^2 \\
&\overset{\text{(Lem. 8(e))}}{\leq} 105 mn.
\end{aligned}$$

Applying Lemma 13 with $k = 4, t = \gamma_4$ and any positive $r$ such that $\sqrt{105}\gamma_4 \leq \epsilon/100r^2$ yields the result.

## B.4 Proof of Lemma 10

We now derive the different expressions for derivatives and prove the bounds for Hessians of $x \mapsto \varphi_{x,i}$, $i \in [m]$ and $x \mapsto \Psi_x$. In this section we use the simpler notation $H_x := \nabla^2 F_x$.

### B.4.1 Gradient of $\sigma$

Using $s_{x+h,i} = (b_i - a_i^\top (x+h)) = s_{x,i} - a_i^\top h$, we define,

$$\Delta_{x,h}^H := H_{x+h} - H_x = \sum_{i=1}^m a_i a_i^\top \left( \frac{1}{(s_{x,i} - a_i^\top h)^2} - \frac{1}{s_{x,i}^2} \right). \tag{52}$$

We have (upto second order terms)

$$\frac{1}{s_{x+h,i}^2} = \frac{1}{s_{x,i}^2} \left[ 1 + \frac{2a_i^\top h}{s_{x,i}} + \frac{3(a_i^\top h)^2}{s_{x,i}^2} \right], \tag{53a}$$

$$\Delta_{x,h}^H = \sum_{i=1}^m \frac{a_i a_i^\top}{s_{x,i}^2} \left[ \frac{2a_i^\top h}{s_{x,i}} + \frac{3(a_i^\top h)^2}{s_{x,i}^2} \right], \tag{53b}$$

$$a_i^T H_{x+h}^{-1} a_i = a_i^\top H_x^{-1} a_i - a_i^\top H_x^{-1} \Delta_{x,h}^H H_x^{-1} a_i + a_i^\top H_x^{-1} \Delta_{x,h}^H H_x^{-1} \Delta_{x,h}^H H_x^{-1} a_i. \tag{53c}$$

Collecting different first order terms in $\sigma_{x+h,i} - \sigma_{x,i}$ we have

$$\begin{aligned}
\sigma_{x+h,i} - \sigma_{x,i} &= 2\, \frac{a_i^\top H_x^{-1} a_i}{s_{x,i}^2} \frac{a_i^\top h}{s_{x,i}} - 2\, \frac{a_i^\top H_x^{-1} \left( \sum_{j=1}^m \frac{a_j a_j^\top}{s_{x,j}^2} \frac{a_j^\top h}{s_{x,j}} \right) H_x^{-1} a_i}{s_{x,i}^2} \\
&= 2 \left[ \sigma_{x,i} \frac{a_i^\top h}{s_{x,i}} - \sum_{j=1}^m \sigma_{x,i,j}^2 \frac{a_j^\top h}{s_{x,j}} \right] \\
&= 2\, [(\Sigma_x - \Upsilon_x^{(2)}) S_x^{-1} A]_i\, h.
\end{aligned}$$

Letting $h \to 0$, we have the desired result.

### B.4.2 Gradient of $\varphi$

Using the chain rule and the fact that $\nabla s_{x,i} = -a_i$, we find that

$$\begin{aligned}
\nabla \varphi_{x,i} &= \frac{\nabla \sigma_{x,i}}{s_{x,i}^2} - 2\,(\sigma_{x,i} + \beta) \frac{\nabla s_{x,i}}{s_{x,i}^3} \\
&= \frac{2}{s_{x,i}^2} A^\top S_x^{-1} \left[ 2\Sigma_x + \beta\, \mathbb{I} - \Upsilon_x^{(2)} \right] e_i,
\end{aligned}$$

as claimed.

### B.4.3 Gradient of $\Psi$

We recall equations (33) and (50)

$$\hat{a}_i = \frac{1}{s_{x,i}} V_x^{-1/2} a_i, \quad \text{and} \quad \sum_{i=1}^m (\sigma_{x,i} + \beta)\, \hat{a}_i \hat{a}_i^\top = \mathbb{I}_n.$$

For a unit vector $h$, we have

$$h^\top \nabla \log\det V_x = \lim_{\delta \to 0} \frac{1}{\delta} \left[ \operatorname{trace}\log \left( \sum_{i=1}^m \frac{(\sigma_{x+\delta h,i} + \beta)}{\left(1 - \delta a_i^\top h / s_{x,i}\right)^2} \hat{a}_i \hat{a}_i^\top \right) - \operatorname{trace}\log \left( \sum_{i=1}^m (\sigma_{x,i} + \beta)\, \hat{a}_i \hat{a}_i^\top \right) \right]. \tag{54}$$

Let $\log L$ denote the logarithm of the matrix $L$. Keeping track of the first order terms on RHS of equation (54), we find that

$$
\begin{aligned}
&\operatorname{trace}\left[\log\left(\sum_{i=1}^{m}\left(\sigma_{x+\delta h,i}+\beta\right)\frac{\hat{a}_i\hat{a}_i^\top}{\left(1-\delta a_i^\top h/s_{x,i}\right)^2}\right)\right]-\operatorname{trace}\left[\log\left(\sum_{i=1}^{m}\left(\sigma_{x,i}+\beta\right)\hat{a}_i\hat{a}_i^\top\right)\right]\\
&=\operatorname{trace}\left[\log\left(\sum_{i=1}^{m}\left(\sigma_{x+\delta h,i}+\beta+\delta h^\top\nabla\sigma_{x,i}\right)\left(1+2\delta\frac{a_i^\top h}{s_{x,i}^2}\right)\right)\right]-\operatorname{trace}\left[\log\left(\sum_{i=1}^{m}\left(\sigma_{x,i}+\beta\right)\hat{a}_i\hat{a}_i^\top\right)\right]\\
&=\operatorname{trace}\left[\sum_{i=1}^{m}\delta\left(2\left(\sigma_{x,i}+\beta\right)\frac{a_i^\top h}{s_{x,i}^2}+h^\top\nabla\sigma_{x,i}\right)\hat{a}_i\hat{a}_i^\top\right]\\
&=\delta\left(\sum_{i=1}^{m}\left(2\left(\sigma_{x,i}+\beta\right)\frac{a_i^\top h}{s_{x,i}^2}+h^\top\nabla\sigma_{x,i}\right)\theta_i\right),
\end{aligned}
$$

where we have used the fact $\operatorname{trace}(\log\mathbb{I})=0$. Substituting expression of $h^\top\nabla\sigma_x$ from part (a), we obtain

$$
h^\top\nabla\log\det V_x = A_x^\top\left(4\Sigma_x+2\beta\mathbb{I}-2\Upsilon_x^{(2)}\right)\Theta_x h.
$$

### B.4.4 Bound on Hessian $\nabla^2\varphi$

Let $E_{ii}=e_ie_i^\top$. We claim that for any $h\in\mathbb{R}^n$, we have

$$
\begin{aligned}
h^\top\nabla^2\varphi_{x,i}h=\frac{2}{s_{x,i}^2}h^\top A_x^\top\Big[&E_{ii}\left(3\left(\Sigma_x+\beta\mathbb{I}\right)+7\Sigma_x-8\operatorname{diag}(\Upsilon_x^{(2)}e_i)\right)E_{ii}\\
&+\operatorname{diag}(\Upsilon_xe_i)(4\Upsilon_x-3\mathbb{I})\operatorname{diag}(\Upsilon_xe_i)\Big]A_xh. \qquad (55)
\end{aligned}
$$

Note that

$$
\varphi_{x+h,i}-\varphi_{x,i}=\underbrace{\left(\frac{a_i^\top H_{x+h,i}^{-1}a_i}{s_{x+h,i}^4}-\frac{a_i^\top H_{x,i}^{-1}a_i}{s_{x,i}^4}\right)}_{=:A_1}+\underbrace{\beta\left(\frac{1}{s_{x+h,i}^2}-\frac{1}{s_{x,i}^2}\right)}_{=:A_2}. \qquad (56)
$$

The second order Taylor expansion of $1/s_{x,i}^4$ is given by

$$
\frac{1}{s_{x+h,i}^4}=\frac{1}{s_{x,i}^4}\left[1+\frac{4a_i^\top h}{s_{x,i}}+\frac{10(a_i^\top h)^2}{s_{x,i}^2}\right].
$$

Let $B_1$ and $B_2$ denote the second order terms, i.e., the terms that are of order $O(\|h\|_2^2)$, in Taylor expansion of $A_1$ and $A_2$ around $x$, respectively. Borrowing terms from equations (53a)-(53c) and simplifying we obtain

$$
B_1=10\sigma_{x,i}\frac{(a_i^\top h)^2}{s_{x,i}^2}-8\frac{a_i^\top h}{s_{x,i}}\sum_{j=1}^{m}\frac{\sigma_{x,i,j}^2}{s_{x,i}^2}\frac{a_j^\top h}{s_{x,j}}-3\sum_{j=1}^{m}\frac{\sigma_{x,i,j}^2}{s_{x,i}^2}\frac{(a_j^\top h)^2}{s_{x,j}^2}+4\sum_{j=1}^{m}\sum_{l=1}^{m}\frac{\sigma_{x,i,j}}{s_{x,i}}\sigma_{x,j,l}\frac{\sigma_{x,l,i}}{s_{x,i}}\frac{a_j^\top h}{s_{x,j}}\frac{a_l^\top h}{s_{x,l}},
$$

and $B_2=3\beta\dfrac{(a_i^\top h)^2}{s_{x,i}^2}$.

Observing that the second order term in the Taylor expansion of $\varphi_{x+h,i}$ around $x$, is exactly $\frac{1}{2}h^\top \nabla^2 \varphi_{x,i} h$ yields the claim (55). We now turn to prove the bound on the directional Hessian. Recall $\eta_{x,i} = a_i^\top h / s_{x,i}$. We have

$$
\begin{aligned}
&s_{y,i}^2 \left| \frac{1}{2} h^\top \nabla^2 \varphi_{x,i} h \right| \\
&\quad = \left| 3\left(\sigma_{x,i}+\beta\right)\eta_{x,i}^2 + 7\sigma_{x,i}\eta_{x,i}^2 - 8\sum_{j=1}^m \sigma_{x,i,j}^2 \eta_{x,j}\eta_{x,i} - 3\sum_{j=1}^m \sigma_{x,i,j}^2 \eta_{x,j}^2 + 4\sum_{j,k=1}^m \sigma_{x,i,j}\sigma_{x,j,k}\sigma_{x,k,i}\eta_{x,j}\eta_{x,k} \right| \\
&\quad \overset{(i)}{\leq} 10\left(\sigma_{x,i}+\beta\right)\eta_{x,i}^2 + 8\sum_{j=1}^m \sigma_{x,i,j}^2 \left|\eta_{x,i}\eta_{x,j}\right| + 7\sum_{j=1}^m \sigma_{x,i,j}^2 \eta_{x,j}^2 \\
&\quad \overset{(ii)}{\leq} 10\left(\sigma_{x,i}+\beta\right)\eta_{x,i}^2 + 4\sum_{j=1}^m \sigma_{x,i,j}^2 \left(\eta_{x,i}^2 + \eta_{x,j}^2\right) + 7\sum_{j=1}^m \sigma_{x,i,j}^2 \eta_{x,j}^2 \\
&\quad \overset{(iii)}{\leq} 10\left(\sigma_{x,i}+\beta\right)\eta_{x,i}^2 + 4\sum_{j=1}^m \sigma_{x,i}\eta_{x,i}^2 + 4\sum_{j=1}^m \sigma_{x,i,j}^2 \eta_{x,j}^2 + 7\sum_{j=1}^m \sigma_{x,i,j}^2 \eta_{x,j}^2, \\
&\quad \overset{(iv)}{\leq} 14\left(\sigma_{x,i}+\beta\right)\eta_{x,i}^2 + 11\sum_{j=1}^m \sigma_{x,i,j}^2 \eta_{x,j}^2,
\end{aligned}
$$

where step $(i)$ follows from the fact that $\operatorname{diag}(\Upsilon_y e_i)\Upsilon_y \operatorname{diag}(\Upsilon_y e_i) \preceq \operatorname{diag}(\Upsilon_y e_i)\operatorname{diag}(\Upsilon_y e_i)$ since $\Upsilon_y$ is an orthogonal projection matrix; step $(ii)$ follows from AM-GM inequality; step $(iii)$ follows from the symmetry of indices $i$ and $j$ and Lemma 8(a), and step $(iv)$ from the fact that $\sigma_{x,i} \leq \sigma_{x,i} + \beta$.

### B.4.5 Bound on Hessian $\nabla^2 \Psi$

We have

$$
\begin{aligned}
\frac{1}{2}h^\top \nabla^2 \Psi_x h = &\frac{1}{2}\lim_{\delta\to 0}\frac{1}{\delta^2}\Bigg[\operatorname{trace}\log\left(\sum_{i=1}^m \frac{\left(\sigma_{x+\delta h,i}+\beta\right)}{\left(1-\delta a_i^\top h/s_{x,i}\right)^2}\hat{a}_i\hat{a}_i^\top\right) + \operatorname{trace}\log\left(\sum_{i=1}^m \frac{\left(\sigma_{x-\delta h,i}+\beta\right)}{\left(1+\delta a_i^\top h/s_{x,i}\right)^2}\hat{a}_i\hat{a}_i^\top\right) \\
&-2\operatorname{trace}\log\left(\sum_{i=1}^m \left(\sigma_x+\beta\right)\hat{a}_i\hat{a}_i^\top\right)\Bigg]. \qquad (57)
\end{aligned}
$$

Upto second order terms, we have

$$
\begin{aligned}
&\operatorname{trace}\left[\log\left(\sum_{i=1}^m \left(\sigma_{x+\delta h,i}+\beta\right)\frac{\hat{a}_i\hat{a}_i^\top}{\left(1-\delta a_i^\top h/s_{x,i}\right)^2}\right)\right] \\
&= \operatorname{trace}\left[\log\left(\sum_{i=1}^m \left(\sigma_{x,i}+\beta+\delta h^\top\nabla\sigma_{x,i}+\frac{1}{2}\delta^2 h^\top\nabla^2\sigma_{x,i}h\right)\left(1+2\delta\frac{a_i^\top h}{s_{x,i}}+3\delta^2\left(\frac{a_i^\top h}{s_{x,i}}\right)^2\right)\hat{a}_i\hat{a}_i^\top\right)\right] \\
&= \operatorname{trace}\left[\sum_{i=1}^m \left(\sigma_{x,i}+\beta+\delta h^\top\nabla\sigma_{x,i}+\frac{1}{2}\delta^2 h^\top\nabla^2\sigma_{x,i}h\right)\left(1+2\delta\frac{a_i^\top h}{s_{x,i}}+3\delta^2\left(\frac{a_i^\top h}{s_{x,i}}\right)^2\right)\hat{a}_i\hat{a}_i^\top\right] \\
&\quad - \operatorname{trace}\left[\frac{1}{2}\left(\sum_{i=1}^m \left(\sigma_{x,i}+\beta+\delta h^\top\nabla\sigma_{x,i}+\frac{1}{2}\delta^2 h^\top\nabla^2\sigma_{x,i}h\right)\left(1+2\delta\frac{a_i^\top h}{s_{x,i}}+3\delta^2\left(\frac{a_i^\top h}{s_{x,i}}\right)^2\right)\hat{a}_i\hat{a}_i^\top\right)^2\right].
\end{aligned}
$$

We can similarly obtain the second order expansion of the term trace log $\left(\sum_{i=1}^m \frac{(\sigma_{x-\delta h,i}+\beta)}{(1+\delta a_i^\top h/s_{x,i})^2}\hat{a}_i\hat{a}_i^\top\right)$. Recall $\eta_{x,i} = \frac{a_i^\top h}{s_{x,i}}$. Using part (a) to substitute $h^\top \nabla \sigma_{x,i}$, we obtain

$$\begin{aligned}
\frac{1}{2}h^\top \nabla^2 \Psi_x h = &\sum_{i=1}^m \left(3\left(\sigma_{x,i}+\beta\right)\eta_{x,i}^2 + 4\left(\sigma_{x,i}\eta_{x,i}^2 - \sum_{j=1}^m \sigma_{x,i,j}^2 \eta_{x,i}\eta_{x,j}\right) + \frac{1}{2}h^\top \nabla^2 \sigma_{x,i} h\right)\theta_i \\
&- 2\Bigg[\sum_{i,j=1}^m \left(2\sigma_{x,i}+\beta\right)\left(2\sigma_{x,j}+\beta\right)\eta_{x,i}\eta_{x,j}\theta_{x,i,j}^2 - 2\sum_{i,j,k=1}^m \left(2\sigma_{x,i}+\beta\right)\sigma_{x,j,k}^2\theta_{x,i,k}^2\eta_{x,i}\eta_{x,j} \\
&+ \sum_{i,j,k,l=1}^m \sigma_{x,i,l}^2\sigma_{x,j,k}^2\theta_{x,k,l}^2\eta_{x,i}\eta_{x,j}\Bigg]. \qquad (58)
\end{aligned}$$

We claim that the directional Hessian $h^\top \nabla^2 \sigma_{x,i} h$ is given by

$$h^\top \nabla^2 \sigma_{x,i} h = 2\,h^\top A_x^\top \left[E_{ii}(3\Sigma_x - 4\,\mathrm{diag}(\Upsilon_x^{(2)}e_i))E_{ii} + \mathrm{diag}(\Upsilon_x e_i)(4\Upsilon_x - 3\mathbb{I})\,\mathrm{diag}(\Upsilon_x e_i)\right] A_x h. \qquad (59)$$

Assuming the claim at the moment we now bound $\left|h^\top \nabla^2 \Psi_x h\right|$. To shorten the notation, we drop the $x$-dependence of the terms $\sigma_{x,i}, \sigma_{x,i,j}, \theta_{x,i}$ and $\eta_{x,i}$. Since $\Upsilon_x$ is an orthogonal projection matrix, we have

$$\mathrm{diag}(\Upsilon_x e_i)\Upsilon_x\,\mathrm{diag}(\Upsilon_x e_i) \preceq \mathrm{diag}(\Upsilon_x e_i)\,\mathrm{diag}(\Upsilon_x e_i).$$

Using this fact and substituting the expression for $h^\top \nabla^2 \sigma_{x,i} h$ from equation (59) in equation (55), we obtain

$$\begin{aligned}
&\left|h^\top \nabla^2 \Psi_x h\right| \\
&\leq \sum_{i=1}^m \left[3\Big(\sigma_i + \beta\Big)\eta_i^2 + 4(\sigma_i\eta_i^2 + \sum_{j=1}^m \sigma_{i,j}^2\eta_i\eta_j) + 3\sigma_i\eta_i^2 + 4\sum_{j=1}^m \sigma_{i,j}^2\eta_i\eta_j + 7\sum_{j=1}^m \sigma_{i,j}^2\eta_j^2\right]\theta_i \\
&\quad + \left[8\sum_{i,j=1}^m \left(\sigma_i+\beta\right)\left(\sigma_j+\beta\right)\eta_i\eta_j\theta_{i,j}^2 + 8\sum_{i,j,k=1}^m \left(\sigma_i+\beta\right)\sigma_{j,k}^2\theta_{i,k}^2\eta_i\eta_j + 2\sum_{i,j,k,l=1}^m \sigma_{i,l}^2\sigma_{j,k}^2\theta_{k,l}^2\eta_i\eta_j\right].
\end{aligned}$$

Rearranging terms, we find that

$$\begin{aligned}
&\left|h^\top \nabla^2 \Psi_x h\right| \\
&\leq \sum_{i=1}^m \left[10\left(\sigma_i+\beta\right)\eta_i^2 + 8\sum_{j=1}^m \sigma_{i,j}^2\eta_i\eta_j + 7\sum_{j=1}^m \sigma_{i,j}^2\eta_j^2\right]\theta_i \\
&\quad + \left[8\sum_{i,j=1}^m \left(\sigma_i+\beta\right)\left(\sigma_j+\beta\right)\eta_i\eta_j\theta_{i,j}^2 + 8\sum_{i,j,k=1}^m \left(\sigma_i+\beta\right)\sigma_{j,k}^2\theta_{i,k}^2\eta_i\eta_j + 2\sum_{i,j,k,l=1}^m \sigma_{i,l}^2\sigma_{j,k}^2\theta_{k,l}^2\eta_i\eta_j\right] \\
&\overset{(i)}{\leq} \sum_{i=1}^m \left[10\left(\sigma_i+\beta\right)\eta_i^2 + 4\sum_{j=1}^m \sigma_{i,j}^2\left(\eta_i^2+\eta_j^2\right) + 7\sum_{j=1}^m \sigma_{i,j}^2\eta_j^2\right]\theta_i \\
&\quad + \left[4\sum_{i,j=1}^m (\sigma_i+\beta)\left(\sigma_j+\beta\right)\theta_{i,j}^2(\eta_i^2+\eta_j^2) + 4\sum_{i,j,k=1}^m \left(\sigma_i+\beta\right)\sigma_{j,k}^2\theta_{i,k}^2(\eta_i^2+\eta_j^2) + \sum_{i,j,k,l=1}^m \sigma_{i,l}^2\sigma_{j,k}^2\theta_{k,l}^2(\eta_i^2+\eta_j^2)\right]
\end{aligned}$$

where in step $(i)$ we have used the AM-GM inequality. Simplifying further, we obtain

$$\begin{aligned}\left|h^\top \nabla^2 \Psi_y h\right| &\leq \sum_{i=1}^m \left[14\left(\sigma_i + \beta\right)\eta_i^2 + 11\sum_{j=1}^m \sigma_{i,j}^2 \eta_j^2\right]\theta_i + \left[\sum_{i=1}^m 12\left(\sigma_i+\beta\right)\theta_i\eta_i^2 + \sum_{i,j=1}^m 6\sigma_{i,j}^2\theta_i\eta_j^2\right] \\ &= 26\sum_{i=1}^m \left(\sigma_i + \beta\right)\theta_i\eta_i^2 + 17\sum_{i,j=1}^m \sigma_{i,j}^2\theta_i\eta_j^2.\end{aligned}$$

Diving both sides by 2 completes the proof.

**Proof of claim** (59)**:** In order to compute the directional Hessian of $x \mapsto \sigma_{x,i}$, we need to track the second order terms in equations (53a)-(53c). Collecting the second order terms (denoted by $\sigma_h^{(2)}$) in the expansion of $\sigma_{x+h,i} - \sigma_{x,i}$, we obtain

$$\sigma_h^{(2)} = 3\,\frac{a_i^\top H_x^{-1} a_i}{s_{x,i}^2}\,\frac{(a_i^\top h)^2}{s_{x,i}^2} - 4\,\frac{a_i^\top H_x^{-1}\left(\sum_{j=1}^m \frac{a_j a_j^\top}{s_{x,j}^2}\frac{a_j^\top h}{s_{x,j}}\right)H_x^{-1}a_i}{s_{x,i}^2}\,\frac{a_i^\top h}{s_{x,i}} - 3\,\frac{a_i^\top H_x^{-1}\left(\sum_{j=1}^m \frac{a_j a_j^\top}{s_{x,j}^2}\frac{(a_j^\top h)^2}{s_{x,j}^2}\right)H_x^{-1}a_i}{s_{x,i}^2} \tag{60}$$

$$+\,4\,\frac{a_i^\top H_x^{-1}\left(\sum_{j=1}^m \frac{a_j a_j^\top}{s_{x,j}^2}\frac{a_j^\top h}{s_{x,j}}\right)H_x^{-1}\left(\sum_{l=1}^m \frac{a_l a_l^\top}{s_{x,l}^2}\frac{a_l^\top h}{s_{x,l}}\right)a_i}{s_{x,i}^2}. \tag{61}$$

We simply each term on the RHS one by one. Simplifying the first term, we obtain

$$3\,\frac{a_i^\top H_x^{-1} a_i}{s_{x,i}^2}\,\frac{(a_i^\top h)^2}{s_{x,i}^2} = 3\,\sigma_{x,i}\eta_{x,i}^2 = h^\top 3\,A_x^\top E_{ii}\Sigma_x E_{ii} A_x\,h.$$

For the second term, we have

$$\begin{aligned}4\,\frac{a_i^\top H_x^{-1}\left(\sum_{j=1}^m \frac{a_j a_j^\top}{s_{x,j}^2}\frac{a_j^\top h}{s_{x,j}}\right)H_x^{-1}a_i}{s_{x,i}^2}\,\frac{a_i^\top h}{s_{x,i}} &= 4\,\eta_{x,i}\sum_{j=1}^m \sigma_{x,i,j}^2\,\eta_{x,j} \\ &= 4\,h^\top A_x^\top E_{ii}\,\mathrm{diag}\left(\Upsilon_x^{(2)} e_i\right)E_{ii}A_x h.\end{aligned}$$

The third term can be simplified as follows:

$$\begin{aligned}3\,\frac{a_i^\top H_x^{-1}\left(\sum_{j=1}^m \frac{a_j a_j^\top}{s_{x,j}^2}\frac{(a_j^\top h)^2}{s_{x,j}^2}\right)H_x^{-1}a_i}{s_{x,i}^2} &= 3\sum_{j=1}^m \sigma_{x,i,j}^2\eta_{x,j}^2 \\ &= 3\,h^\top A_x^\top\,\mathrm{diag}\left(\Upsilon_x e_i\right)\mathrm{diag}\left(\Upsilon_x e_i\right)A_x h\end{aligned}$$

For the last term, we find that

$$\begin{aligned}4\,\frac{a_i^\top H_x^{-1}\left(\sum_{j=1}^m \frac{a_j a_j^\top}{s_{x,j}^2}\frac{a_j^\top h}{s_{x,j}}\right)H_x^{-1}\left(\sum_{l=1}^m \frac{a_l a_l^\top}{s_{x,l}^2}\frac{a_l^\top h}{s_{x,l}}\right)a_i}{s_{x,i}^2} &= 4\sum_{j,l=1}^m \sigma_{x,i,j}\,\sigma_{x,j,l}\,\sigma_{x,l,i}\,\eta_{x,j}\,\eta_{x,l} \\ &= 4\,h^\top A_x^\top\,\mathrm{diag}\left(\Upsilon_x e_i\right)\Upsilon_x\,\mathrm{diag}\left(\Upsilon_x e_i\right)A_x h.\end{aligned}$$

Putting together the pieces yields the expression (59).

# References

[ABW12] Sungjin Ahn, Anoop Korattikara Balan, and Max Welling. Bayesian posterior sampling via stochastic gradient Fisher scoring. In *ICML*, 2012.

[Ans00] Kurt M Anstreicher. The volumetric barrier for semidefinite programming. *Mathematics of Operations Research*, 25(3):365–380, 2000.

[BEL15] Sébastien Bubeck, Ronen Eldan, and Joseph Lehec. Sampling from a log-concave distribution with projected Langevin Monte Carlo. *arXiv preprint arXiv:1507.02564*, 2015.

[BF87] Imre Bárány and Zoltán Füredi. Computing the volume is difficult. *Discrete & Computational Geometry*, 2(4):319–326, 1987.

[BGJM11] Steve Brooks, Andrew Gelman, Galin L Jones, and Xiao-Li Meng. *Handbook of Markov Chain Monte Carlo*. Chapman and Hall/CRC, 2011.

[Bha13] Rajendra Bhatia. *Matrix Analysis*, volume 169. Springer Science & Business Media, 2013.

[Bus73] Peter J Bushell. Hilbert's metric and positive contraction mappings in a Banach space. *Archive for Rational Mechanics and Analysis*, 52(4):330–338, 1973.

[BV04] Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

[CV14] Ben Cousins and Santosh Vempala. A cubic algorithm for computing Gaussian volume. In *Proceedings of the Twenty-Fifth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 1215–1228. Society for Industrial and Applied Mathematics, 2014.

[Dal16] Arnak S Dalalyan. Theoretical guarantees for approximate sampling from smooth and log-concave densities. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 2016.

[DDH07] James Demmel, Ioana Dumitriu, and Olga Holtz. Fast linear algebra is stable. *Numerische Mathematik*, 108(1):59–91, 2007.

[DFK91] Martin Dyer, Alan Frieze, and Ravi Kannan. A random polynomial-time algorithm for approximating the volume of convex bodies. *Journal of the ACM (JACM)*, 38(1):1–17, 1991.

[DM16] Alain Durmus and Eric Moulines. Sampling from strongly log-concave distributions with the unadjusted Langevin algorithm. *arXiv preprint arXiv:1605.01559*, 2016.

[Ele86] György Elekes. A geometric inequality and the complexity of computing volume. *Discrete & Computational Geometry*, 1(1):289–292, 1986.

[Has70] W Keith Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1):97–109, 1970.

[HJ12] Roger A Horn and Charles R Johnson. *Matrix Analysis.* Cambridge University Press, 2012.

[HM13] Kuo-Ling Huang and Sanjay Mehrotra. An empirical evaluation of walk-and-round heuristics for mixed integer linear programs. *Computational Optimization and Applications*, 55(3):545–570, 2013.

[Iss18] Leon Isserlis. 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, 1918.

[Jan97] Svante Janson. *Gaussian Hilbert Spaces*, volume 129. Cambridge University Press, 1997.

[Kha93] Leonid Khachiyan. Complexity of polytope volume computation. In *New Trends in Discrete and Computational Geometry*, pages 91–101. Springer, 1993.

[KN12] Ravindran Kannan and Hariharan Narayanan. Random walks on polytopes and an affine interior point method for linear programming. *Mathematics of Operations Research*, 37(1):1–20, 2012.

[KV06] Adam Tauman Kalai and Santosh Vempala. Simulated annealing for convex optimization. *Mathematics of Operations Research*, 31(2):253–266, 2006.

[Law91] Jim Lawrence. Polytope volume computation. *Mathematics of Computation*, 57(195):259–271, 1991.

[Lov99] László Lovász. Hit-and-run mixes fast. *Mathematical Programming*, 86(3):443–461, 1999.

[LS90] László Lovász and Miklós Simonovits. The mixing rate of Markov chains, an isoperimetric inequality, and computing the volume. In *Proceedings of 31st Annual Symposium on Foundations of Computer Science, 1990*, pages 346–354. IEEE, 1990.

[LS92] László Lovász and Miklós Simonovits. On the randomized complexity of volume and diameter. In *33rd Annual Symposium on Foundations of Computer Science, 1992*, pages 482–492. IEEE, 1992.

[LS93] László Lovász and Miklós Simonovits. Random walks in a convex body and an improved volume algorithm. *Random Structures & Algorithms*, 4(4):359–412, 1993.

[LV03] László Lovász and Santosh Vempala. Hit-and-run is fast and fun. *Tehnical Report, Microsoft Research*, 2003.

[LV06a] László Lovász and Santosh Vempala. Fast algorithms for logconcave functions: Sampling, rounding, integration and optimization. In *47th Annual IEEE Symposium on Foundations of Computer Science, 2006*, pages 57–68. IEEE, 2006.

[LV06b] László Lovász and Santosh Vempala. Hit-and-run from a corner. *SIAM Journal on Computing*, 35(4):985–1005, 2006.

[LV06c] László Lovász and Santosh Vempala. Simulated annealing in convex bodies and an $O^*(n^4)$ volume algorithm. *Journal of Computer and System Sciences*, 72(2):392–417, 2006.

[LV07] László Lovász and Santosh Vempala. The geometry of logconcave functions and sampling algorithms. *Random Structures & Algorithms*, 30(3):307–358, 2007.

[LV16] Yin Tat Lee and Santosh S Vempala. Geodesic walks in polytopes. *arXiv preprint arXiv:1606.04696*, 2016.

[MMPS83] Paul Meakin, H Metiu, RG Petschek, and DJ Scalapino. The simulation of spinodal decomposition in two dimensions: a comparison of Monte Carlo and Langevin dynamics. *The Journal of Chemical Physics*, 79(4):1948–1954, 1983.

[MRR$^+$53] Nicholas Metropolis, Arianna W Rosenbluth, Marshall N Rosenbluth, Augusta H Teller, and Edward Teller. Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21(6):1087–1092, 1953.

[Nar16] Hariharan Narayanan. Randomized interior point methods for sampling and optimization. *The Annals of Applied Probability*, 26(1):597–641, 2016.

[NN94] Yurii Nesterov and Arkadii Nemirovskii. *Interior-point polynomial algorithms in convex programming.* SIAM, 1994.

[NR10] Hariharan Narayanan and Alexander Rakhlin. Random walk approach to regret minimization. In *Advances in Neural Information Processing Systems*, pages 1777–1785, 2010.

[PW15a] Mert Pilanci and Martin J Wainwright. Newton sketch: A linear-time optimization algorithm with linear-quadratic convergence. *arXiv preprint arXiv:1505.02250*, 2015.

[PW15b] Mert Pilanci and Martin J Wainwright. Randomized sketches of convex programs with sharp guarantees. *IEEE Transactions on Information Theory*, 61(9):5096–5115, 2015.

[Rob04] Christian P Robert. *Monte Carlo methods.* Wiley Online Library, 2004.

[RR01] Gareth O Roberts and Jeffrey S Rosenthal. Optimal scaling for various Metropolis-Hastings algorithms. *Statistical Science*, 16(4):351–367, 2001.

[RRT17] Maxim Raginsky, Alexander Rakhlin, and Matus Telgarsky. Non-convex learning via stochastic gradient Langevin dynamics: a nonasymptotic analysis. *arXiv preprint arXiv:1702.03849*, 2017.

[RT96] Gareth O Roberts and Richard L Tweedie. Exponential convergence of Langevin distributions and their discrete approximations. *Bernoulli*, pages 341–363, 1996.

[SV16] Sushant Sachdeva and Nisheeth K Vishnoi. The mixing time of the Dikin walk in a polytope—a simple proof. *Operations Research Letters*, 44(5):630–634, 2016.

[VA93] Pravin M Vaidya and David S Atkinson. A technique for bounding the number of iterations in path following algorithms. In *Complexity in Numerical Optimization*, pages 462–489. World Scientific, 1993.

[Vai89] Pravin M Vaidya. A new algorithm for minimizing convex functions over convex sets. In *30th Annual Symposium on Foundations of Computer Science, 1989*, pages 338–343. IEEE, 1989.

[Vem05] Santosh Vempala. Geometric random walks: a survey. *Combinatorial and Computational Geometry*, 52(573-612):2, 2005.

[WT11] Max Welling and Yee W Teh. Bayesian learning via stochastic gradient Langevin dynamics. In *Proceedings of the 28th International Conference on Machine Learning 2011*, pages 681–688, 2011.

[ZLC17] Yuchen Zhang, Percy Liang, and Moses Charikar. A hitting time analysis of stochastic gradient Langevin dynamics. *arXiv preprint arXiv:1702.05575*, 2017.