Journal of Machine Learning Research 18 (2017) 1-29 Submitted 6/14; Revised 11/16; Published 10/17

Efficient Sampling from Time-Varying Log-Concave Distributions

Hariharan Narayanan HARIN@UW.EDU
Department of Statistics and Department of Mathematics
University of Washington

Alexander Rakhlin RAKHLIN@WHARTON.UPENN.EDU
Department of Statistics
University of Pennsylvania

Editor: Robert McCulloch

Abstract

We propose a computationally efficient random walk on a convex body which rapidly mixes with respect to a fixed log-concave distribution and closely tracks a time-varying log-concave distribution. We develop general theoretical guarantees on the required number of steps; this number can be calculated on the fly according to the distance from and the shape of the next distribution. We then illustrate the technique on several examples. Within the context of exponential families, the proposed method produces samples from a posterior distribution which is updated as data arrive in a streaming fashion. The sampling technique can be used to track time-varying truncated distributions, as well as to obtain samples from a changing mixture model, fitted in a streaming fashion to data. In the setting of linear optimization, the proposed method has oracle complexity with best known dependence on the dimension for certain geometries. In the context of online learning and repeated games, the algorithm is an efficient method for implementing no-regret mixture forecasting strategies. Remarkably, in some of these examples, only one step of the random walk is needed to track the next distribution.[1]

1. Introduction

Let $\mathcal{K}$ be a compact convex subset of $\mathbb{R}^d$ with non-empty interior. Let $\mu_0, \ldots, \mu_t, \ldots$ be a sequence of probability measures with support on $\mathcal{K}$. Suppose each probability distribution μ_t has a density

$$\frac{d\mu_t(x)}{dx} = \frac{e^{-s_t(x)}}{Z_t}\ , \qquad Z_t = \int_{x\in\mathcal{K}} e^{-s_t(x)}dx \tag{1}$$

with respect to the Lebesgue measure, where each $s_t(x)$ is a convex function on $\mathcal{K}$. This paper proposes a Markov Chain Monte Carlo method for sequentially sampling from these distributions. The method comes with strong mixing time guarantees, and is shown to be applicable to a variety of problems. Observe that, by definition, the distributions μ_t are

1. An extended abstract containing partial results appeared in the proceedings of the NIPS 2010 conference (Narayanan and Rakhlin, 2010).

log-concave, and thus our work falls within the emerging body of literature on sampling from log-concave distributions.

The problem of sampling from distributions arises in many areas of statistics, most notably in Bayesian inference (Robert and Casella, 2004). In particular, Sequential Monte Carlo methods (Doucet et al., 2001) aim to sample from time-varying distributions. The need for such methods arises, for instance, in the case of online arrival of data: it is desirable to be able to update the posterior distribution at a low computational cost. If the distributions are changing "slowly" with time, sequential methods can re-use samples from the previous distribution and perform certain re-weighting to track the next distribution, thus saving computational resources. These ideas are exploited in particle filtering methods (see (Chopin, 2002; Doucet et al., 2001) and references therein). Beyond Bayesian inference, other applications of sampling from distributions include simulated annealing, global optimization, and regret minimization.

The main critique of the MCMC methods is, in many situations, the lack of mixing time analysis. In practice, the number of steps of the chain required to obtain an honest sample from a distribution is mostly calculated based on heuristics. There is a growing body of literature that presents exceptions to these heuristic approaches (Diaconis, 2013). Coupling methods, spectral gap methods, as well as the more recent study of positive Ricci curvature, yield geometric decrease of the distance to the desired stationary distribution – a property known as *geometric ergodicity*. The most well-understood cases in this context are those with a finite or countable state space (see (Meyn and Tweedie, 2009; Diaconis, 2009)). In contrast, we are interested in a random walk on a non-discrete set.

This paper is focused on a particular circle of problems defined via log-concave distributions. These distributions constitute an important subset of the set of unimodal distributions, a fact that has been recognized within Statistics (see e.g. (Walther, 2009)). We are not the first to study mixing times for such distributions: this line of work started with the breakthrough paper of (Dyer et al., 1991), followed by a series of improvements (Frieze et al., 1994; Lovász, 1999; Lovász and Vempala, 2006, 2007). However, the recent advances in (Kannan and Narayanan, 2012) on sampling from convex bodies give an edge to obtaining stronger guarantees. In particular, a variant of the Dikin walk studied in this paper is analyzed in (Narayanan, 2016) for the case of log-linear distributions. Extending the study of this random walk, we show rapid mixing to a log-concave distribution, and further show that we can provably track a *changing* log-concave distribution with a small number (or even only *one step*) of a random walk, provided that the distribution changes slowly enough. *Such a result seems out of reach with other random walk methods due to the lack of scale-free bounds on conductance.*

We assume that we can compute a *self-concordant barrier* (see Section 5 and Appendix 8) for the set $\mathcal{K}$, a requirement that is satisfied in many cases of interest. For instance, the self-concordant barrier can be readily computed in closed form if $\mathcal{K}$ is defined via linear and quadratic constraints. While the availability of the barrier is a stronger assumption than, for instance, access to a separation oracle for $\mathcal{K}$, the barrier gives a better handle on the geometry of the space and yields fast mixing of the Markov chain.

In Section 5, we illustrate the method within several diverse application domains. As one of the examples, we consider the problem of updating the posterior with respect to a conjugate prior in an exponential family, where the parameter is taking values in a space of a

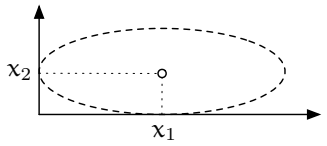
fixed dimensionality given by the sufficient statistics. The constraints then constitute a prior knowledge about the possible location of the parameter. As another example, we consider sampling from a time-varying truncated distribution, as well as the extension to sampling from mixture models fitted to streaming data. We employ the sampling technique to the classical problem of linear optimization via simulated annealing. The final example concerns the problem of regret minimization where the log-concave distribution arises naturally from the exponential weighting scheme.

The paper is organized as follows. In the next section we study the geometry of the set $\mathcal{K}$ induced by a self-concordant barrier and prove a key isoperimetric inequality in the corresponding Riemannian metric. The Markov chain for a given log-concave distribution is defined in Section 3. Conditions on the size of a step are introduced in Section 3.1, and a lower bound on the conductance of the chain is proved in Section 3.2. Section 4 contains main results about tracking time-varying distributions given appropriate measures of change between time steps. Section 5 is devoted to applications. Finally, Sections 6 and 7 contain all the remaining proofs.

2. Geometry Induced by the Self-Concordant Barrier

The Markov chain studied in this paper uses as a proposal a Gaussian distribution with a covariance that approximates well the local geometry of the set $\mathcal{K}$ at the current point. This local geometry plays a crucial role in the theory of interior point methods for optimization, yet for our purposes a handle on the local geometry yields a good lower bound on *conductance* of the Markov chain. Further intriguing similarities between optimization and sampling will be pointed out throughout the paper.

We refer to (Nemirovskii, 2004) for an introduction to the theory of interior point methods, a subject centered around the notion of a self-concordant barrier. Once we have defined a self-concordant barrier for $\mathcal{K}$, the local geometry is defined through the Hessian of the barrier at the current point. For the reader unfamiliar with the literature on self-concordant barriers, it is useful to think of the simple 2-dimensional example of $F(x) = -\log x_1 - \log x_2$ defined on the positive quadrant $\mathcal{K} = [0, \infty)^2$, with $x = (x_1, x_2)$. The Hessian of F at x is $\begin{bmatrix} 1/x_1^2 & 0 \\ 0 & 1/x_2^2 \end{bmatrix}$, and a unit ball centered at x and reshaped according to this matrix is the ellipse that touches the axes, as shown below. That is, the ellipsoid corresponds well to the



local geometry of $\mathcal{K}$ at the point x.

For a function F on the interior $int(\mathcal{K})$ having continuous derivatives of order k, for vectors $h_1, \ldots, h_k \in \mathbb{R}^d$ and $x \in int(\mathcal{K})$, for $k \geq 1$, we recursively define

$$D^k F(x)[h_1, \ldots, h_k] \triangleq \lim_{\epsilon \to 0} \frac{D^{k-1}(x + \epsilon h_k)[h_1, \ldots, h_{k-1}] - D^{k-1}(x)[h_1, \ldots, h_{k-1}]}{\epsilon}$$

where $D^0F(x) \triangleq F(x)$. Let F be a self-concordant barrier of $\mathcal{K}$ with a parameter ν (see Appendix 8 for the precise definition and Section 5 for examples). The value of the parameter ν depends on the shape of $\mathcal{K}$ and the choice of F. The barrier parameter will enter many of the results in this paper, but it is worth mentioning that there always exists a barrier with $\nu = O(d)$.

The barrier induces a Riemannian metric whose metric tensor is the Hessian of F (Nesterov and Todd, 2008). In other words, the metric tensor on the tangent space at x assigns to a vector v the length

$$\|v\|_x^2 \triangleq D^2F(x)[v, v],$$

and to a pair of vectors v, w, the inner product

$$\langle v, w\rangle_x \triangleq D^2F(x)[v, w] \ .$$

The unit ball in $\|\cdot\|_x$ around a point x is called the *Dikin ellipsoid* (Nemirovskii, 2004), and it describes well the local geometry of $\mathcal{K}$ at x in the following sense: (i) the unit Dikin ellipsoid at any point is contained in $\mathcal{K}$, and (ii) the Dikin ellipsoid of radius $r = 2(1 + 3\nu)$ contains $\mathrm{sym}(\mathcal{K}, x) = \{\mathcal{K} - x\} \cap -\{\mathcal{K} - x\}$.

The random walk, introduced in the next section, is *anisotropic*, i.e. the steps change in size and shape from point to point. It is then useful to connect the properties of this random walk directly to the Riemannian distance that is defined in terms of the Hessian of F. More precisely, for $x, y \in \mathcal{K}$, let $\rho(x, y)$ be the Riemannian distance $\rho(x, y) = \inf_\Gamma \int_z \|d\Gamma\|_z$ where the infimum is taken over all rectifiable paths Γ from x to y. Let $\mathcal{M}$ be the metric space whose point set is $\mathcal{K}$ and metric is ρ, and define $\rho(S_1, S_2) = \inf_{x\in S_1, y\in S_2} \rho(x, y)$. The first main ingredient of the analysis is an isoperimetric inequality.

Theorem 1 *Let S_1 and S_2 be measurable subsets of $\mathcal{K}$ and μ a probability measure supported on $\mathcal{K}$ that possesses a density whose logarithm is concave. Then it holds that*

$$\mu((\mathcal{K} \setminus S_1) \setminus S_2) \geq \frac{1}{2(1 + 3\nu)}\rho(S_1, S_2)\mu(S_1)\mu(S_2).$$

The theorem ensures that two subsets well-separated in ρ distance must have a large mass between them. A lower bound on conductance of our Markov chain will follow from this isoperimetric inequality. We remark that convexity of the set $\mathcal{K}$ is crucial for the above property. A classical example of a non-convex shape with a "bottleneck" is a dumbbell. For this body, the above statement clearly fails, and a "local" random walk on such a body gets trapped in either of the two parts for a long time.

2.1 Connections to Interior Point Methods

Interestingly, the idea of tracking a changing distribution with only one step of a random walk parallels the technique of following a central path in the theory of interior point methods for optimization, as discussed in (Narayanan and Rakhlin, 2010). In the analysis of interior point methods, the local (according to the Dikin ellipsoid introduced in the next section) quadratic convergence of the Damped Newton step counters the slowly changing "temperature parameter" of the barrier to ensure sufficiency of one optimization step; in our method, the geometric ergodicity of the scale-free random walk (which is based on the

local shape of the Dikin ellipsoid) balances the additive change in the distribution due to the changing temperature. Once again, the scale-free property of the Dikin walk, introduced below, is crucial for drawing this parallel between one-step interior point methods and our one-step random walk.

We remark that a different parallel between sampling and interior point methods has been recently outlined in (Abernethy and Hazan, 2016). The authors show that the *mean* of the log-concave distribution coincides with the central path when the barrier is chosen to be entropic. Since one still needs to sample from the distribution, this does not imply an "algorithmic equivalence" of the two methods unless one proves a scale-free bound on conductance, as we do in this paper.

3. The Markov Chain

Let $\mathcal{B}$ be the Borel σ-field on $\mathcal{K}$. Given an initial probability measure on $\mathcal{K}$, a Markov chain is specified by a collection of one-step transition probabilities

$$\{\mathsf{P}(x, B), x \in \mathcal{K}, B \in \mathcal{B}\}$$

such that $x \mapsto \mathsf{P}(x, B)$ is a measurable map for any $B \in \mathcal{B}$ and $\mathsf{P}_x(\cdot) \triangleq \mathsf{P}(x, \cdot)$ is a probability measure on $\mathcal{K}$ for any $x \in \mathcal{K}$.

For $x \in int(\mathcal{K})$, let G_x^r denote the unique Gaussian probability density function on $\mathbb{R}^d$ such that

$$G_x^r(y) \propto \exp\left(-\frac{d\|x-y\|_x^2}{r^2} + V(x)\right), \qquad V(x) \triangleq \frac{1}{2}\ln\det D^2F(x)$$

and r is a parameter that is chosen according to a condition specified below. The covariance of this distribution is given by the Hessian of F at point x, and thus the contour lines are scaled Dikin ellipsoids.

The Markov chain considered in this paper is based on the Dikin Walk introduced by Kannan and Narayanan (2012). Adapted to sampling from log-concave distributions in this paper, the Markov chain is parametrized by a convex function s and a step size r. Rather than writing out the unwieldy explicit form of the transition kernel P_x, we can give it implicitly as the following random walk:

With probability $1/2$, set $w := x$.

With probability $1/2$, sample z from G_x^r and

If $z \notin \mathcal{K}$, let $w := x$.

If $z \in \mathcal{K}$, let $w := \begin{cases} z & \text{with prob.} \quad \min\left(1,\ \frac{G_z^r(x)\exp(s(x))}{G_x^r(z)\exp(s(z))}\right) \\ x & \text{otherwise.} \end{cases}$

The Markov chain is *lazy*, as it stays at the current point with probability at least $1/2$. This ensures uniqueness of the stationary distribution (Lovász and Simonovits, 1993). Furthermore, a simple calculation shows that the detailed balance conditions are satisfied with

respect to a stationary distribution μ whose density (with respect to the Lebesgue measure) is proportional to $\exp(-s(x))$. Indeed, to see that $\mu(x)\mathsf{P}_x(dz) = \mu(z)\mathsf{P}_z(dx)$, it suffices to observe that

$$\exp(-s(x))G_x^r(z)\min\left(1,\ \frac{G_z^r(x)\exp(s(x))}{G_x^r(z)\exp(s(z))}\right) \\ = \exp(-s(z))G_z^r(x)\min\left(1,\ \frac{G_x^r(z)\exp(s(z))}{G_z^r(x)\exp(s(x))}\right).$$

Therefore the Markov chain is reversible and has the desired stationary measure μ.

The value of r has a specific meaning: most of the y's sampled from G_x^r are within a thin "Dikin shell" of radius proportional to $(\mathbb{E}\|x-y\|_x^2)^{1/2} = r$ by measure-concentration arguments. We will therefore refer to r as the effective "step size". An important and non-trivial result from the theory of interior point methods is that the unit Dikin ellipsoid is contained in the set $\mathcal{K}$ and gives a good approximation to the local geometry of the set (see Figure 1 below). Thanks to this fact, the sampling procedure has in general better mixing properties than the Ball Walk (Lovász and Simonovits, 1993; Vempala, 2005).

3.1 Step Size Conditions

The analysis of the Markov chain requires the steps r to be not too large to ensure that different enough transition probability functions happen only for far away points. The precise upper bounds on r depend on the convex function $s(x)$ and can be calculated on the fly when we move to the setting of a time-varying function. We give four conditions:

Sufficient Condition 1 (Linear Functions) *If s is linear, we may set $r = 1/d$.*

Sufficient Condition 2 (Lipschitz Functions) *For a function s that is L-Lipschitz with respect to the Euclidean norm, we may set the step size $r = \min\left\{\frac{1}{d}, \frac{1}{L}\right\}$.*

Sufficient Condition 3 (Smooth Functions) *Suppose s has Lipschitz-continuous gradients: there exists $\sigma > 0$ such that $\|\nabla s(x) - \nabla s(y)\| \le \sigma\|x-y\|$. We may then set the step size to be $\min\left\{\frac{1}{d}, \frac{1}{\sqrt{\sigma}}\right\}$.*

These three conditions can be shown to follow from a more general sufficient step size condition that is based on "local" information:

Sufficient Condition 4 (General Condition) *Fix constants $C, C' > 0$. Given the convex function $s(x)$, the step size $r \le \min\left\{\frac{1}{d}, r^*\right\}$ is a valid choice if there exists a linear function $< g, x >$ such that*

$$r^* \le \sup\left\{\ r\ :\ \forall z, w \in \mathcal{K}\ \ \textit{with}\ \ \|z-w\|_z \le C'r, \quad \Big|s(z) - s(w) - \langle g, z-w\rangle\Big| < C\right\}$$

The condition says that for two points, with one being inside the $O(r)$-Dikin ellipsoid around the other point, the function is within a constant of being linear. It follows from the last condition that, for instance, if $s(x) = \langle b, x\rangle + a(x)$ is a sum of a linear and a non-linear Lipschitz part, the step size is only affected by the Lipschitz constant of the non-linear part.

It is simple to verify that the step size in Condition 2 satisfies Condition 4. Indeed, for any w such that $\|z-w\|_z \le C'r$, we have $\|z-w\| \le C''rR$ (where R is the radius of the largest ball contained in $\mathcal{K}$). Take g_z and g_w to be any subgradients of s at z and w, respectively. We then have

$$|s(z) - s(w) - \langle g_w, z-w\rangle| \le \langle g_z - g_w, z-w\rangle \le 2L\|z-w\| \le 2 \ .$$

Notice that for Condition 3, the above calculation becomes

$$\langle g_z - g_w, z-w\rangle \le \sigma\|z-w\|^2 \le 1 \ .$$

In the remainder of this paper, C will denote a universal constant that may change from line to line. The exact value of the final constant in Lemma 4 below can be traced in the proofs; we omit this calculation for the sake of brevity.

3.2 Conductance of the Markov Chain

In order to show rapid mixing of the proposed Markov chain, we prove a lower bound on its *conductance*

$$\Phi \triangleq \inf_{\mu(S_1)\le\frac{1}{2}} \frac{\int_{S_1} \mathsf{P}_x(\mathcal{K}\setminus S_1)d\mu(x)}{\mu(S_1)}, \tag{2}$$

where P_x is the one-step transition function defined earlier. Once such a lower bound is established, the following general result on the reduction of distance between distributions will imply exponentially fast convergence.

Theorem 2 ((Lovász and Simonovits, 1993)) *Let γ_0 be the initial distribution for a lazy reversible ergodic Markov chain whose conductance is Φ and stationary measure is γ. For every bounded f, let $\|f\|_\gamma \triangleq \sqrt{\int_{\mathcal{K}} f(x)^2 d\gamma(x)}$. For any fixed f, let Ef be the map that takes x to $\int_{\mathcal{K}} f(y)d\mathsf{P}_x(y)$. Then if $\int_{\mathcal{K}} f(x)d\gamma(x) = 0$, it holds that*

$$\|E^k f\|_\gamma \le \left(1 - \frac{\Phi^2}{2}\right)^{k/2} \|f\|_\gamma \ .$$

To prove a lower bound on conductance Φ, we first relate the Riemannian metric ρ to the proposed Markov Chain. Intuitively, the following result says that for close-by points, their transition distributions cannot be far apart in the total variation distance d_{TV}.

Lemma 3 *If $x, y \in \mathcal{K}$ and $\rho(x,y) \le \frac{r}{C\sqrt{d}}$ for some constant C, then*

$$d_{TV}(\mathsf{P}_x, \mathsf{P}_y) \le 1 - \frac{1}{C'}$$

for some constant C'.

Lemma 3 together with the isoperimetric inequality of Theorem 1 give a lower bound on conductance of the Markov Chain.

Lemma 4 *Let μ be a log-concave distribution with support on $\mathcal{K}$ whose density with respect to the Lebesgue measure is proportional to $\exp\{-s(x)\}$, and suppose an appropriate step size condition (Section 3.1) for the Markov chain is satisfied. Then there exists a constant $C > 0$ such that the conductance of the above Markov chain is bounded below as*

$$\Phi \geq \frac{r}{C\nu\sqrt{d}} \,.$$

We remark that the step size r enters the lower bound on Φ. While we would like the steps to be large, the conditions outlined earlier dictate a limitation on how large r can be. In particular, we always have $r \leq 1/d$. The step size needs to be even smaller for functions s for which a linear approximation is poor.

We also remark that Lemma 4 establishes a *scale-free* bound on the conductance, that is, a bound that does not depend on the measure of the set S_1 in the definition (2). Such a scale-free conductance is needed for the geometric ergodicity of the chain.

4. Tracking the Distributions

Having specified the Markov chain and the step size, we now turn to the problem of tracking a sequence of distributions $\mu_1, \ldots, \mu_t, \ldots$. For each $t \geq 1$, define a Markov chain with parameters r_t and s_t, and let its transition kernel be denoted by $\mathsf{P}_t(x, B)$ for $x \in \mathcal{K}$ and $B \in \mathcal{B}$. Let Φ_t denote the conductance of this chain. The chain will be run for τ_t steps starting from the end of the chain at time $t-1$. Formally, let the i-th step of the t-th chain be denoted by the random variable $X_{t,i}$. Define $\tau_0 = 0$ and let $\sigma_{0,0}$ be the initial distribution of $X_{0,0}$. Then $X_{t,i}$ has distribution

$$\sigma_{0,0}\mathsf{P}_1^{\tau_1} \cdots \mathsf{P}_{t-1}^{\tau_{t-1}}\mathsf{P}_t^i$$

and we have made the identification $X_{s,\tau_s} = X_{s+1,0}$, gluing the successive chains together. Let the distribution of $X_{t,i}$ be denoted by $\sigma_{t,i}$. By the definition of the chain, $\sigma_{t,i}$ is a distribution with bounded density, supported on $\mathcal{K}$.

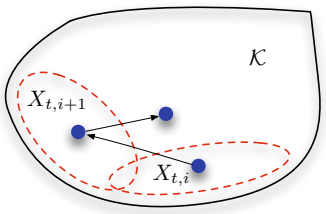


Figure 1: Steps of the Dikin Walk. The next point is sampled from a Gaussian distribution with a shape (contours depicted with dashed lines) corresponding to Dikin ellipsoids. These ellipsoids approximate well the local geometry.

4.1 Measuring the Change

Let $\|\cdot\|_t$ denote the $\mathcal{L}_2$ norm with respect to the measure μ_t, defined as $\|f\|_t = \left(\int_{\mathcal{K}} f^2 d\mu_t\right)^{1/2}$ for a measurable function $f : \mathcal{K} \to \mathbb{R}$. Further, let $\|\cdot\|_{\mathcal{K}}$ denote the supremum norm

$\|f\|_{\mathcal{K}} = \sup_{x\in\mathcal{K}} |f(x)|$ and let

$$\beta_{t+1} = \max\left\{\|d\mu_t/d\mu_{t+1}\|_{\mathcal{K}}\,, \|d\mu_{t+1}/d\mu_t\|_{\mathcal{K}}\right\}\ . \tag{3}$$

This ratio provides an upper bound on the point-wise change of the density function. A straightforward way to upper bound β_{t+1} is by writing

$$\sup_{x\in\mathcal{K}} \frac{e^{-s_t(x)}}{e^{-s_{t+1}(x)}} \frac{\int_{\mathcal{K}} e^{-s_{t+1}(x)}dx}{\int_{\mathcal{K}} e^{-s_t(x)}dx} \le \sup_{x\in\mathcal{K}} e^{2|s_t(x)-s_{t+1}(x)|}$$

and, hence,

$$\log\beta_{t+1} \le\ 2\|s_t(x) - s_{t+1}(x)\|_{\mathcal{K}}\ . \tag{4}$$

Another way to measure the change in successive distributions is with respect to the $\mathcal{L}_2$ norm:

$$\alpha_{t+1} = \|d\mu_t/d\mu_{t+1}\|_{t+1}\ . \tag{5}$$

In contrast to the point-wise change, the ratio α_{t+1} is more difficult to calculate. In this respect, the following result, which follows from the proof of (Lovász and Vempala, 2006; Kalai and Vempala, 2006), will be useful:

Lemma 5 *Let s_t be a convex function and $s_{t+1} = (1-\delta)^{-1} s_t$. Let μ_t and μ_{t+1} be defined as in* (1)*. Then*

$$\alpha_{t+1} \le \left(1 + \frac{\delta^2}{1-2\delta}\right)^{d/2}$$

In particular, if $\delta \le d^{-1/2} \le 1/3$, then $\alpha_{t+1} \le 5$.

We remark that the ratio between μ_t and μ_{t+1} measured in the supremum norm may be exponentially large, while the $\mathcal{L}_2$ change is small. As in (Lovász and Vempala, 2006; Kalai and Vempala, 2006), this fact will be crucial in this paper when we study simulated annealing.

4.2 Tracking the Distributions: Main Results

Denote the error in approximating the stationary distribution at the end of t-th chain by

$$\xi_t\ \triangleq\ \left\|\frac{d\sigma_{t,\tau_t}}{d\mu_t} - 1\right\|_t \tag{6}$$

and let

$$\Delta_t\ \triangleq\ \frac{r_t^2}{C\nu^2 d}\ .$$

Theorem 6 *The errors ξ_t satisfy the recurrence*

$$\xi_t \le (1-\Delta_t)^{\tau_t}(\beta_t^{3/2}\xi_{t-1} + \sqrt{\beta_t}(\beta_t - 1)) \tag{7}$$

for any $t \ge 1$.

Proof [Proof of Theorem 6] We iteratively apply Theorem 2 with $f = \frac{d\sigma_{t,j}}{d\mu_t} - 1$ and the stationary distribution $\gamma = \mu_t$, and observe that Ef takes $\sigma_{t,j}$ to $\sigma_{t,j+1}$. Then from Lemma 4, for $t \geq 1$ and $i \geq 1$,

$$\left\| \frac{d\sigma_{t,i}}{d\mu_t} - 1 \right\|_t \leq \left\| \frac{d\sigma_{t,0}}{d\mu_t} - 1 \right\|_t \cdot (1 - \Delta_t)^i$$

Using the first part of Lemma 11 (see Section 6)

$$\left\| \frac{d\sigma_{t,0}}{d\mu_t} - 1 \right\|_t \leq \beta_t^{3/2} \left\| \frac{d\sigma_{t,0}}{d\mu_{t-1}} - 1 \right\|_{t-1} + \sqrt{\beta_t}(\beta_t - 1),$$

concluding the proof. An alternative recurrence, using the second part of Lemma 11, is

$$\xi_t \leq (1 - \Delta_t)^{\tau_t}(\sqrt{\beta_t}\xi_{t-1} + \sqrt{\beta_t - 1}),$$

which is better for large β_t but worse for $\beta_t \approx 1$. ■

We would like to adaptively choose τ_t to make the right-hand side (7) small. While the value of the error ξ_{t-1} at the previous round is not available for this purpose, let us maintain an upper bound u_{t-1} on this error. Thus, we may write τ_t as a function $\tau_t(u_{t-1}, s_t, r_t, \beta_t)$. Suppose at round $t = 0$ we ensure that $\xi_0 \leq u_0$. Then, recursively, we may compute u_t as the upper bound in (7):

$$u_t \geq (1 - \Delta_t)^{\tau_t}(\beta_t^{3/2} u_{t-1} + \sqrt{\beta_t}(\beta_t - 1)) \tag{8}$$

Then, given the initial condition, we have $\xi_t \leq u_t$ for all $t \geq 0$.

Let us consider some consequences of Theorem 6. In particular, we are interested in situations when we can track the distributions with only *one step* of the random walk.

Corollary 7 *Let $\tau_t = 1$ for all $t \geq 1$ and suppose $\xi_0 \leq u_0 = \sqrt{\beta_0}(\beta_0 - 1)/\Delta_0$ with $\Delta_0 = \frac{1}{Cd^3\nu^2} \leq \frac{1}{2}$. Assume that β_t is non-decreasing and Δ_t is non-increasing in t, and suppose*

$$\beta_t^{3/2} \leq 1 + \frac{\Delta_t^2}{1 - \Delta_t} \tag{9}$$

for all $t \geq 1$. Then we have

$$\xi_t \leq u_t = \frac{\sqrt{\beta_t}(\beta_t - 1)}{\Delta_t}$$

for all $t \geq 0$. In particular, (9) is satisfied whenever $\beta_t - 1 \leq 0.4\Delta_t^2$.

The proof of the above corollary follows from the more general result:

Corollary 8 *Fix a sequence $\epsilon_0, \ldots, \epsilon_t, \ldots$ of positive target accuracies and assume $\xi_0 \leq \epsilon_0$. It is then enough to set*

$$\tau_t = \left\lceil \frac{1}{\Delta_t} \log \left(\beta_t^{3/2} \cdot \frac{\epsilon_{t-1}}{\epsilon_t} + \frac{\sqrt{\beta_t}(\beta_t - 1)}{\epsilon_t} \right) \right\rceil \tag{10}$$

in order to ensure $\xi_t \leq \epsilon_t$ for each $t \geq 0$.

Proof Immediate by writing

$$u_t = (1-\Delta_t)^{\tau_t}(\beta_t^{3/2}\epsilon_{t-1} + \sqrt{\beta_t}(\beta_t - 1)) \leq \epsilon_t,$$

solving for τ_t, and using the approximation $\log(1/(1-\Delta_t)) \geq \log(1+\Delta_t) \geq \Delta_t$. ■

We now consider the case when one has control on the $\mathcal{L}_2$ norm α_t of the change between successive distributions. First, observe that closeness of the distributions in the norm $\|\cdot\|_t$ implies closeness in total variation distance as

$$\int |d\sigma_{t,i} - d\mu_t| = \int \left| \frac{d\sigma_{t,i}}{d\mu_t} - 1 \right| d\mu_t \leq \left\| \frac{d\sigma_{t,i}}{d\mu_t} - 1 \right\|_t . \tag{11}$$

Proposition 9 *Fix a sequence $\epsilon_0, \ldots, \epsilon_t, \ldots$ of positive target accuracies and assume $d_{TV}(\sigma_{0,0}, \mu_0) \leq \epsilon_0$. Suppose we set*

$$\tau_t = \left\lceil \frac{1}{\Delta_t} \log\left(\frac{\alpha_t}{\epsilon_t}\right) \right\rceil . \tag{12}$$

Then the total variation distance between σ_{t,τ_t} and μ_t is bounded as

$$d_{TV}(\sigma_{t,\tau_t}, \mu_t) \leq \sum_{s=0}^{t} \epsilon_s \tag{13}$$

for each $t \geq 0$.

Proof For any $t \geq 1$, let us write

$$\sigma_{t,\tau_t} = \mu_t + \gamma_t \tag{14}$$

with a signed measure $\gamma_t = \sigma_{t,\tau_t} - \mu_t$. By way of induction, suppose (13) holds for time t. Consider the operator E_{t+1} corresponding to the random walk of the $t+1$-st chain. The operator acts on a function f by taking f to $\int_{\mathcal{K}} f(y) d\mathsf{P}_{t+1}(x,y)$. Then applying Theorem 2 to the function $d\mu_t/d\mu_{t+1} - 1$, we have

$$\left\| E_{t+1}^{\tau_{t+1}} \left(\frac{d\mu_t}{d\mu_{t+1}} - 1 \right) \right\|_{t+1} \leq \epsilon_{t+1}$$

by the choice of τ_{t+1} and the definition of α_{t+1}. That is, upon the action of $E_{t+1}^{\tau_{t+1}}$, μ_t is mapped to μ_{t+1} within an error of at most ϵ_{t+1} in the $\mathcal{L}_2$ sense (and, hence, in the total variation sense). Since the operator E_{t+1} is non-expanding in the $\mathcal{L}_1$ sense, total variation of γ_t does not increase under the action of $E_{t+1}^{\tau_{t+1}}$. In view of the inductive hypothesis for step t, we conclude $d_{TV}(\sigma_{t+1,\tau_{t+1}}, \mu_{t+1}) \leq \sum_{s=0}^{t} \epsilon_s + \epsilon_{t+1}$, as desired. ■

5. Applications

Before diving into the applications of the random walk, let us give several examples of sets $\mathcal{K}$ for which the self-concordant barrier F and its Hessian can be easily calculated. In the following examples, assume that $\mathcal{K}$ has non-empty interior.

Example 1 *Suppose $\mathcal{K}$ is given by m linear constraints of the form $\langle a_j, x\rangle \le b_j$, $j = 1, \dots, m$. Then $F(x) = -\sum_{j=1}^m \log(b_j - \langle a_j, x\rangle)$ is a self-concordant barrier with parameter $\nu = m$. The Hessian is easily computable:*

$$D^2 F(x) = \sum_{j=1}^m \frac{a_j a_j^\mathsf{T}}{(b_j - \langle a_j, x\rangle)^2} \ .$$

Example 2 *Let $\mathcal{K} = \{x \in \mathbb{R}^d : f_j(x) \le 0, j = 1, \dots, m\}$ where each f_j is a convex quadratic form. Then $F(x) = -\sum_{j=1}^m \log(-f_j(x))$ is a self-concordant barrier with parameter m. As an example, the function $-\log(R - \|x\|^2)$ is a self-concordant barrier for the unit Euclidean sphere $\{x : \|x\|^2 - 1 \le 0\}$, with parameter $\nu = 1$, and the Hessian is given by*

$$D^2 F(x) = \frac{2}{1 - \|x\|^2} I + \frac{4}{(1 - \|x\|^2)^2} x x^\mathsf{T} \ .$$

Importantly, there always exists a self-concordant barrier with $\nu = O(d)$; yet, for some convex sets (such as the sphere) the parameter can even be constant.

Self-concordant barriers can be combined: if F_j is ν_j-self-concordant for $\mathcal{K}_j$, $j = 1, \dots, m$, then $\sum_j F_j$ is $\sum_j \nu_j$-self-concordant for the intersection $\cap_i \mathcal{K}_i$, given that it has nonempty interior. Thus, closed forms for the Hessian of the barrier, required for defining G_x^r in our Markov chain, can be calculated for many sets $\mathcal{K}$ of interest. We refer to (Nemirovskii, 2004; Nesterov and Nemirovskii, 1994) for further powerful methods for constructing the barriers.

5.1 Sampling from Posterior in Exponential Families

Suppose data $y_1, y_2, \dots \in \mathcal{Y}$ are distributed i.i.d. according to a member of an exponential family with natural parameter x:

$$p(y|x) = \exp\{\langle x, T(y)\rangle - A(x)\} h(y)$$

where $A(x) = \int h(y) \exp\{\langle x, T(y)\rangle\}$ is a convex function and $T : \mathcal{Y} \mapsto \mathbb{R}^d$ is a sufficient statistic. Suppose $x \in \mathcal{K}$; that is, we have some knowledge about the support of the parameter. We have in mind the situation where data arrive one at a time and we are interested in sampling from the associated posterior distributions. The likelihood function after seeing $y_1, \dots, y_t$ is

$$\ell(x) \propto \exp\left\{ \left\langle x, \sum_{i=1}^t T(y_i) \right\rangle - tA(x) \right\}$$

and, together with a conjugate prior $\pi_{\kappa_1,\kappa_2}(x) \propto \exp\{\langle x, \kappa_1\rangle - \kappa_2 A(x)\}$ for some $(\kappa_1, \kappa_2) \in \mathbb{R}^{d+1}$, we obtain the posterior distribution at time t

$$p_t(x|y) \propto \exp\left\{\left\langle x, \kappa_1 + \sum_{i=1}^{t} T(y_i)\right\rangle - (t+\kappa_2)A(x)\right\} .$$

We apply the sampling technique to this scenario by defining

$$s_0(x) = -\langle x, \kappa_1\rangle + \kappa_2 A(x), \qquad s_t(x) = -\left\langle x, \kappa_1 + \sum_{i=1}^{t} T(y_i)\right\rangle + (t+\kappa_2)A(x) .$$

It remains to calculate the number of steps required to track the distributions as additional data arrive one-by-one. Let L be the Lipschitz constant of $A(x)$ over $\mathcal{K}$ with respect to Euclidean norm, and let us assume L to be finite. Then Condition 4 is satisfied with $r = \min\left\{\frac{1}{(t+\kappa_2)L}, \frac{1}{d}\right\}$. Furthermore, we may set

$$\beta_t = \sup_{x\in\mathcal{K}} \exp\left\{2|\langle x, T(y_t)\rangle - A(x)|\right\},$$

a quantity that depends on the observed data. Importantly, we do not need to provide an a priori data-independent bound of this type, which might not be finite.

Suppose we would like to maintain a constant level $\epsilon > 0$ of accuracy at each step t. Corollary 8 guarantees this accuracy if each chain is run for

$$\tau_t = \mathcal{O}\left(\nu^2 d \max\{(t+\kappa_2)^2 L^2, d^2\} + \log(1/\epsilon)\right) .$$

One of the features of this bound is a relatively benign dependence on the dimension d, especially if the geometry of the set $\mathcal{K}$ allows the parameter $\nu = \mathcal{O}(1)$, as in the case of a sphere. On the negative side, the number of steps needed after seeing t data points is proportional to t^2. Such an adverse dependence, however, is to be expected as the posterior distribution becomes concentrated very quickly.

We now demonstrate that stronger results can be achieved under additional assumptions via Condition 3. Suppose that A is smooth: there exists $H \succeq 0$ such that

$$A(x) \le A(w) + \langle \nabla A(x), w - x\rangle + (w-x)^\intercal H (w-x)$$

for any $w, x \in \mathcal{K}$. This is a natural assumption, as the second derivative of the log normalization function A corresponds to the variance of the random variable with the given parameter; furthermore, A is differentiable. Let $\lambda_{\max}$ be the largest eigenvalue of H. Then the condition yields $r_t = \frac{C}{\sqrt{(t+\kappa_2)\lambda_{\max}}}$. To obtain ϵ-accuracy, it suffices to set

$$\tau_t = \mathcal{O}\left(\nu^2 d \max\{(t+\kappa_2)\lambda_{\max}, d^2\} + \log(1/\epsilon)\right),$$

which has only linear dependence on the size of the data seen so far.

We remark that each step of the random walk requires evaluation of the log-partition function $A(x)$. If this function is not available in closed form, we may approximate the value $A(x)$ for each query x. In order to do this, we may run an additional sampling procedure with $s'(x) = \langle x, T(y)\rangle$. Alternatively, we may appeal to known methods for this problem, such as Hit-and-Run (Vempala, 2005).

5.2 Examples of Sampling from Drifting Truncated Distributions

In the previous example, we employed the Markov chain to sample a parameter from a log-concave posterior. We now turn to the question of sampling from a log-concave distribution restricted to a convex set. This problem has a long history (see e.g. (Devroye, 1986; Gilks and Wild, 1992)), and it is recognized that sampling from truncated distributions is difficult even for nice forms such as the Normal distribution. One successful approach to this problem is the Gibbs sampling method (Robert, 1995; Damien and Walker, 2001), yet the rate of convergence is not generally available. The MCMC method of this paper yields a provably fast algorithm for such situations. Furthermore, we can track a drifting distribution over $\mathcal{K}$ with a small number of steps.

For illustration purposes, we study a simple example of a truncated Normal distribution; the same techniques, however, apply more generally. To simplify calculations, suppose the distributions μ_t are defined to be $\mathcal{N}(\mathbf{c}_t, \frac{1}{d}I)$ over a convex compact set $\mathcal{K} \subset \mathbb{R}^d$ and suppose the mean $\mathbf{c}_t$ is drifting within a Euclidean ball of radius R. With the definition in (1) we have $s_t(x) = \frac{1}{2}\|x - \mathbf{c}_t\|^2$. Define the drift $\delta_t = \|\mathbf{c}_t - \mathbf{c}_{t-1}\|$. In view of (4),

$$\log \beta_t \leq \sup_{x \in \mathcal{K}} \|\mathbf{c}_t - \mathbf{c}_{t-1}\| \cdot \|2x - \mathbf{c}_t - \mathbf{c}_{t-1}\| \leq C_{R,\mathcal{K}} \delta_t$$

where $C_{R,\mathcal{K}}$ depends on the radius R and the radius of a smallest Euclidean ball enclosing $\mathcal{K}$. In the same manner, the Lipschitz constant of $s_t(x)$ over $\mathcal{K}$ can be upper bounded by $L_{R,\mathcal{K}}$ that depends solely on the two radii. We may thus set the step size to be $r_t = \min\{\frac{1}{d}, \frac{1}{L_{R,\mathcal{K}}}\}$. If we aim for a fixed target accuracy ϵ for all t, by Corollary 8, it is enough to make

$$\tau_t = \left\lceil \frac{1}{\Delta_t} \log \left(\beta_t^{3/2} + \frac{\sqrt{\beta_t}(\beta_t - 1)}{\epsilon} \right) \right\rceil \tag{15}$$

steps. In the case that the drift δ_t is small enough, only one step is sufficient. To quantify the regime when this happens, observe that $\beta_t \leq \exp\{C_{R,\mathcal{K}}\delta\} \leq 1 + C\delta_t$, and it is then enough to require

$$\delta_t = \mathcal{O}\left(\Delta_t^2\right) = \mathcal{O}\left(\frac{\min\{1/d^2, 1/L_{R,\mathcal{K}}^2\}}{\nu^2 d} \right)$$

in view of (9). It is quite remarkable that the one-step random walk can track the changing distribution up to the accuracy $\mathcal{O}\left(\delta_t \frac{\nu^2 d}{r_t^2}\right)$, proportional to the size of the drift. Of course, better accuracy can be achieved by performing more steps, as per Corollary 8.

Another related application is to modeling with mixtures of log-concave distributions. Such models have been successful in clustering (McLachlan and Peel, 2000; Walther, 2009), with a mixture of normal distributions being a classical example (Fraley and Raftery, 2002). A mixture of parametric log-concave distributions can be written as $\sum_{i=1}^k \alpha_i \pi_i(\theta_i; x)$; here α_i are positive mixing weights summing to one, and π_i are a distributions on $\mathcal{K}$ parametrized by θ_i. A classical method for fitting models to data is the EM algorithm. Given that the parameters $\{\theta_i\}_{i=1}^k$ and the mixing weights $\{\alpha_i\}_{i=1}^k$ have been estimated from data, one may require random samples from this model for integration or other purposes. Given our procedure for sampling from a single log-concave distribution, one may simply pick the mixture according to the weights α_i and then sample from the component. The situation

becomes interesting in the case of online arrival of data, when we need to re-compute the EM solution in light of additional data. By the arguments of (Rakhlin and Caponnetto, 2006; Caponnetto and Rakhlin, 2006), the solution to clustering problems (the analysis was performed for square loss) is *stable* in the following sense: addition of $o(\sqrt{n})$ new data to a sample of size n is unlikely to drastically move the solution (the argument is based on uniqueness of the maximum of an empirical process). This in turn implies that the parameters $\{\theta_i\}$ are unlikely to change by a large amount, and we may thus use the method of sampling from a drifting distribution described earlier. We also remark that the method can be easily parallelized since the Markov chains for the k components do not interact.

5.3 Simulated Annealing for Convex Optimization

Let $f(x)$ be a proper convex 1-Lipschitz function. The aim of convex optimization is to find $\tilde{x}$ with the property $f(\tilde{x}) - \min_{x\in\mathcal{K}} f(x) \leq \epsilon$ for a given target accuracy $\epsilon > 0$. We consider the special case of linear function $f(x) = \langle \ell, x\rangle$, known as Linear Optimization. Complexity of an optimization procedure is often measured in terms of *oracle calls* – queries about the unknown function. A query about the function value is known as the zero-th order information, while a query about a subgradient at a point – as the first order information. In the case that the oracle answer is given without noise, it is known that the complexity scales as $\mathcal{O}\left(\mathrm{poly}(d, \log(1/\epsilon))\right)$. The state-of-the-art result here is the method of (Kalai and Vempala, 2006; Lovász and Vempala, 2006) which attains the $d^{4.5}$ dependence on the dimension.

We now apply our machinery to obtain a $\mathcal{O}\left(\nu^2 d^{3.5}\log(1/\epsilon)\right)$ method. In particular, this yields an improved $d^{3.5}$ dependence on the dimension for the case when $\mathcal{K}$ has a favorable geometry: there exists a self-concordant barrier with a parameter $\nu = \mathcal{O}(1)$.

We use the annealing scheme of (Kalai and Vempala, 2006). To this end, we set $s_t = \left(1 - d^{-1/2}\right)^{-t} f$ and observe that the assumption of Lemma 5 is satisfied with $\delta = d^{-1/2}$. Since functions are linear, we may set the step size $r_t = 1/d$ for all t. Hence, $\alpha_t \leq 5$ whenever $d > 8$ (and a different constant can be obtained for smaller d from the proof). By Proposition 9 with a constant accuracy $\epsilon_t = \epsilon \cdot (\sqrt{d}\log(d/\epsilon))^{-1}$, by making

$$\tau_t = \left\lceil C d^3 \nu^2 \log\left(\frac{5\sqrt{d}\log(d/\epsilon)}{\epsilon}\right)\right\rceil \tag{16}$$

steps for $t = 1, \ldots, k$, we guarantee

$$d_{TV}(\sigma_{k,\tau_k}, \mu_k) \leq k\epsilon(\sqrt{d}\log(d/\epsilon))^{-1} \ . \tag{17}$$

According to (Kalai and Vempala, 2006, Lemma 4.1), if X is chosen from a distribution with density proportional to $\exp\{-T^{-1}\langle \ell, x\rangle\}$, with $\|\ell\| = 1$ and some temperature $T > 0$, then

$$\mathbb{E}(\langle \ell, X\rangle) - \min_{x\in\mathcal{K}} \langle \ell, x\rangle \leq dT.$$

Hence, we take the desired temperature to be $T = \epsilon/d$, and the number of chains that permits the annealing schedule to reach this temperature can be calculated as $k = \sqrt{d}\log(\frac{d}{\epsilon})$. In view of (17), the final output of the procedure is an ϵ-accurate solution to the optimization problem. The complexity of the method is then $\mathcal{O}(d^{3.5}\nu^2\log^2(d/\epsilon))$.

This result can be extended to Lipschitz convex functions beyond linear optimization. However, the step size condition for convex Lipschitz functions requires the steps to be $\mathcal{O}(1/\epsilon)$ towards the end of the annealing schedule. This in turn implies only a suboptimal $\tilde{\mathcal{O}}(\nu^2 d/\epsilon^2)$ complexity. It is an open question of whether Dikin Walk can handle such annealing schedules in a more graceful manner.

5.4 Sequential Prediction

Another application of the proposed sampling technique is to the problem of *sequential prediction* with convex cost functions. Within this setting, the learner (or, the Statistician) is tasked with making a series of predictions while observing a sequence of outcomes on which we place no distributional assumptions. The goal of the learner is to incur cost comparable to that of a fixed strategy chosen in hindsight after observing the data. Initially studied by Hannan (Hannan, 1957), Blackwell (Blackwell, 1956), and Cover (Cover, 1965), the problem of achieving low *regret* for all sequences has received much attention in the last two decades, and we refer the reader to (Cesa-Bianchi and Lugosi, 2006) for a comprehensive treatment. As we show in this section, a strategy that exponentially down-weighs the decisions with large costs is a good regret-minimization strategy, and this exponential form is amenable to the sampling technique of this paper whenever the costs are convex.

More specifically, let $\mathcal{K} \subset \mathbb{R}^d$ be a convex compact set of decisions of the learner. Let $\boldsymbol{\ell}_1, \ldots, \boldsymbol{\ell}_T$ be a sequence of unknown cost functions $\boldsymbol{\ell}_t : \mathcal{K} \to \mathbb{R}$. On round t, the learner chooses a distribution (or, a *mixed strategy*) μ_{t-1} supported on $\mathcal{K}$ and "plays" a decision $Y_t \sim \mu_{t-1}$.[2] Nature then reveals the next cost function $\boldsymbol{\ell}_t$. For example, in the well-studied problem of sequential probability assignment, the Statistician predicts the probability $x_t \in [0,1] = \mathcal{K}$ of the next outcome $\{0,1\}$ and incurs the cost $\boldsymbol{\ell}_t(x_t) = |x_t - y_t|$ with respect to the actual outcome y_t. A randomized strategy Y_t then incurs a cost $\boldsymbol{\ell}_t(Y_t)$. The goal of the learner is to minimize *expected regret*

$$\mathrm{Reg}_T(U) \triangleq \mathbb{E}\left[\sum_{t=1}^{T} \boldsymbol{\ell}_t(Y_t) - \sum_{t=1}^{T} \boldsymbol{\ell}_t(U)\right]$$

with respect to all randomized strategies defined by $p_U \in \mathcal{P}$, for some collection of distributions $\mathcal{P}$. A procedure that guarantees sublinear growth of regret with respect to any distribution $p_U \in \mathcal{P}$ and for any sequence of cost functions $\boldsymbol{\ell}_1, \ldots, \boldsymbol{\ell}_T$ will be called *consistent* with respect to $\mathcal{P}$.

Let $L_t(x) = \sum_{s=1}^{t} \boldsymbol{\ell}_s(x)$ denote the cumulative cost functions, and let $\eta > 0$ be a parameter called *the learning rate*. Fix $R(x)$ to be some convex function that defines the prior, let

$$s_t(x) = \eta L_t(x) + R(x), \qquad s_0(x) = R(x) \tag{18}$$

and define the probability distributions μ_t as in (1). It turns out that this choice of μ_t is indeed a good regret-minimization strategy, as we show next. The method is similar to the Mixture Forecaster used in the prediction context (Yamanishi, 1998; Vovk, 2001; Azoury and Warmuth, 2001; Kakade and Ng, 2005), and for a discrete set of decisions it is known

2. The index $t-1$ on μ_{t-1} reflects the fact that Y_t is chosen without the knowledge of $\boldsymbol{\ell}_t$.

as the celebrated Exponential Weights Algorithm (Vovk, 1990; Littlestone and Warmuth, 1994).

Let $D(p||q)$ stand for the Kullback-Leibler (KL) divergence between distributions p and q.

Lemma 10 *For each $t \geq 1$, let Y_t be a random variable with distribution μ_{t-1} as defined in (1). The expected regret with respect to U with distribution p_U is*

$$Reg_T(U) = \eta^{-1} \left(D(p_U||\mu_0) - D(p_U||\mu_T) \right) + \eta^{-1} \sum_{t=1}^{T} D(\mu_{t-1}||\mu_t).$$

Specializing to the case $\boldsymbol{\ell}_t : \mathcal{K} \mapsto [0,1]$ for all t,

$$Reg_T(U) \leq \eta^{-1} D(p_U||\mu_0) + T\eta/8.$$

If the KL divergence between the comparator distribution p_U and the prior μ_0 is bounded for all $p_U \in \mathcal{P}$, the second statement of the lemma yields consistency and a $\mathcal{O}(\sqrt{T})$ rate of regret growth. To bound the divergence between a continuous initial μ_0 and a point distribution at some $x^* \in \mathcal{K}$, the analysis can be carried out in two stages: comparison to a "small-covariance" Gaussian centered at x^*, followed by an observation that the loss of the "small-covariance" Gaussian strategy is not very different from the loss of the deterministic strategy x^*. The analysis can be found in (Cesa-Bianchi and Lugosi, 2006, p. 326) and gives a near-optimal $\mathcal{O}(\sqrt{T \log T})$ regret bound.

The easy proof of Lemma 10 appeared in (Narayanan and Rakhlin, 2010) and we include it in Section 6 for completeness. Having exhibited a good prediction strategy, a natural question is whether there exists a computationally efficient algorithm for producing a random draw from a distribution close to the desired mixed strategy μ_{t-1}. To this end, we use the sampling method proposed in this paper.

As a concrete example, consider linear functions $\boldsymbol{\ell}_1, \ldots, \boldsymbol{\ell}_T$ and let $R \equiv 0$. For simplicity assume boundedness $\boldsymbol{\ell}_t : \mathcal{K} \mapsto [0,1]$. In this case, we may choose $\eta = \mathcal{O}(1/\sqrt{T})$. Then

$$\beta_t \leq \exp\left\{2\eta \|\boldsymbol{\ell}_t\|_{\mathcal{K}}\right\} \leq 1 + C\eta$$

for large enough T. Further, we set $r_t = 1/d$ according to Condition 1, and the requirement (9) is seen to be satisfied for large enough T. With these choices of the parameters, the sequence of distributions $\mu_1, \ldots, \mu_t$ can be tracked with only one step of a random walk per iteration. The quality of this approximation is $\mathcal{O}\left(\eta d^3 \nu^2\right)$ at each step. Therefore, regret of the proposed random walk method is within $\mathcal{O}\left(T\eta d^3 \nu^2\right)$ from the ideal procedure of Lemma 10, as can be seen by writing

$$|\mathbb{E}\boldsymbol{\ell}_t(Y_t) - \mathbb{E}\boldsymbol{\ell}_t(X_{t-1,1})| \leq \int_{x \in \mathcal{K}} |\boldsymbol{\ell}_t(x)| \cdot |d\sigma_{t-1,1}(x) - d\mu_{t-1}(x)| \leq C\eta d^3 \nu^2 \ .$$

By choosing $\eta = \frac{1}{d^{3/2} \nu \sqrt{T}}$,

$$\mathrm{Reg}_T(U) \leq C d^{3/2} \nu D(p_U||\mu_0) \sqrt{T}. \tag{19}$$

A similar results holds for nonzero R, under the assumption that the L_2 distance between $d\mu_0(x) \propto \exp\{-R(x)\}dx$ and the uniform distribution on $\mathcal{K}$ is bounded.

We now discuss interesting parallels between the proposed randomized method and the known deterministic optimization-based regret minimization methods. First, the statement of Lemma 10 bears striking similarity to upper bounds on regret in terms of Bregman divergences for the Follow the Regularized Leader and Mirror Descent methods (Rakhlin, 2008; Beck and Teboulle, 2003), (Cesa-Bianchi and Lugosi, 2006, Therem 11.1). Yet, the randomized method operates in the (infinite-dimensional) space of distributions while the deterministic methods work directly with the set $\mathcal{K}$. Second, deterministic methods of online convex optimization face the difficulty of projections back to the set $\mathcal{K}$. This issue does not arise when dealing with distributions, but instead translates into the *difficulty of sampling*. We find these parallels between sampling and optimization intriguing. Third, a single step of the proposed random walk requires sampling from a Gaussian distribution with covariance given by the Hessian of the self-concordant barrier. This step can be implemented efficiently whenever the Hessian can be computed. The computation time exactly matches (Abernethy et al., 2008, Algorithm 2): it is the same as time spent inverting a Hessian matrix, which is $\mathcal{O}(d^3)$ or less. Finally, as already mentioned, the idea of following a time-varying distribution is inspired by the method of following the central path in the theory of interior point methods (Nesterov and Nemirovskii, 1994; Nemirovskii, 2004). Similarly to the fast convergence of the chain under the lower bound on conductance, one has fast quadratic local convergence of interior point methods. One may therefore make parallels between conductance and local curvature. A further investigation of these connections is needed, especially in view of the recent developments on positive Ricci curvature of Markov chains (Ollivier, 2009).

6. Proofs

Lemma 11 *For any t and $i \geq 0$, it holds that*

$$\left\| \frac{d\sigma_{t,i}}{d\mu_t} - 1 \right\|_t \leq \beta_t^{3/2} \left\| \frac{d\sigma_{t,i}}{d\mu_{t-1}} - 1 \right\|_{t-1} + \sqrt{\beta_t}(\beta_t - 1)$$

and, alternatively,

$$\left\| \frac{d\sigma_{t,i}}{d\mu_t} - 1 \right\|_t \leq \beta_t^{1/2} \left\| \frac{d\sigma_{t,i}}{d\mu_{t-1}} - 1 \right\|_{t-1} + \sqrt{\beta_t - 1}$$

Proof Let us use the shorthand $d\sigma = d\sigma_{t+1,i}$ and $\beta = \beta_{t+1}$. Using (3), we may write

$$\begin{aligned}
\left\| \frac{d\sigma}{d\mu_{t+1}} - 1 \right\|_{t+1} &\leq \sqrt{\beta} \left\| \frac{d\sigma}{d\mu_{t+1}} - 1 \right\|_t \\
&\leq \sqrt{\beta} \left(\left\| \frac{d\sigma}{d\mu_{t+1}} - 1 \right\|_t - \left\| \frac{d\sigma}{d\mu_t} - 1 \right\|_t + \left\| \frac{d\sigma}{d\mu_t} - 1 \right\|_t \right) .
\end{aligned}$$

By the triangle inequality,

$$\left\| \frac{d\sigma}{d\mu_{t+1}} - 1 \right\|_t - \left\| \frac{d\sigma}{d\mu_t} - 1 \right\|_t \leq \left\| \frac{d\sigma}{d\mu_{t+1}} - \frac{d\sigma}{d\mu_t} \right\|_t .$$

For any function $f : \mathcal{K} \to \mathbb{R}$, let $f^+(x) = \max(0, f(x))$ and $f^-(x) = \min(0, f(x))$. In view of (3),

$$\begin{aligned}\left\|\frac{d\sigma}{d\mu_{t+1}} - \frac{d\sigma}{d\mu_t}\right\|_t^2 &= \left\|\left(\frac{d\sigma}{d\mu_{t+1}} - \frac{d\sigma}{d\mu_t}\right)^+\right\|_t^2 + \left\|\left(\frac{d\sigma}{d\mu_{t+1}} - \frac{d\sigma}{d\mu_t}\right)^-\right\|_t^2 \\ &\le \left\|\frac{d\sigma}{d\mu_t}(\beta-1)\mathbf{1}\left[1 < \frac{d\mu_t}{d\mu_{t+1}}\right]\right\|_t^2 + \left\|\frac{d\sigma}{d\mu_t}\left(1-\frac{1}{\beta}\right)\mathbf{1}\left[1 \ge \frac{d\mu_t}{d\mu_{t+1}}\right]\right\|_t^2 \\ &\le (\beta-1)^2\left\|\frac{d\sigma}{d\mu_t}\right\|_t^2 .\end{aligned}$$

Therefore,

$$\left\|\frac{d\sigma}{d\mu_{t+1}} - 1\right\|_t - \left\|\frac{d\sigma}{d\mu_t} - 1\right\|_t \le (\beta-1)\left\|\frac{d\sigma}{d\mu_t}\right\|_t \le (\beta-1)\left(1 + \left\|\frac{d\sigma}{d\mu_t} - 1\right\|_t\right) .$$

The first statement follows by rearranging the terms.

Alternatively, we can obtain an inequality that is slightly weaker for $\beta - 1 \approx 0$ and stronger for large β by simply writing

$$\begin{aligned}\left\|\frac{d\sigma}{d\mu_{t+1}} - 1\right\|_{t+1}^2 &= \int_{\mathcal{K}}\left(\frac{d\sigma}{d\mu_{t+1}} - 1\right)^2 d\mu_{t+1} \\ &= \int_{\mathcal{K}}\frac{d\sigma^2}{d\mu_{t+1}} - 1 = \int_{\mathcal{K}}\frac{d\sigma^2}{d\mu_t^2}\frac{d\mu_t}{d\mu_{t+1}}d\mu_t - 1 .\end{aligned}$$

Using β as an upper bound on the one-sided change $\|d\mu_t/d\mu_{t+1}\|_{\mathcal{K}}$ leads to

$$\beta\int_{\mathcal{K}}\frac{d\sigma^2}{d\mu_t^2}d\mu_t - 1 = \beta\left\|\frac{d\sigma}{d\mu_t} - 1\right\|_t^2 + \beta - 1$$

and subadditivity of the square root function concludes the proof. ■

Proof [**Proof of Theorem 1**]

Given interior points x, y in $int(\mathcal{K})$, suppose p, q are the ends of the chord in $\mathcal{K}$ containing x, y and p, x, y, q lie in that order. Denote the *cross ratio* by

$$\sigma(x, y) = \frac{|x-y||p-q|}{|p-x||q-y|},$$

and for two sets S_1 and S_2 let

$$\sigma(S_1, S_2) \triangleq \inf_{x\in S_1, y\in S_2}\sigma(x, y).$$

A result due to Lovász and Vempala (2007) states the following. If S_1 and S_2 are measurable subsets of $\mathcal{K}$ and μ a probability measure supported on $\mathcal{K}$ that possesses a density whose logarithm is concave, then

$$\mu((\mathcal{K} \setminus S_1) \setminus S_2) \ge \sigma(S_1, S_2)\mu(S_1)\mu(S_2).$$

This is a non-trivial isoperimetric inequality which says that for any partition of the convex set $\mathcal{K}$ into S_1, S_2 and S_3, the "volume" of S_3 is large relative to that of S_1 and S_2 whenever S_1 and S_2 are separated. Given this isoperimetric result, to prove the theorem it only remains to show that the σ-distance can be lower bounded (up to a multiplicative constant) by the Riemannian metric ρ. The proof of this fact goes through the Hilbert (projective) metric, which is defined by

$$d_H(x,y) \triangleq \ln\left(1+\sigma(x,y)\right).$$

Further, for $x \in \mathcal{K}$ and a vector v, let

$$|v|_x \triangleq \sup_{x \pm \alpha v \in \mathcal{K}} \alpha.$$

The following two relations between the introduced notions hold. The first one (see Nesterov and Nemirovskii (Nesterov and Nemirovskii, 1994, Theorem 2.3.2 (iii))) is

$$|h|_x \le \|h\|_x \le 2(1+3\nu)|h|_x \tag{20}$$

for all $h \in \mathbb{R}^d$ and $x \in int(\mathcal{K})$, where ν is the self-concordance parameter of F. The second relation (see Nesterov and Todd (Nesterov and Todd, 2008, Lemma 3.1)) states that

$$\|x-y\|_x - \|x-y\|_x^2 \le \rho(x,y) \le -\ln(1-\|x-y\|_x). \tag{21}$$

whenever $\|x-y\|_x < 1$.

For any z on the segment $\overline{xy}$ an easy computation shows that $d_H(x,z) + d_H(z,y) = d_H(x,y)$. Therefore it suffices to prove the result infinitesimally. From (21), $\lim_{y\to x} \frac{\rho(x,y)}{\|x-y\|_x} = 1$, and a direct computation shows that

$$\lim_{y\to x} \frac{d_H(x,y)}{|x-y|_x} = \lim_{y\to x} \frac{\sigma(x,y)}{|x-y|_x} \ge 1.$$

Hence, in view of (20), the Hilbert metric and the Riemannian metric satisfy

$$\rho(x,y) \le 2(1+3\nu) d_H(x,y).$$

Using $\ln(1+x) \le x$ concludes the proof. ■

Proof [**Proof of Lemma 4**] The argument roughly follows the standard path, which is explained, for instance, in (Vempala, 2005). Let S_1 be a measurable subset of $\mathcal{K}$ such that $\mu(S_1) \le \frac{1}{2}$ and $S_2 = \mathcal{K} \setminus S_1$ be its complement. Fix a $C > 1$ and let

$$S_1' = S_1 \cap \{x \big| \mathsf{P}_x(S_2) \le 1/C\} \quad \text{and} \quad S_2' = S_2 \cap \{y \big| \mathsf{P}_y(S_1) \le 1/C\}.$$

That is, points in the set S_1' are unlikely to transition to the set S_2, and S_2' is analogously unlikely to reach S_1 in one step. By the reversibility of the chain, which is easily checked,

$$\int_{S_1} \mathsf{P}_x(S_2) d\mu(x) = \int_{S_2} \mathsf{P}_y(S_1) d\mu(y).$$

For any $x \in S_1'$ and $y \in S_2'$,

$$d_{TV}(\mathsf{P}_x, \mathsf{P}_y) = 1 - \int_{\mathcal{K}} \min\left(\frac{d\mathsf{P}_x}{d\mu}(w), \frac{d\mathsf{P}_y}{d\mu}(w)\right) d\mu(w) \geq 1 - \frac{1}{C}.$$

That is, the transition probabilities for a pair in S_1' and S_2' must be dissimilar. But Lemma 3 implies that if $\rho(x,y) \leq \frac{r}{C\sqrt{d}}$, then $d_{TV}(\mathsf{P}_x, \mathsf{P}_y) \leq 1 - \frac{1}{C}$. Therefore

$$\rho(S_1', S_2') \geq \frac{r}{C\sqrt{d}}.$$

We conclude that the sets S_1' and S_2' must be well-separated. Therefore, the isoperimetric result of Theorem 1 implies that

$$\mu((\mathcal{K} \setminus S_1') \setminus S_2') \geq \frac{\rho(S_1', S_2')}{2(1+3\nu)} \min(\mu(S_1'), \mu(S_2')) \geq \frac{r}{C\nu\sqrt{d}} \min(\mu(S_1'), \mu(S_2')).$$

First suppose $\mu(S_1') \geq (1 - \frac{1}{C})\mu(S_1)$ and $\mu(S_2') \geq (1 - \frac{1}{C})\mu(S_2)$. Then,

$$\begin{aligned}
\int_{S_1} \mathsf{P}_x(S_2) d\mu(x) &= \frac{1}{2}\int_{S_1} \mathsf{P}_x(S_2) d\mu(x) + \frac{1}{2}\int_{S_2} \mathsf{P}_x(S_1) d\mu(x) \\
&\geq \frac{1}{2C}\mu((\mathcal{K} \setminus S_1') \setminus S_2') \\
&\geq \frac{r}{2C^2\nu\sqrt{d}} \min(\mu(S_1'), \mu(S_2')) \\
&\geq \frac{1 - 1/C}{2C^2} \frac{r}{\nu\sqrt{d}} \min(\mu(S_1), \mu(S_2)),
\end{aligned}$$

proving the result. Otherwise, without loss of generality, suppose $\mu(S_1') \leq (1 - \frac{1}{C})\mu(S_1)$. Then

$$\begin{aligned}
\int_{S_1} \mathsf{P}_x(S_2) d\mu(x) &= \frac{1}{2}\int_{S_1} \mathsf{P}_x(S_2) d\mu(x) + \frac{1}{2}\int_{S_2} \mathsf{P}_x(S_1) d\mu(x) \\
&\geq \frac{1}{2}\int_{S_1 \setminus S_1'} \mathsf{P}_x(S_2) d\mu(x) \geq \frac{\mu(S_1)}{2C^2},
\end{aligned}$$

concluding the proof. ■

Proof [**Proof of Lemma 5**] The proof closely follows that in (Kalai and Vempala, 2006). By definition,

$$\|d\mu_t / d\mu_{t+1}\|_{t+1}^2 = \int_{\mathcal{K}} \left(\frac{d\mu_t}{d\mu_{t+1}}\right)^2 d\mu_{t+1} = \int_{\mathcal{K}} \frac{d\mu_t^2}{d\mu_{t+1}} = \int_{\mathcal{K}} \frac{\exp\{-2s_t\}}{Z_t^2} \cdot \frac{Z_{t+1}}{\exp\{-s_{t+1}\}}.$$

Writing out the normalization terms,

$$\|d\mu_t / d\mu_{t+1}\|_{t+1}^2 = \frac{\int_{\mathcal{K}} \exp\{-s_{t+1}\} \int_{\mathcal{K}} \exp\{s_{t+1} - 2s_t\}}{\left(\int_{\mathcal{K}} \exp\{-s_t\}\right)^2} = \frac{Y(1)Y(-1+2(1-\delta))}{Y(1-\delta)Y(1-\delta)}$$

where $Y(a) = \int_{\mathcal{K}} \exp\{-as_{t+1}\}$. As shown in (Kalai and Vempala, 2006, Lemma 3.1), the function $a^d Y(a)$ is log-concave in a, and thus

$$\frac{Y(a)Y(b)}{Y\left(\frac{a+b}{2}\right)^2} \leq \left(\frac{\left(\frac{a+b}{2}\right)^2}{ab}\right)^d .$$

Applying this inequality with $a = 1$ and $b = -1 + 2(1-\delta)$,

$$\|d\mu_t/d\mu_{t+1}\|_{t+1}^2 \leq \left(1 + \frac{\delta^2}{1-2\delta}\right)^d .$$

In particular, if $\delta \leq d^{-1/2} \leq 1/3$ (that is, $d > 8$), we obtain an upper bound of $\exp\left\{\frac{d}{d-2\sqrt{d}}\right\} \leq 21$. ■

Proof [**Proof of Lemma 10**] Observe that $D(\mu_{t-1}||\mu_t)$ can be written as

$$\int_{\mathcal{K}} d\mu_{t-1} \log \frac{q_{t-1}Z_t}{Z_{t-1}q_t} = \log \frac{Z_t}{Z_{t-1}} + \int_{\mathcal{K}} \eta \boldsymbol{\ell}_t(x) d\mu_{t-1}(x) = \log \frac{Z_t}{Z_{t-1}} + \eta \mathbb{E}\boldsymbol{\ell}_t(Y_t). \tag{22}$$

Rearranging, canceling the telescoping terms, and using the fact that $Z_0 = 1$

$$\eta \mathbb{E} \sum_{t=1}^T \boldsymbol{\ell}_t(Y_t) = \sum_{t=1}^T D(\mu_{t-1}||\mu_t) - \log Z_T.$$

Let U be a random variable with a probability distribution p_U. Then

$$-\sum_{t=1}^T \mathbb{E}\boldsymbol{\ell}_t(U) = \eta^{-1} \int_{\mathcal{K}} -\eta L_T(u) dp_U(u) = \eta^{-1} \int_{\mathcal{K}} dp_U(u) \log \frac{q_T(u)}{q_0(u)}$$

Combining,

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^T \boldsymbol{\ell}_t(Y_t) - \sum_{t=1}^T \boldsymbol{\ell}_t(U)\right] &= \eta^{-1} \int_{\mathcal{K}} dp_U(u) \log \frac{q_T(u)/Z_T}{q_0(u)} + \eta^{-1} \sum_{t=1}^T D(\mu_{t-1}||\mu_t) \\
&= \eta^{-1} \left(D(p_U||\mu_0) - D(p_U||\mu_T)\right) + \eta^{-1} \sum_{t=1}^T D(\mu_{t-1}||\mu_t).
\end{aligned}$$

Now, from Eq. (22), the KL divergence can be also written as

$$D(\mu_{t-1}||\mu_t) = \log \frac{\int_{\mathcal{K}} e^{-\eta \boldsymbol{\ell}_t(x)} q_{t-1}(x) dx}{\int_{\mathcal{K}} q_{t-1}(x) dx} + \eta \mathbb{E}\boldsymbol{\ell}_t(Y_t) = \log \mathbb{E} e^{-\eta(\boldsymbol{\ell}_t(Y_t) - \mathbb{E}\boldsymbol{\ell}_t(Y_t))}$$

By representing the divergence in this form, one can obtain upper bounds via known methods, such as *log-Sobolev inequalities* (e.g. (Boucheron et al., 2003)). In the simplest case of bounded loss, it is easy to show that $D(\mu_{t-1}||\mu_t) \leq O(\eta^2)$, and the particular constant $1/8$ can be obtained by, for instance, applying Lemma A.1 in (Cesa-Bianchi and Lugosi, 2006). This proves the second part of the lemma. ■

7. Smooth Variation of the Transition Kernel

In this section, we study the transition $x \to y$. For this purpose, it is enough to assume that x is the origin and that the Dikin ellipsoid at x is a unit Euclidean ball. This can be achieved by an affine transformation, leading to no loss of generality since the resulting statement about measures on $\mathcal{K}$ is invariant with respect to affine transformations. Hence, in what follows, for the particular x we have $< \cdot, \cdot >_x = < \cdot, \cdot >$ and $\| \cdot \|_x = \| \cdot \|$. Since x is the origin, we have $\mathbb{E}\|z\|_x^2 = r^2$ for z sampled from G_x^r. Further, without loss of generality, we may also assume $s(x) = 0$.

Proof [**Proof of Lemma 3**]

In view of the first inequality in Eq. (21),

$$\|x - y\|_x - \|x - y\|_x^2 \leq \rho(x, y) \leq \frac{r}{C\sqrt{d}}.$$

Without loss of generality, assume $\frac{r}{C\sqrt{d}} \leq \frac{1}{8}$. First, we claim that $\|x - y\|_x$ must be small. For the sake of contradiction, suppose $\|x - y\|_x > 1/2$ and consider a point y' with $\|x - y'\|_x = 1/2$ and lying on the geodesic path between x and y with respect to the Riemannian metric. Clearly, $\rho(x, y') \leq \frac{r}{C\sqrt{d}} \leq \frac{1}{8}$, yet by Eq. (21) we have $\frac{1}{4} \leq \rho(x, y')$, contradicting our assumption. Hence, $\|x - y\|_x \leq 1/2$, and, therefore, $\|x - y\|_x \leq \frac{2r}{C\sqrt{d}}$.

It remains to show that if $x, y \in \mathcal{K}$ and

$$\|x - y\|_x \leq \frac{2r}{C\sqrt{d}},$$

then

$$d_{TV}(\mathsf{P}_x, \mathsf{P}_y) = 1 - \frac{1}{C}.$$

By definition, we have that

$$1 - d_{TV}(\mathsf{P}_x, \mathsf{P}_y) = \mathbb{E}_z \left[\min \left\{ 1, \frac{G_y^r(z)}{G_x^r(z)}, \frac{G_z^r(x) \exp(s(x))}{G_x^r(z) \exp(s(z))}, \frac{G_z^r(y) \exp(s(y))}{G_x^r(z) \exp(s(z))} \right\} \right],$$

where the expectation is taken over a random point z having density G_x^r. Thus, it suffices to prove that for some $C > 1$

$$\mathbb{P} \left[\min \left\{ \frac{G_y^r(z)}{G_x^r(z)}, \frac{G_z^r(x) \exp(s(x))}{G_x^r(z) \exp(s(z))}, \frac{G_z^r(y) \exp(s(y))}{G_x^r(z) \exp(s(z))} \right\} > \frac{1}{C} \right] \geq \frac{1}{C}.$$

By our assumption, x is the origin and $D^2 F(x) = I$, the latter implying that $V(x) = 0$. Thus,

$$\frac{G_y^r(z)}{G_x^r(z)} = \exp \left\{ -\frac{d\|y - z\|_y^2}{r^2} + V(y) + \frac{d\|z\|^2}{r^2} \right\},$$

$$\frac{G_z^r(x) \exp(s(x))}{G_x^r(z) \exp(s(z))} = \exp \left\{ -\frac{d\|z\|_z^2}{r^2} + V(z) + \frac{d\|z\|^2}{r^2} + (s(x) - s(z)) \right\},$$

and

$$\frac{G_z^r(y) \exp(s(y))}{G_x^r(z) \exp(s(z))} = \exp \left\{ -\frac{d\|y - z\|_z^2}{r^2} + V(z) + \frac{d\|z\|^2}{r^2} + (s(y) - s(z)) \right\}.$$

Thus, it remains to prove that there exists a constant C such that

$$\mathbb{P}\Big[\max\Big\{ d\|y-z\|_y^2 - r^2 V(y), \quad d\|z\|_z^2 + r^2(s(z)-s(x)) - r^2 V(z), \\ d\|z-y\|_z^2 + r^2(s(z)-s(y)) - r^2 V(z)\Big\} < d\|z\|^2 + r^2 C\Big] \geq \frac{1}{C}.$$

This fact is shown in technical Lemmas 13 and 14 below. ■

In proving the technical lemmas, we will use the fact that $\|x-y\|_x \leq \frac{2r}{C\sqrt{d}}$ as shown above, and that $\|x-z\|_x$ (for z sampled from G_x^r) is likely to be bounded above by a multiple of r by straightforward concentration arguments.

Lemma 12 *There exists a constant $C > 0$ such that*

$$\mathbb{P}\left[\max\left(-V(y), -V(z)\right) < C\right] > 0.9$$

Proof Fix a constant c. First, notice that over a Euclidean ball of radius c/d around the origin, the Hessians $D^2F(u)$ are lower-bounded by a factor of $(1-c/d)^2$ from the Hessian at the origin (the identity) by (24). Hence, the determinant function can decrease from 1 by at most a constant factor. Thus $-V(u) < C'$ for some constant C' for any u with $\|x-u\|_x \leq c/d$. Now recall that y is deterministically within the $1/d$ ball, while z is in the ball of radius c/d with high probability. ■

Lemma 13 *Under step size Condition 4, for any*

$$\mathbb{P}\left[\max\Big\{s(z)-s(x), s(z)-s(y)\Big\} < C\right] > 0.32.$$

Proof Since with large enough probability $\|x-y\|_x < C'r$ and $\|x-z\|_x < C'r$, we also have $\|z-y\|_x < 2C'r$. Then, by (24), the norms at z and x are within a multiplicative constant, and thus the pairs (z,x) and (z,y) are subject to the step size choice specified in the condition. That is, there exists a g such that

$$s(z) - s(x) = s(z) - s(x) - \langle g, z-x\rangle + \langle g, z-x\rangle \leq C + \langle g, z-x\rangle$$

and similarly

$$s(z) - s(y) = s(z) - s(y) - \langle g, z-y\rangle + \langle g, z-y\rangle \leq C + \langle g, z-y\rangle$$

Then, assuming (without loss of generality) $x = 0$,

$$\mathbb{P}\left[\max\left\{\langle g, z-x\rangle, \langle g, z-y\rangle\right\} < 0\right] = \mathbb{P}\left[\langle g, z\rangle \leq \min\left\{0, \langle g, y\rangle\right\}\right].$$

Observe that $\langle g, z\rangle$ is a Gaussian random variable whose standard deviation is larger than $\|g\|\,\|y\|$. Therefore,

$$\mathbb{P}\left[\langle g, z\rangle \leq \min\left\{0, \langle g, y\rangle\right\}\right] \geq \operatorname{erfc}\left(1/\sqrt{2}\right) > 0.32,$$

where $\operatorname{erfc}(x) \triangleq \frac{2}{\sqrt{\pi}}\int_x^\infty e^{-t^2}dt$ is the usual complementary error function.

■

The following probabilistic upper bound completes the proof.

Lemma 14 *There exists a constant $C > 0$ such that*

$$\mathbb{P}\left[\max\left\{\|y-z\|_y^2, \|z\|_z^2, \|z-y\|_z^2\right\} - \|z\|^2 < \frac{Cr^2}{d}\right] > 0.9$$

Proof [**Proof of Lemma 14**] Since $\|y\| < \frac{Cr}{\sqrt{d}}$, $\|y\|_y$ and $\|y\|_z$ are less than $\frac{Cr}{\sqrt{d}}$. So it suffices to show that

$$\mathbb{P}\left[\max\left\{\|z\|_y^2 - \|z\|^2, \|z\|_z^2 - \|z\|^2, \langle y, z\rangle_y, \langle y, z\rangle_z\right\} < \frac{Cr^2}{d}\right] > 0.9$$

We proceed to do so by proving probabilistic upper bounds on each of the terms

$$\text{(a) } \|z\|_y^2 - \|z\|^2\,, \qquad \text{(b) } \|z\|_z^2 - \|z\|^2\,, \qquad \text{(c) } \langle y, z\rangle_y\,, \quad \text{and} \quad \text{(d) } \langle y, z\rangle_z$$

separately, and finally applying the union bound. We first prove an upper bound on (a) and (b). Note that $r \le \frac{1}{d}$ and thus $r^3 \le \frac{r^2}{d}$. It suffices to observe that by (24)

$$\|z\|_z^2 - \|z\|^2 \le \left(\left(\frac{1}{1-\|z\|}\right)^2 - 1\right)\|z\|^2 \le 8\|z\|^3,$$

whenever $\|z\| < 1/2$. Similarly, for $\|y\| < 1/2$,

$$\|z\|_y^2 - \|z\|^2 \le \left(\left(\frac{1}{1-\|y\|}\right)^2 - 1\right)\|z\|^2 \le 8\|z\|^3.$$

There exists a constant C such that the quantity $\|z\|^3$ is bounded by Cr^3 with probability at least 0.99.

We now turn to bounding (c) and (d). Let $[0, u]$ denote the line segment between the origin and u. By the mean-value theorem,

$$\langle y, z\rangle_y = \langle y, z\rangle + (\langle y, z\rangle_y - \langle y, z\rangle) \le \langle y, z\rangle + \sup_{y' \in [0,y]} D^3F(y')[y, y, z]$$

$$\langle y, z\rangle_z = \langle y, z\rangle + (\langle y, z\rangle_z - \langle y, z\rangle) \le \langle y, z\rangle + \sup_{z' \in [0,z]} D^3F(z')[y, z, z]$$

Observe that

$$\langle y, z\rangle \le \frac{C\|y\|\|z\|}{\sqrt{d}}$$

with probability at least 0.99 by a measure-concentration argument. Indeed, most of the vectors z are almost perpendicular to the given vector y. Now, using (23),

$$\sup_{y' \in [0,y]} D^3F(y')[y, y, z] \le \sup_{y' \in [0,y]} 2\|y\|_{y'}^2\|z\|_{y'} \le \frac{Cr^2}{d}$$

and

$$\sup_{z' \in [0,z]} D^3F(z')[y, z, z] \le \sup_{z' \in [0,z]} 2\|y\|_{z'}\|z\|_{z'}^2 \le \frac{Cr^3}{\sqrt{d}} \le \frac{Cr^2}{d}$$

with probability at least 0.99. Therefore, there exists a constant $C > 0$ such that

$$\mathbb{P}\left[\langle y, z\rangle_y < \frac{Cr^2}{d}\right] > 0.98$$

and the same statement holds for $\langle y, z\rangle_z$. We also have that

$$\mathbb{P}\left[\frac{\|y\|\|z\|}{\sqrt{d}} + \sup_{z' \in [0,z]} 2\|y\|_{z'}\|z\|_{z'}^2 \leq \frac{Cr^2}{d}\right] > 0.99$$

Therefore,

$$\mathbb{P}\left[\langle y, z\rangle_z < \frac{Cr^2}{d}\right] > 0.98.$$

■

8. Self-concordant barriers

Let $\mathcal{K}$ be a convex subset of $\mathbb{R}^d$ that is not contained in any $(d-1)$-dimensional affine subspace and $int(\mathcal{K})$ denote its interior. Following Nesterov and Nemirovskii, we call a real-valued function $F : int(\mathcal{K}) \to \mathbb{R}$, a regular self-concordant barrier if it satisfies the conditions stated below. For convenience, if $x \notin int(\mathcal{K})$, we define $F(x) = \infty$.

1. (Convex, Smooth) F is a convex thrice continuously differentiable function on $int(\mathcal{K})$.
2. (Barrier) For every sequence of points $\{x_i\} \in int(\mathcal{K})$ converging to a point $x \notin int(\mathcal{K})$, $\lim_{i\to\infty} f(x_i) = \infty$.
3. (Differential Inequalities) For all $h \in \mathbb{R}^d$ and all $x \in int(\mathcal{K})$, the following inequalities hold.

   (a) $D^2F(x)[h,h]$ is 2-Lipschitz continuous with respect to the local norm, which is equivalent to
   $$D^3F(x)[h,h,h] \leq 2(D^2F(x)[h,h])^{\frac{3}{2}}.$$

   (b) $F(x)$ is ν-Lipschitz continuous with respect to the local norm defined by F,
   $$|DF(x)[h]|^2 \leq \nu D^2F(x)[h,h].$$
   We call the smallest positive integer ν for which this holds, *the self-concordance parameter* of the barrier.

The following results can be found, for instance, in (Nesterov and Nemirovskii, 1994; Nemirovskii, 2004; Nemirovski and Todd, 2008). First,

$$|D^3F(x)[h_1, \ldots, h_k]| \leq 2\|h_1\|_x\|h_2\|_x\|h_3\|_x \ . \tag{23}$$

Second, if $\delta = \|h\|_x < 1$, then

$$(1-\delta)^2 D^2F(x) \preceq D^2F(x+h) \preceq (1-\delta)^{-2} D^2F(x) \ . \tag{24}$$

References

J. Abernethy, E. Hazan, and A. Rakhlin. Competing in the dark: An efficient algorithm for bandit linear optimization. In *Proceedings of The Twenty First Annual Conference on Learning Theory*, 2008.

J. D. Abernethy and E. Hazan. Faster convex optimization: Simulated annealing with an efficient universal barrier. In *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016*, pages 2520–2528, 2016.

K. S. Azoury and M. K. Warmuth. Relative loss bounds for on-line density estimation with the exponential family of distributions. *Machine Learning*, 43(3):211–246, June 2001.

A. Beck and M. Teboulle. Mirror descent and nonlinear projected subgradient methods for convex optimization. *Oper. Res. Lett.*, 31(3):167–175, 2003.

D. Blackwell. An analog of the minimax theorem for vector payoffs. *Pac. J. Math.*, 6:1–8, 1956.

S. Boucheron, G. Lugosi, and P. Massart. Concentration inequalities using the entropy method. *Annals of Probability*, 31:1583–1614, 2003.

A. Caponnetto and A. Rakhlin. Stability properties of empirical risk minimization over Donsker classes. *Journal of Machine Learning Research*, 6:2565–2583, 2006.

N. Cesa-Bianchi and G. Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, 2006.

N. Chopin. A sequential particle filter method for static models. *Biometrika*, 89(3):539–552, 2002.

T. Cover. Behaviour of sequential predictors of binary sequences. In *Proc. 4th Prague Conf. Inform. Theory, Statistical Decision Functions, Random Processes*, 1965.

P. Damien and S. Walker. Sampling truncated normal, beta, and gamma densities. *Journal of Computational and Graphical Statistics*, 10(2), 2001.

L. Devroye. *Non-uniform random variate generation (1986)*. Springer Verlag, 1986.

P. Diaconis. The markov chain monte carlo revolution. *Bulletin of the American Mathematical Society*, 46(2):179–205, 2009.

P. Diaconis. Some things we've learned (about Markov chain Monte Carlo). *Bernoulli*, 19 (4):1294–1305, 2013.

A. Doucet, N. De Freitas, N. Gordon, et al. *Sequential Monte Carlo methods in practice*. Springer New York, 2001.

M. Dyer, A. Frieze, and R. Kannan. A random polynomial-time algorithm for approximating the volume of convex bodies. *Journal of the ACM (JACM)*, 38(1):1–17, 1991.

C. Fraley and A. Raftery. Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association*, 97(458):611–631, 2002.

A. Frieze, R. Kannan, and N. Polson. Sampling from log-concave distributions. *The Annals of Applied Probability*, pages 812–837, 1994.

W. Gilks and P. Wild. Adaptive rejection sampling for gibbs sampling. *Applied Statistics*, pages 337–348, 1992.

J. Hannan. Approximation to Bayes risk in repeated play. *Contributions to the Theory of Games*, 3:97–139, 1957.

S. Kakade and A. Ng. Online bounds for Bayesian algorithms. In *Proceedings of Neural Information Processing Systems (NIPS 17)*, 2005.

A.T. Kalai and S. Vempala. Simulated annealing for convex optimization. *Mathematics of Operations Research*, 31(2):253–266, 2006.

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

N. Littlestone and M. K. Warmuth. The weighted majority algorithm. *Information and Computation*, 108(2):212–261, 1994.

L. Lovász. Hit-and-run mixes fast. *Mathematical Programming*, 86(3):443–461, 1999. ISSN 0025-5610.

L. Lovász and M. Simonovits. Random walks in a convex body and an improved volume algorithm. *Random Structures and Algorithms*, 4(4):359–412, 1993.

L. Lovász and S. Vempala. Simulated annealing in convex bodies and an $o^*(n^4)$ volume algorithm. *J. Comput. Syst. Sci.*, 72(2):392–417, 2006.

L. Lovász and S. Vempala. The geometry of logconcave functions and sampling algorithms. *Random Struct. Algorithms*, 30(3):307–358, 2007.

GJ McLachlan and D Peel. Finite mixture models. 2000.

S. Meyn and R. L. Tweedie. *Markov chains and stochastic stability*. Cambridge University Press, 2009.

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

H. Narayanan and A. Rakhlin. Random walk approach to regret minimization. In *Advances in Neural Information Processing Systems*, 2010.

A. S. Nemirovski and M. J. Todd. Interior-point methods for optimization. *Acta Numerica*, pages 191—234, 2008.

A. S. Nemirovskii. Interior point polynomial time methods in convex programming, 2004.

Y. E. Nesterov and A. S. Nemirovskii. *Interior Point Polynomial Algorithms in Convex Programming*. SIAM, Philadelphia, 1994.

Y.E. Nesterov and M. J. Todd. On the Riemannian geometry defined by self-concordant barriers and interior-point methods. *Foundations of Computational Mathematics*, 2(4): 333–361, 2008.

Y. Ollivier. Ricci curvature of markov chains on metric spaces. *Journal of Functional Analysis*, 256(3):810–864, 2009.

A. Rakhlin. Lecture notes on online learning, 2008. http://stat.wharton.upenn.edu/~rakhlin/papers/online_learning.pdf.

A. Rakhlin and A. Caponnetto. Stability of K-means clustering. In *Advances in Neural Information Processing Systems 19*, pages 1121–1128. MIT Press, 2006.

C. P. Robert. Simulation of truncated normal variables. *Statistics and computing*, 5(2): 121–125, 1995.

C. P. Robert and G. Casella. *Monte Carlo statistical methods*, volume 319. 2004.

S. Vempala. Geometric random walks: A survey. *In Combinatorial and computational geometry. Math. Sci. Res. Inst. Publ*, 52:577–616, 2005.

V. Vovk. Aggregating strategies. In *Proceedings of the Third Annual Workshop on Computational Learning Theory*, pages 372–383. Morgan Kaufmann, 1990.

V. Vovk. Competitive on-line statistics. *International Statistical Review*, 69:213–248, 2001.

G. Walther. Inference and modeling with log-concave distributions. *Statistical Science*, pages 319–327, 2009.

K. Yamanishi. Minimax relative loss analysis for sequential prediction algorithms using parametric hypotheses. In *COLT' 98*, pages 32–43, New York, NY, USA, 1998. ACM.