Faster Convex Optimization: Simulated Annealing with an Efficient Universal Barrier

Jacob Abernethy
Department of Computer Science
University of Michigan
jabernet@umich.edu

Elad Hazan
Department of Computer Science
Princeton University
ehazan@princeton.edu

November 6, 2015

Abstract

This paper explores a surprising equivalence between two seemingly-distinct convex optimization methods. We show that simulated annealing, a well-studied random walk algorithms, is *directly equivalent*, in a certain sense, to the central path interior point algorithm for the the entropic universal barrier function. This connection exhibits several benefits. First, we are able improve the state of the art time complexity for convex optimization under the membership oracle model. We improve the analysis of the randomized algorithm of Kalai and Vempala [7] by utilizing tools developed by Nesterov and Nemirovskii [16] that underly the central path following interior point algorithm. We are able to tighten the temperature schedule for simulated annealing which gives an improved running time, reducing by square root of the dimension in certain instances. Second, we get an efficient randomized interior point method with an efficiently computable universal barrier for any convex set described by a membership oracle. Previously, efficiently computable barriers were known only for particular convex sets.

1 Introduction

Convex optimization is by now a well established field and a cornerstone of the fields of algorithms and machine learning. Polynomial time methods for convex optimization belong to relatively few classes: the oldest and perhaps most general is the ellipsoid method with roots back to Kachiyan in the 50s [4]. Despite its generality and simplicity, the ellipsoid method is known to perform poorly in practice.

A more recent family of algorithms are the celebrated interior point methods, initially developed by Karmarkar in the context of linear programming, and generalized in the seminal work of Nesterov and Nemirovskii [16]. These methods are known to perform well in practice and come with rigorous theoretical guarantees of polynomial running time, but with a significant catch: the underlying constraints must admit an efficiently-computable *self-concordant barrier function*. Barrier functions are defined as satisfying certain differential inequality conditions that facilitate the path-following scheme developed by Nesterov and Nemirovskii [16], in particular it guarantees that the Newton step procedure maintains feasibility of the iterates. Indeed the iterative path following scheme essentially reduces the optimization problem to the construction of a barrier function, and in many nice scenarios a self-concordant barrier is easy to obtain; e.g., for polytopes the simple logarithmic barrier suffices. Yet up to the present there is no known universal *efficient* construction of a barrier that applies to any convex set. The problem is seemingly even more difficult in the *membership oracle model* where our access to $\mathcal{K}$ is given only via queries of the form "is $x \in \mathcal{K}$?".

The most recent polynomial time algorithms are random-walk based methods, originally pioneered in the work of Dyer et al. [3] and greatly advanced by Lovász and Vempala [13]. These algorithms apply in full generality of convex optimization and require only a membership oracle. The state of the art in polynomial time convex optimization is the random-walk based algorithm of simulated annealing and the specific temperature schedule analyzed in the breakthrough of Kalai and Vempala [7]. Improvements have been given in certain cases, most notably in the work of Narayanan and Rakhlin [14] where barrier functions were utilized.

In this paper we tie together two of the three known methodologies for convex optimization, give an efficiently computable universal barrier for interior point methods, and derive a faster algorithm for convex optimization in the membership oracle model. Specifically,

1. We define the **heat path** of a simulated annealing method as the (determinisitc) curve formed by the mean of the annealing distribution as the temperature parameter is continuously decreased. We show that the heat path coincides with the **central path** of an interior point algorithm with the entropic universal barrier function. This intimately ties the two major convex optimization methods together and shows they are approximately equivalent over *any* convex set.

   We further enhance this connection by showing that the central path following interior point method applied with the universal entropic barrier is a first-order approximation of simulated annealing.

2. Using the connection above, we give an efficient randomized interior point method with an efficiently computable universal barrier for any convex set described by a membership oracle. Previously, efficiently computable barriers were known only for particular convex sets.

3. We give a new temperature schedule for simulated annealing inspired by interior point methods. This gives rise to an algorithm for general convex optimization with running time of $\tilde{O}(\sqrt{\nu}n^4)$, where ν is the self-concordance parameter of the entropic barrier. The previous state of the

art is $\tilde{O}(n^{4.5})$ by [7]. We note that our algorithm does not need explicit access to the entropic barrier, it only appears in the analysis of the temperature schedule.

It was recently shown by Bubeck and Eldan [2] that the entropic barrier satisfies all require self-concordance properties and that the associated barrier parameter satisfies $\nu \leq n(1 + o(1))$, although this is generally not tight. Our algorithm improves the previous annealing run time by a factor of $\tilde{O}(\sqrt{\frac{n}{\nu}})$ which in many cases is $o(1)$. For example, in the case of semi-definite programming over matrices in $\mathbb{R}^{n\times n}$, the entropic barrier is identically the standard log-determinant barrier [5], exhibiting a parameter $\nu = n$, rather than n^2, which an improvement of n compared to the state-of-the-art. More details are given in section 4.4.

The Problem of Convex Optimization For the remainder of the paper, we will be considering the following algorithmic optimization problem. Assume we are given access to an arbitrary bounded convex set $\mathcal{K} \subset \mathbb{R}^n$, and we shall assume without loss of generality that $\mathcal{K}$ lies in a 2-norm ball of radius 1. Assume we are also given as input a vector $\hat{\theta} \in \mathbb{R}^n$. Our goal is to solve the following:

$$\min_{x\in\mathcal{K}} \hat{\theta}^\top x. \tag{1}$$

We emphasize that this is, in a certain sense, the most general convex optimization problem one can pose. While the objective is linear in x, we can always reduce non-linear convex objectives to the problem (1). If we want to solve $\min_{x\in\mathcal{K}} f(x)$ for some convex $f : \mathcal{K} \to \mathbb{R}$, we can instead define a new problem as follows. Letting $\mathcal{K}' := \{(x, c) \in \mathcal{K} \times \mathbb{R} : f(x) - c \leq 0\}$, this non-linear problem is equivalent to solving the linear problem $\min_{\{(x,c)\in\mathcal{K}'\}} c$. This equivalence is true in in the membership oracle model, since a membership oracle for $\mathcal{K}$ immediately implies an efficient membership oracle for $\mathcal{K}'$.

1.1 Preliminaries

This paper ties together notions from probability theory and convex analysis, most definitions are deferred to where they are first used. We try to follow the conventions of interior point literature as in the excellent text of Nemirovski [15], and the simulated annealing and random-walk notation of [7].

For some constant C, we say a distribution P is *C-isotropic* if for any vector $v \in \mathbb{R}^d$ we have

$$\frac{1}{C}\|v\|^2 \leq \mathop{\mathbb{E}}_{X\sim P}[(v^\top X)^2] \leq C\|v\|^2.$$

Let P, P' be two distributions on $\mathbb{R}^n$ with means μ, μ', respectively. We say P is C-isotropic with respect to P' if

$$\frac{1}{C}\mathop{\mathbb{E}}_{X\sim P'}[(v^\top X)^2] \leq \mathop{\mathbb{E}}_{X\sim P}[(v^\top X)^2] \leq C\mathop{\mathbb{E}}_{X\sim P'}[(v^\top X)^2].$$

One measure of the distance between two distributions, often referred to as the ℓ_2 norm, is given by

$$\left\|\frac{\mu}{\pi}\right\| \equiv \mathop{\mathbb{E}}_{x\sim\mu}\left(\frac{\mu(x)}{\pi(x)}\right) = \int_{x\sim\mu}\left(\frac{\mu(x)}{\pi(x)}\right) d\mu(x).$$

We note that this distance is not symmetric in general.

For a differentiable convex function $f : \mathbb{R}^n \to \mathbb{R}$, the *Bregman divergence* $D_f(x, y)$ between points $x, y \in \mathrm{dom}(f)$ is the quantity

$$D_f(x, y) \equiv f(x) - f(y) - \nabla f(y)^\top (x - y).$$

Further, we can always define the *Fenchel conjugate* (or *Fenchel dual*) [17] which is the function f^* defined as

$$f^*(\theta) := \sup_{x \in \mathrm{dom}(f)} \theta^\top x - f(x). \tag{2}$$

It is easy to see that $f^*(\cdot)$ is also convex, and under weak conditions one has $f^{**} = f$. A classic duality result (see e.g. [17]) states that when f^* is smooth and strictly convex on its domain and tends to infinity at the boundary, we have a characterization of the gradients of f and f^* in terms of maximizers:

$$\nabla f^*(\theta) = \operatorname*{arg\,max}_{x \in \mathrm{dom}(f)} \theta^\top x - f(x) \quad \text{and} \quad \nabla f(x) = \operatorname*{arg\,max}_{\theta \in \mathrm{dom}(f^*)} \theta^\top x - f^*(\theta). \tag{3}$$

1.2 Structure of this paper

We start by an overview of random-walk methods for optimization in the next section, and introduce the notion of the *heat path* for simulated annealing. The following section surveys the important notions from interior point methods for optimization and the entropic barrier function. In section 4 we tie the two approaches together formally by proving that the heat path and central path are the same for the entropic barrier. We proceed to give a new temperature schedule for simulated annealing as well as prove its convergence properties. In the appendix we describe the Kalai-Vempala methodology for analyzing simulated annealing and its main components for completeness.

2 An Overview of Simulated Annealing

Consider the following distribution over the set $\mathcal{K}$ for an arbitrary input vector $\theta \in \mathbb{R}^n$.

$$P_\theta(x) := \frac{\exp(-\theta^\top x)}{\int_K \exp(-\theta^\top x')\, dx'}. \tag{4}$$

This is often referred to as the *Boltzmann distribution* and is a natural exponential family parameterized by θ. It was observed by [7] that the optimization objective (1) can be reduced to sampling from these distributions. That is, if we choose some scaling quantity $t > 0$, usually referred to as the *temperature*, then any sample X from the distribution $P_{\hat{\theta}/t}$ must be nt-optimal in expectation. More precisely, [7] show that

$$\mathop{\mathbb{E}}_{X \sim P_{\hat{\theta}/t}} [\hat{\theta}^\top X] - \min_{x \in \mathcal{K}} \hat{\theta}^\top x \leq nt. \tag{5}$$

As we show later, our connection implies an even stronger statement, replacing n above by the self-concordant parameter of the entropic barrier, as we will define in the next section equation (10).

It is quite natural that for small temperature parameter $t \in \mathbb{R}$, samples from the $P_{\frac{\theta}{t}}$ are essentially optimal – the exponential nature of the distribution will eventually concentrate all probability mass on a small neightborhood around the maximizing point $x^* \in \mathcal{K}$. The problem, of course, is that sampling from a point mass around x^* is precisely as hard as finding x^*.

This brings us to the technique of so-called *simulated annealing*, originally introduced by Kirkpatrick et al. [9] for solving generic problems of the form $\min_{x \in \mathcal{K}} f(x)$, for arbitrary (potentially non-convex) functions f. At a very high level, simulated annealing would begin by sampling from a "high-entropy" distribution (t very close to 0), and then continue by slowly "turning down the temperature" on the distribution, i.e. decreasing t, which involves sampling according to the pdf $Q_{f,t}(x) \propto \exp(-\frac{1}{t} f(x))$. The intuition behind annealing is that, as long as t'/t is a small constant, then the distributions $Q_{f,t'}$ and $Q_{f,t}$ will be "close" in the sense that a random walk starting from the initial distribution $Q_{f,t'}$ will mix quickly towards its stationary distribution $Q_{f,t}$.

Since its inception, simulated annealing is generally referred to as a heuristic for optimization, as polynomial-time guarantees have been difficult to establish. However, the seminal work of Kalai and Vempala [7] exhibited a poly-time annealing method with formal guarantees for solving linear optimization problems in the form of (1). Their technique possessed a particularly nice feature: the sampling algorithm utilizes a well-studied random walk (Markov chain) known as HitAndRun [18, 10, 13], and the execution of this Markov chain requires only access to a **membership oracle** on the set $\mathcal{K}$. That is, HitAndRun does not rely on a formal description of $\mathcal{K}$ but only the ability to (quickly) answer queries "$x \in \mathcal{K}$?" for arbitrary $x \in \mathbb{R}^d$.

Let us now describe the HitAndRun algorithm in detail. We note that this Markov chain was first introduced by Smith [18], a poly-time guarantee was given by Lovász [10] for uniform sampling, and this was generalized to arbitrary log-concave distributions by Lovász and Vempala [11]. HitAndRun requires several inputs, including: (a) the distribution parameter θ, (b) an estimate of the covariance matrix Σ of P_θ, (c) the membership oracle $\mathcal{O}_\mathcal{K}$, for $\mathcal{K}$, (d) a starting point X_0, and (e) the number of iterations N of the random walk.

Algorithm 1: HitAndRun$(\theta, \mathcal{O}_\mathcal{K}, N, \Sigma, X_0)$ for approximately sampling P_θ

1 **Inputs:** parameter vector θ, oracle $\mathcal{O}_\mathcal{K}$ for $\mathcal{K}$, covariance matrix Σ, #iterations N, initial $X_0 \in \mathcal{K}$.
2 **for** $i = 1, 2, \ldots, N$ **do**
3 Sample a random direction $u \sim N(0, \Sigma)$
4 Querying $\mathcal{O}_\mathcal{K}$, determine the line segment $R = \{X_{i-1} + \rho u \ : \ \rho \in \mathbb{R}\} \cap \mathcal{K}$
5 Sample a point X_i from R according to the distribution P_θ restricted to R
6 Return X_N

The first key fact of HitAndRun(θ) is that the stationary distribution of this Markov chain is indeed the desired P_θ; a proof can be found in [19]. The difficulty for this and many other random walk techniques is to show that the Markov chain "mixes quickly", in that the number of steps N needn't be too large as a function of n. This issue has been the subject of much research will be discussed below. Before proceeding, we note that a single step of HitAndRun can be executed quite efficiently. Sampling a random gaussian vector with covariance Σ (line 3) can be achieved by simply sampling a standard gaussian vector z and returning $\Sigma^{1/2} z$. Computing the line segment R (line 4) requires simply finding the two locations where the line $\{X_{i-1} + \rho u \ : \ \rho \in \mathbb{R}\}$ intersects with the boundary of $\mathcal{K}$, but an ϵ-approximation of these points can be found via binary search using $O\left(\log \frac{1}{\epsilon}\right)$ queries to $\mathcal{O}_\mathcal{K}$. Sampling from P_θ restricted to the line segment R can also be achieved efficiently, and we refer the reader to Vempala [19].

The analysis for simulated annealing in [7] proceeds by imagining a sequence of distributions

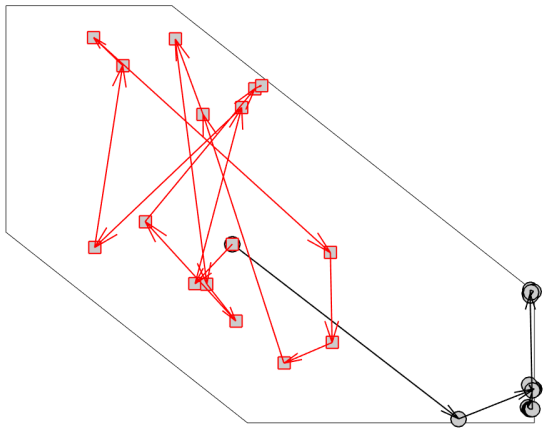

Figure 1: The progression of two Hit-and-Run random walks for a high temperature (red squares) and a low temperature (black circles). Notice that at low temperature the walk coverges very quickly to a corner of $\mathcal{K}$.

$P_{\theta_k} = P_{\hat{\theta}/t_k}$ where $t_1 = R$ is the diameter of the set $\mathcal{K}$ and $t_k := \left(1 - \frac{1}{\sqrt{n}}\right)^k$. Let $k = O(\sqrt{n} \log \frac{n}{\epsilon})$, then sampling from P_{θ_k} is enough to achieve the desired optimization guarantee. That is, via Equation 5, we see a sample from P_{θ_k} is ϵ-optimal in expectation.

To sample from P_{θ_k}, [7] construct a recursive sampling oracle using HitAndRun. The idea is that samples from $P_{\theta_{k+1}}$ can be obtained from a warm start by sampling from P_{θ_k} according to a carefully chosen temperature schedule. The details are given in Algorithm 2.

Algorithm 2: SimulatedAnnealing with HitAndRun – Kalai and Vempala [7]

Input: temperature schedule $\{t_k, k \in [T]\}$, objective $\hat{\theta}$.
Set $X_0 = 0$, $\Sigma_1 = I$, $t_1 = R$
for $k = 1, ..., T$ **do**
    $\theta_k \leftarrow \frac{\hat{\theta}}{t_k}$
    $X_k \leftarrow \text{HitAndRun}(\theta_k, \mathcal{O}_\mathcal{K}, N, \Sigma_k, X_{k-1})$
    for $j = 1, .., n$ **do**
        $Y_k^j = \text{HitAndRun}(\theta_k, \mathcal{O}_\mathcal{K}, N, \Sigma_k, Y_{k-1}^j)$
    Estimate covariance: $\Sigma_{k+1} := \text{CovMtx}(Y_1^k, \ldots, Y_n^k)$
Return X_T

The Kalai and Vempala [7] analysis leans on a number of technical but crucial facts which they prove. The temperature update schedule that they devise, namely $t_k = (1 - \frac{1}{\sqrt{n}})t_{k-1}$, is shown to satisfy these iterative rules and thus return an approximate solution.

Theorem 1 (Key result of Kalai and Vempala [7] and Lovász and Vempala [11])**.** *Fix k and consider the* HitAndRun *walk used in Algorithm 2 to compute X_k and Y_k^j for each j. Assume we choose the temperature schedule in order that successive distributions $P_{\theta_k}, P_{\theta_{k-1}}$ are close in ℓ_2:*

$$\max\left\{\left\|\frac{P_{\theta_k}}{P_{\theta_{k-1}}}\right\|_2, \left\|\frac{P_{\theta_{k-1}}}{P_{\theta_k}}\right\|_2\right\} \leq 10. \tag{6}$$

Then, as long as the warm start samples X_{k-1} and Y_{k-1}^j are (approximately) distributed according to $P_{\theta_{k-1}}$, the random walk HitAndRun *mixes to P_{θ_k} with $N = \tilde{O}(n^3)$ steps. That is, the output samples X_k and Y_k^j are distributed according to P_{θ_k} up to error $\leq \epsilon$.*

In the appendix we sketch the proof of this theorem for completeness.

Corollary 1. *The temperature schedule $t_k := (1 - 1/\sqrt{n})^k t_1$ satisfies condition (6), and thus Algorithm 2 with this schedule returns an ϵ-approximate solution in time $\tilde{O}(n^{4.5})$.*

Proof. By equation (5), to achieve ϵ error it suffice that $\frac{1}{t} \geq \frac{n}{\epsilon}$, or in other words k needs to be large enough such that $(1-\frac{1}{\sqrt{n}})^k \leq \frac{\epsilon}{n}$ for which $k = 8\sqrt{n}\log\frac{n}{\epsilon}$ suffices: $(1-\frac{1}{\sqrt{n}})^k \leq e^{-\frac{k}{2\sqrt{n}}} = e^{-4\log\frac{n}{\epsilon}} \leq \frac{\epsilon}{n}$. Hence the temperature schedule need be applied with $T = \tilde{O}(\sqrt{n})$ iterations. Each iteration requires $O(n)$ applications of HitAndRun that cost $O(n^3)$, for the total running time of $\tilde{O}(n^{4.5})$. □

In later sections we proceed to give a more refined temperature schedule that satisfies the Kalai-Vempala conditions, and thus results in a faster algorithm. Our temperature schedule is based on new observations in interior point methods, which we describe next.

2.1 The heat path for simulated annealing

Our main result follow from the observation that the path-following interior point method has an analogue in the random walk world. Simulated annealing incorporates a carefully chosen temperature schedule to reach its objective from a near-uniform distribution. We can think of all temperature schedules as performing a random process whose changing mean is a single well-defined curve. For a given convex set $\mathcal{K} \subseteq \mathbb{R}^d$ and objective $\hat{\theta}$, define the heat path as the following set of points, parametrized by the temperature $t \in [0, \infty]$ as follows:

$$\textsc{HeatPath}(t) = \mathop{\mathbb{E}}_{x \sim P_{\hat{\theta}/t}}[x]$$

We can now define the heat path as $\textsc{HeatPath} = \cup_{t \geq 0}\{\textsc{HeatPath}(t)\}$. At this point it is not yet clear why this set of points is even a continuous curve in space, let alone equivalent to an analogous notion in the interior point world. We will return to this equivalence in section 4.

3 An Overview of Interior Point Methods for Optimization

Let us now review the vast literature on Interior Point Methods (IPMs) for optimization, and in particular the use of the Iterative Newton Step technique. The first instance of polynomial time algorithms for convex optimization using interior point machinery was the linear programming algorithm of Karmarkar [8]. The pioneering book of Nesterov and Nemirovskii [16] brought up

techniques in convex analysis that allowed for polynomial time algorithms for much more general convex optimization, ideas that are reviewed in great detail and clarity in [15] .

The goal remains the same, to solve the linear optimization problem posed in Equation (1). The intuition behind IPMs is that iterative update schemes such as gradient descent for solving (1) can fail because the boundary of $\mathcal{K}$ can be difficult to manage, and "moving in the direction of descent" will fail to achieve a fast rate of convergence. Thus one needs to "smooth out" the objective with the help of an additional function. In order to produce an efficient algorithm, a well-suited type of function is known as a *self-concordant barrier* which have received a great deal of attention in the optimization literature.

A self-concordant barrier function $\varphi : \text{int}(\mathcal{K}) \to \mathbb{R}$, with *barrier parameter*, ν is a convex function satisfying two differential conditions: for any $h \in \mathbb{R}^n$ and any $x \in \mathcal{K}$,

$$\begin{aligned} \nabla^3\varphi[h,h,h] &\leq 2(\nabla^2\varphi[h,h])^{3/2}, \text{ and} \\ \nabla\varphi[h] &\leq \sqrt{\nu\nabla^2\varphi[h,h]}, \end{aligned} \tag{7}$$

in addition to the property that the barrier should approach infinity when approaching the boundary of $\mathcal{K}$. Such function possess very desirable properties from the perspective of optimization, several of which we discuss in the present section. We note an important fact: while the existence of such a function φ for general sets $\mathcal{K}$ has been given by Nesterov and Nemirovskii [16]—the *universal barrier* with parameter parameter $\nu = O(n)$, to be discussed further in Section 4.4—an efficient construction of such a function has remained elusive and was considered an important question in convex optimization. This indeed suggests that the annealing results we previously outlined are highly desirable, as HitAndRun requires only a membership oracle on $\mathcal{K}$. However, one of the central results of the present work is the equivalence between annealing and IPMs, where we show that sampling gives one implicit access to a particular barrier function. This will be discussed at length in Section 4.

Let us now assume we are given such a function φ with barrier parameter ν. A standard approach to solving (1) is to add the function $\varphi(x)$ to the primary objective, scaling the linear term by a "temperature" parameter $t > 0$:

$$\min_{x\in\mathcal{K}} \{t\hat{\theta}^\top x + \varphi(x)\}. \tag{8}$$

As the the temperature t tends to ∞ the solution of (8) will tend towards the optimal solution to 1. This result is proved for completeness in Theorem 2.

Towards developing in detail the iterative Newton algorithm, let us define the following for every positive integer k:

$$\begin{aligned} t_k &:= \left(1 + \tfrac{c}{\sqrt{\nu}}\right)^k \qquad \text{for some } c > 0, \\ \Phi_k(x) &:= t_k\hat{\theta}^\top x + \varphi(x) \\ \bar{x}_k &:= \arg\min_x \Phi_k(x) \end{aligned} \tag{9}$$

As φ is a barrier function, it is clear that $\bar{x}_k$ is in the interior of K and, in particular, $\nabla\Phi_k(\bar{x}_k) = 0 \implies \nabla\varphi(\bar{x}_k) = t_k\hat{\theta}$. It is shown in [15] (Equation 3.6) that any ν-SCB (Self-Concordant Barrier) φ satisfies $\nabla\varphi(x)^\top(y - x) \leq \nu$, whence we can bound the difference in objective value between $\bar{x}_k$ and the optimal point x^*:

$$\hat{\theta}^\top(x^* - \bar{x}_k) = \frac{\nabla\varphi(\bar{x}_k)^\top(x^* - \bar{x}_k)}{t_k} \leq \frac{\nu}{t_k}. \tag{10}$$

We see that the approximation point $\bar{x}_k$ becomes exponentially better as k increases. Indeed, setting $k = \lceil \frac{\sqrt{\nu}}{c} \log(\nu/\epsilon) \rceil$ guarantees that the error is bounded by ϵ.

The iterative Newton's method technique actually involves approximating $\bar{x}_k$ with $\hat{x}_k$, a near-optimal maximizer of Φ_k, at each iteration k. For an arbitrary $v \in \mathbb{R}^n$, $x \in \text{int}(\mathcal{K})$, and any $k \geq 1$, following [15] we define:

$$\|v\|_x := \sqrt{v^\top \nabla^2 \varphi(x) v}, \qquad \text{the "local norm" of } v \text{ w.r.t } x; \tag{11}$$

$$\|v\|_x^* := \sqrt{v^\top \nabla^{-2} \varphi(x) v}, \qquad \text{the corresponding dual norm of } v, \tag{12}$$

$$\lambda(x, t_k) := \|\nabla \Phi_k(x)\|_x^*, \qquad \text{the } \textit{Newton decrement} \text{ of } x \text{ for temperature } t_k. \tag{13}$$

Note that, for a fixed point $x \in K$, the norms $\|\cdot\|_x$ and $\|\cdot\|_x^*$ are dual to each other[1]. It will turn out that $\lambda(x, t_k)$ will be used both as a quantity in the algorithm, and as a kind of potential that we need to keep small.

In Algorithm 3 we describe the damped newton update algorithm, henceforth called IterativeNewtonStep. We note that the

Algorithm 3: IterativeNewtonStep

Input: $\hat{\theta} \in \mathbb{R}^d$, $\mathcal{K}$ and barrier function φ
Solve: $\hat{x}_0 = \arg\max_{x \in \mathcal{K}} \hat{\theta}^\top x + \varphi(x)$
for $k = 1, 2, \ldots$ **do**
  $\hat{x}_k \leftarrow \hat{x}_{k-1} - \frac{1}{1+\lambda(\hat{x}_{k-1}, t_k)} \nabla^{-2} \varphi(\hat{x}_{k-1}) \nabla \Phi_k(\hat{x}_{k-1})$

The most difficult part of the analysis is in the following two lemmas, which are crucial elements of the IterativeNewtonStep analysis. A full exposition of these results is found in the excellent survey [15]. The first lemma tells us that when we update the temperature, we don't perturb the Newton decrement too much. The second lemma establishes the *quadratic convergence* of the Newton Update for a fixed temperature.

Lemma 1. *Let c be the constant chosen in the definition (9). Let $t > 0$ be arbitrary and let $t' = t\left(1 + \frac{c}{\sqrt{\nu}}\right)$. Then for any $x \in int(\mathcal{K})$, we have $\lambda(x, t') \leq (1 + c)\lambda(x, t) + c$.*

Lemma 2. *Let k be arbitrary and assume we have some $\hat{x}_{k-1}$ such that $\lambda(\hat{x}_{k-1}, t_k)$ is finite. The Newton update $\hat{x}_k$ satisfies $\lambda(\hat{x}_k, t_k) \leq 2\lambda^2(\hat{x}_{k-1}, t_k)$.*

The previous two lemmas can be combined to show that the following invariant is maintained. Neither the constant bound of $1/3$ on the Newton decrement nor the choice of $c = 1/20$ are particularly fundamental; they are convenient for the analysis but alternative choices are possible.

Lemma 3. *Assume we choose $c = 1/20$ for the parameter in (9). Then for all k we have $\lambda(\hat{x}_k, t_k) < \frac{1}{3}$.*

[1]Technically, for $\|\cdot\|_x$ and its dual to be a norm, we need $\nabla^2 \varphi$ to be positive definite and φ to be strictly convex. One can verify this is the case for bounded sets, which is the focus of this paper.

Proof. We give a simple proof by induction. The base case is satisfied since we assume that $\lambda(\hat{x}_0, t_0) = 0$, as $t_0 = 1$.[2] For the inductive step, assume $\lambda(\hat{x}_{k-1}, t_{k-1}) < 1/3$. Then

$$\begin{aligned}\lambda(\hat{x}_k, t_k) &\leq 2\lambda^2(\hat{x}_{k-1}, t_k) \\ &\leq 2((1+c)\lambda(\hat{x}_{k-1}, t_{k-1}) + c)^2 < 2(0.4)^2 < 1/3.\end{aligned}$$

The first inequality follows by Lemma 2 and the second by Lemma 1, hence we are done. □

Theorem 2. *Let x^* be a solution to the objective* (1)*. For every k, $\hat{x}_k$ is an ϵ_k-approximate solution to* (1)*, where $\epsilon_k = \frac{\nu+\sqrt{\nu}/4}{t_k}$. In particular, for any $\epsilon > 0$, as long as $k > \frac{\sqrt{\nu}}{c}\log(2\nu/\epsilon)$ then $\hat{x}_k$ is an ϵ-approximation solution.*

Proof. Let k be arbitrary. Then,

$$\begin{aligned}\hat{\theta}^\top(\hat{x}_k - x^*) &= \hat{\theta}^\top(\bar{x}_k - x^*) + \hat{\theta}^\top(\hat{x}_k - \bar{x}_k) \\ \text{(By (10))} \quad &\leq \frac{\nu}{t_k} + \hat{\theta}^\top(\hat{x}_k - \bar{x}_k) \\ \text{(Hölder's Inequality)} \quad &\leq \frac{\nu}{t_k} + \|\hat{\theta}\|^*_{\bar{x}_k}\|\bar{x}_k - \hat{x}_k\|_{\bar{x}_k} \\ \text{([15] Eqn. 2.20)} \quad &\leq \frac{\nu}{t_k} + \|\hat{\theta}\|^*_{\bar{x}_k}\frac{\lambda(\hat{x}_k,t_k)}{1-\lambda(\hat{x}_k,t_k)} \\ \text{(Applying Lemma 3 and (14))} \quad &\leq \frac{\nu}{t_k} + \left\|\frac{\nabla\varphi(\bar{x}_k)}{t_k}\right\|^*_{\bar{x}_k}\frac{1}{4} \quad \leq \quad \frac{\nu+\sqrt{\nu}/4}{t_k}\end{aligned}$$

The last equation utilizes a fact that derives immediately from the definition (7), namely

$$\|\nabla\varphi^*(x)\|^*_x = \|\nabla\varphi^*(x)\|_{\theta(x)} \leq \sqrt{\nu} \tag{14}$$

holds for any SCBF φ with parameter ν, and for any $x \in \mathcal{K}$. □

We proceed to give a specific barrier function that applies to any convex set and gives rise to an efficient algorithm.

4 The Equivalence of IterativeNewton and SimulatedAnnealing

We now show that the previous two techniques, Iterative Newton's Method and Simulated Annealing, are in a certain sense two sides of the same coin. In particular, with the appropriate choice of barrier function φ the task of computing the sequence of Newton iterates $\hat{x}_1, \hat{x}_2, \ldots$ may be viewed precisely as estimating the means for each of the distributions $P_{\theta_1}, P_{\theta_2}, \ldots$. This correspondence helps to unify two very popular optimization methods, but also provides three additional novel results:

1. We show that the heat path for simulated annealing is equivalent to the central path with the entropic barrier.

[2] As stated, Algorithm 3 requires finding the minimizer of $\varphi(\cdot)$ on $\mathcal{K}$, but this is not strictly necessary. The convergence rate can be established as long as a "reasonable" initial point $\hat{x}_0$ can be computed—e.g. it suffices that $\lambda(\hat{x}_0, 1) < 1/2$.

2. We show that the running time of Simulated Annealing can be improved to $O(n^4\sqrt{\nu})$ from the previous best of $O(n^{4.5})$. In the most general case we know that $\nu = O(n)$, but there are many convex sets in which the parameter ν is significantly smaller. Notably such is the case for the positive-semi-definite cone, which is the basis of semi-definite programming and a cornerstone of many approximation algorithms. Further examples are surveyed in section 4.4.

3. We show that Iterative Newton's Method, which previously required a provided barrier function on the set $\mathcal{K}$, can now be executed efficiently where the only access to $\mathcal{K}$ is through a membership oracle. This method relies heavily on previously-developed sampling techniques [7]. Discussion is deferred to Appendix C.

4.1 The Duality of Optimization and Sampling

We begin by rewriting our Boltzmann distribution for θ in exponential family form,

$$P_\theta(x) := \exp(-\theta^\top x - A(\theta)) \quad \text{where} \quad A(\theta) := \log \textstyle\int_K \exp(-\theta^\top x')dx'. \tag{15}$$

The function $A(\cdot)$ is known as the *log partition function* of the exponential family, and it has several very natural properties in terms of the mean and variance of P_θ:

$$\nabla A(\theta) = -\mathop{\mathbb{E}}_{X\sim P_\theta}[X] \tag{16}$$

$$\nabla^2 A(\theta) = \mathop{\mathbb{E}}_{X\sim P_\theta}[(X - \mathbb{E}\,X)(X - \mathbb{E}\,X)^\top]. \tag{17}$$

We can also appeal to Convex (Fenchel) Duality [17] to obtain the conjugate

$$A^*(x) := \sup_\theta \theta^\top x - A(\theta). \tag{18}$$

It is easy to establish that A^* is smooth and strictly convex. The domain of $A^*(\cdot)$ is precisely the space of gradients of $A(\cdot)$, and it is straightforward to show that this is the set $\text{int}(-\mathcal{K})$, the interior of the reflection of $\mathcal{K}$ about the origin. However we want a function defined on (the interior of) $\mathcal{K}$, not its reflection, so let us define a new function $A^*_-(x) := A^*(-x)$ whose domain is $\text{int}(\mathcal{K})$. We now present a recent result of Bubeck and Eldan [2].

Theorem 3 ([2])**.** *The function A^*_- is a ν-self-concordant barrier function on $\mathcal{K}$ with $\nu \le n(1 + o(1))$.*

One of the significant drawbacks of barrier/Newton techniques is the need for a readily-available self-concordant barrier function. In their early work on interior point methods, Nesterov and Nemirovskii [16] provided such a function, often referred to as the "universal barrier", yet the actual construction was given implicitly without oracle access to function values or derivatives. Bubeck and Eldan [2] refer to function $A^*_-(\cdot)$ as the *entropic barrier*, a term we will also use, as it relates to a notion of differential entropy of the exponential family of distributions.

It is important to note that, when our set of interest is a cone K, the entropic barrier on K corresponds exactly to the Fenchel dual of the universal barrier on the dual cone K^* [6], which immediately establishes the self-concordance. Indeed, many beautiful properties of the entropic barrier on cones have been developed, and for a number of special cases $A^*_-(\cdot)$ corresponds exactly

to the canonical barrier used in practice; e.g. $A^*_-(\cdot)$ on the PSD cone corresponds to the log-determinant barrier. In many such cases one obtains a much smaller barrier parameter ν than the $n(1+o(1))$ bound. We defer a complete discussion to Section 4.4.

In order to utilize $A^*_-(\cdot)$ as a barrier function as in (8) we must be able to approximately solve objectives of the form $\min_{x\in\mathcal{K}}\{\theta^\top x + A^*_-(x)\}$. One of the key observations of the paper, given in the following Proposition, is that solving this objective correponds to computing the mean of the distribution P_θ.

Proposition 1. *Let $\theta \in \mathbb{R}^n$ be arbitrary, and let P_θ be defined as in* (15)*. Then we have*

$$\mathop{\mathbb{E}}_{X\sim P_\theta}[X] = \mathop{\arg\min}_{x\in\mathrm{int}(\mathcal{K})}\left\{\theta^\top x + A^*_-(x)\right\}. \tag{19}$$

Proof. Let θ be an arbitrary input vector. As we mentioned in (3), standard Fechel duality theory gives us

$$\nabla A(\theta) = \mathop{\arg\max}_{x\in\mathrm{dom}(A^*)}\left\{\theta^\top x - A^*(x)\right\} = \mathop{\arg\min}_{x\in\mathrm{dom}(A^*)}\left\{-\theta^\top x + A^*(x)\right\}.$$

It can be verified that the domain of A^* is precisely the interior of $-\mathcal{K}$, the reflection of $\mathcal{K}$. Thus we have

$$\nabla A(\theta) = \mathop{\arg\min}_{x\in\mathrm{int}(-\mathcal{K})}\left\{-\theta^\top x + A^*(x)\right\} = -\left(\mathop{\arg\min}_{x\in\mathrm{int}(\mathcal{K})}\left\{\theta^\top x + A^*_-(x)\right\}\right).$$

In addition, we noted in (15) that $\nabla A(\theta) = -\mathbb{E}_{X\sim P_\theta}[X]$. Combining the latter two facts gives the result. □

We now have a direct connection between sampling methods and barrier optimization. For the remainder of this section, we shall assume that our chosen $\varphi(\cdot)$ is the entropic barrier $A^*(\cdot)$, and the quantities $\Phi_k(\cdot), \|\cdot\|_x, \lambda(\cdot,\cdot)$ are defined accordingly. We shall also use the notation $x(\theta) := \mathbb{E}_{X\sim P_\theta}[X] = -\nabla A(\theta)$.

Lemma 4. *Let θ, θ' be such that $\|x(\theta') - x(\theta)\|_{x(\theta)} \le \frac{1}{2}$. Then we have*

$$D_{A^*_-}(x(\theta'), x(\theta)) = KL(P'_\theta, P_\theta) = D_A(\theta,\theta') \le 2\|x(\theta') - x(\theta)\|^2_{x(\theta)} \tag{20}$$

Proof. The duality relationship of the Bregman divergence, and its equivalence to Kullback-Leibler divergence, is classical and can be found in, e.g., [20] equation (5.10) The final inequality follow as a direct consequence of [15], Equation 2.4. □

4.2 Equivalence of the heat path and central path

The most appealing observation on the equivalence between random walk optimization and interior point methods is the following geometric equivalence of curves. For a given convex set $\mathcal{K} \subseteq \mathbb{R}^d$ and objective $\hat{\theta}$, define the heat path as the following set of points:

$$\textsc{HeatPath} = \bigcup_{t\ge 0}\{\textsc{HeatPath}(t)\} = \bigcup_{t\ge 0}\{\mathop{\mathbb{E}}_{x\sim P_{\frac{\hat{\theta}}{t}}}[x]\}$$

To see that this set of points is a continuous curve in space, consider the central path w.r.t. barrier function $\varphi(x)$:

$$\textsc{CentralPath}(\varphi) = \bigcup_{t\ge 0}\{\textsc{CentralPath}(t,\varphi)\} = \bigcup_{t\ge 0}\{\mathop{\arg\min}_{x\in\mathcal{K}}\{t\hat{\theta}^\top x + \varphi(x)\}$$

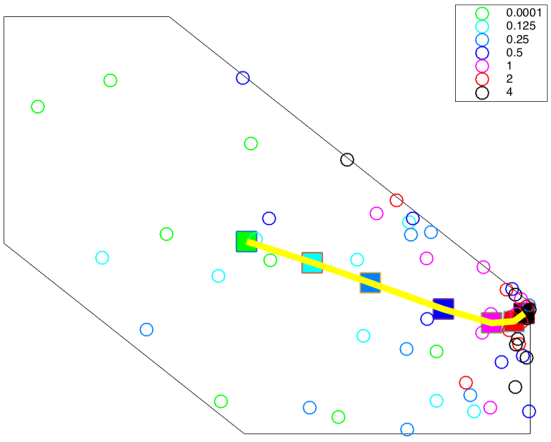

Figure 2: For a set of seven different temperatures t, we used Hit-and-Run to generate and scatter plot several samples from $P_{\theta/t}$ using colored circles. We also computed the true means for each distribution, plotted with squares, giving a curve representing the HeatPath across the seven temperatures. Of course via Corollary 2 this corresponds exactly to the CentralPath for the entropic barrier.

It is well known that the central path is a continuous curve in space for any self-concordant barrier function ϕ. We now have the following immediate corollary of Proposition 1:

Corollary 2. *The central path corresponding to the self-concordant barrier A^* over any set $\mathcal{K}$ is equivalent to the heat path over the same set, i.e.*

$$\textsc{HeatPath} \equiv \textsc{CentralPath}(A^*)$$

This mathematical equivalence is demonstrated in figure 2 generated by simulation over a polytope.

4.3 IPM techniques for sampling and the new schedule

We now prove our main theorem, formally stated as:

Theorem 4. *The temperature schedule of $\theta_1 = R$ where $R = \text{diam}(\mathcal{K})$ and $\theta_k := \left(1 - \frac{1}{4\sqrt{\nu}}\right)\theta_{k-1}$, for ν being the self-concordance parameter of the entropic barrier for the set $\mathcal{K}$, satisfies condition (6) of theorem 1. Therefore algorithm 2 with this schedule returns an ϵ-approximate solution in time $\tilde{O}(\sqrt{\nu}n^4)$.*

Condition (6) is formally proved in the following lemma, which crucially uses the interior point methodology, namely Lemma 3.

Lemma 5. *Consider distributions P_θ and $P_{\theta'}$ where $\theta' = (1+\gamma)\theta$ for $\gamma < \frac{1}{6\sqrt{\nu}}$. Then we have the following bound on the ℓ_2 distance between distributions:*

$$\max\left\{\left\|\frac{P_\theta}{P_{(1+\gamma)\theta}}\right\|_2, \left\|\frac{P_{(1+\gamma)\theta}}{P_\theta}\right\|_2\right\} \leq 10$$

Proof. We first show by elementary linear algebra that

$$\|P_\theta / P_{(1-\gamma)\theta}\| = \|P_\theta / P_{(1+\gamma)\theta}\| = \exp(D_A((1+\gamma)\theta, \theta) + D_A((1-\gamma)\theta, \theta)).$$

Let us consider the log of the 2-norm:

$$\begin{aligned}
\log \|P_\theta / P_{(1+\gamma)\theta}\| &= \log \int_{\mathcal{K}} \frac{\exp(-\theta^\top x - A(\theta))}{\exp(-(1+\gamma)\theta^\top x - A((1+\gamma)\theta))} dP_\theta \\
&= \log \int_{\mathcal{K}} \exp(\gamma\theta^\top x - A(\theta) + A((1+\gamma)\theta)) dP_\theta \\
&= \log \int_{\mathcal{K}} \exp(\gamma\theta^\top x - A(\theta) + A((1+\gamma)\theta)) \exp(-\theta^\top x - A(\theta)) dx \\
&= A((1+\gamma)\theta) - 2A(\theta) + \log \int_{\mathcal{K}} \exp(-(1-\gamma)\theta^\top x) dx \\
&= A((1+\gamma)\theta) - 2A(\theta) + A((1-\gamma)\theta) \\
&= D_A((1+\gamma)\theta, \theta) + D_A((1-\gamma)\theta, \theta).
\end{aligned}$$

Replacing γ by $-\gamma$, we get a completely symmetrical expression. Next, we observe that

$$\|P_{\theta(1+\gamma)} / P_\theta\| = \|P_{\tilde{\theta}} / P_{\tilde{\theta}(1-\tilde{\gamma})}\|$$

where $\tilde{\theta} = \theta(1+\gamma)$ and $1 - \tilde{\gamma} = \frac{1}{1+\gamma} = 1 - \frac{\gamma}{1+\gamma}$, thus $\tilde{\gamma} \in \gamma \times [1, 2]$. By this observation, both sides of the lemma follow if we prove an upper bound

$$\left\|\frac{P_\theta}{P_{(1+\gamma)\theta}}\right\|_2 \leq 10 \quad \text{for} \quad \gamma < \frac{1}{6\sqrt{\nu}} \times 2 = \frac{1}{3\sqrt{\nu}}$$

Lemma 1 implies $\lambda(x(\theta), 1+\gamma) \leq (1+c)\lambda(x(\theta), 1) + c = c \leq \frac{1}{4}$. Applying Lemma 4,

$$\begin{aligned}
D_A((1+\gamma)\theta, \theta) &\leq 2\|x(\theta) - x((1+\gamma)\theta)\|^2_{x((1+\gamma)\theta)} && \text{Lemma 4} \\
&\leq 2\left(\tfrac{\lambda(x(\theta), 1+\gamma)}{1-\lambda(x(\theta), 1+\gamma)}\right)^2 && \text{(2.28) in [15]} \\
&\leq 2\left(\tfrac{c}{1-c}\right)^2 < 2\tfrac{(1/3)^2}{(2/3)^2} = \tfrac{1}{2} && \text{Lemma 1}
\end{aligned}$$

Notice that to apply Lemma 4, we need the point $x((1+\gamma)\theta)$ to lie in the Dikin ellipsoid of $x(\theta)$, which is exactly whats proved in the last two lines of the above proof.

The bound on $D_A((1-\gamma)\theta, \theta)$ follows in precisely the same fashion, by similar change of variables as before (again, the condition for applying Lemma 4 is proven in the last few lines of the equations

below):

$$\begin{aligned} D_A((1-\gamma)\theta,\theta) &= D_A(\tilde{\theta},\tilde{\theta}(1+\tilde{\gamma})) \\ &\le 2\|x(\tilde{\theta}) - x((1+\tilde{\gamma})\tilde{\theta})\|^2_{x(\tilde{\theta})} && \text{Lemma 4} \\ &\le 2\left(\tfrac{\lambda(x(\tilde{\theta}),1+\tilde{\gamma})}{1-\lambda(x(\tilde{\theta}),1+\hat{\gamma})}\right)^2 && \text{(2.27) in [15]} \\ &\le 2\left(\tfrac{c}{1-c}\right)^2 < 2\tfrac{(1/3)^2}{(2/3)^2} = \tfrac{1}{2} && \text{Lemma 1} \end{aligned}$$

It follows that $\|P_\theta / P_{(1+\gamma)\theta}\| \le e^{D_A((1+\gamma)\theta,\theta)+D_A((1-\gamma)\theta,\theta)} \le e^{\frac{1}{2}+\frac{1}{2}} \le 10$. □

4.4 Some history on the entropic barrier and the universal barrier for cones

Let K be a cone in $\mathbb{R}^n$ and let $K^* = \{\theta : \theta^\top x \ge 0 \ \forall x \in K\}$ be its dual cone. We note that a cone K is *homogeneous* if its automorphism group is transitive; that is, for every $x, y \in K$ there is an automorphism $B : K \to K$ such that $Bx = y$. Homogeneous cones are very rare, but one notable example is the PD cone (matrices with all positive eigenvalues). Given any point $x \in K$, we can define a truncated region of K^* as the set $K^*(x) := \{y \in K^* : x^\top y \le 1\}$. Nesterov and Nemirovskii [16] defined the first generic self-concordant barrier function, known as the *universal barrier* in terms of these regions. Namely, they show that the function

$$u_K(x) := \log(\mathrm{vol}(K^*(x)))$$

is a self concordant barrier function with an $O(n)$ parameter.

There is an alternative characterization of the universal barrier in terms of the larg partition function. Let $F_K(x) := \log \int_{K^*} \exp(\theta^\top x) d\theta$ and equivalently let $F_{K^*}(\theta) := \log \int_K \exp(\theta^\top x) dx$. It was shown by Güler [5] that

$$F_K(x) = u_K(x) + n!,$$

that is, the universal barrier corresponds exactly to a log-partition function but defined on *the dual cone* K^*, modulo a simple additive constant. We note that this is not the entropic barrier construction we have here, as our function of interest is $A^*_-(\cdot) \equiv F^*_{K^*}(\cdot)$ (the Fenchel conjugate of $F_{K^*}(x)$), and not $F_K(x)$. However, it was also shown by Güler [5] that, when K is a homogeneous cone, we have the identity $F_K(\cdot) \equiv F^*_{K^*}(\cdot)$; in other words, the universal barrier and the entropic barrier are equivalent for homogeneous cones.

It is worth noting that, following the connection of Güler [5], $A^*_-(\cdot)$ is (up to additive constant) the Fenchel conjugate of the universal barrier u_{K^*} for K^*. It was shown by Nesterov and Nemirovskii [16] (Theorem 2.4.1) that Fenchel conjugation preserves all required self concordance properties and in particular if g is a ν-self-concordant barrier for a cone K, then g^* will be a self-concordant barrier for K^* with the same parameter ν. With this observation it follows immediately that the entropic barrier $F^*_{K^*}(\cdot)$ on K is an $O(n)$-self-concordant barrier. Bubeck and Eldan [2] took this statement further, proving that $F^*_{K^*}(\cdot)$ enjoys an essentially optimal self-concordance parameter $\nu = n(1 + o(1))$.

It is important to note that the assumption that the set of interest is a cone is, roughly speaking, without loss of generality. Given any convex set $U \subset \mathbb{R}^n$ we have the *fitted cone* $K(U) := \{\alpha(x, 1) : x \in U, \alpha \ge 0\} \subseteq \mathbb{R}^{n+1}$. Hence once can always work with the barrier function $F^*_{K(U)^*}(\cdot)$ on $K(U)$,

and take its restriction to the set $U \times \{1\} \subset K(U)$ to obtain a barrier on U (affine restriction preserves the barrier properties).

We conclude by summarizing several results in Güler [5] regarding the entropic barrier for various cones, as well as the associated barrier parameter of each. In these canonical cases the entropic barrier corresponds exactly to the "typical" barrier, up to additive and multiplicative constants. We use the notation $f(\cdot) \cong g(\cdot)$ to denote that f and g are identical up to additive constants.

1. Assume $K := \mathbb{R}^n_+$ the nonnegative orthant. This is a homogeneous cone and we have $F_K(x) \cong F^*_{K^*}(x) \cong -\sum_{i=1}^n \log x_i$. This is the optimal barrier for K and the barrier parameter is $\nu = n$.

2. Assume $K := \{x \in \mathbb{R}^n : x_1^2 + \ldots + x_{n-1}^2 \leq x_n^2\}$ be the *Lorentz cone*. K is a homogeneous self-dual cone. K can also be described by the fitted cone of the radius-1 $L2$ ball, so we may parameterize elements of K as $\alpha(x, 1)$ where $\alpha \geq 0$ and x is vector in $\mathbb{R}^{n-1}$ with $L2$ norm bounded by 1. Then $F_K(\alpha(x,1)) \cong F^*_{K^*}(\alpha(x,1)) \cong -\frac{n+1}{2}\log(\alpha^2 - x^2)$. This barrier has parameter $\nu = n + 1$ which is indeed not optimal, one has the optimal barrer $-\log(\alpha^2 - x^2)$ which has parameter $\nu = 2$, but this is simply a scaled version of the entropic barrier.

3. The PSD cone K of positive semi-definite matrices, i.e. symmetric matrices with non-negative eigenvalues, is a homogeneous self-dual cone. The entropic barrier is $F_K(x) \cong F^*_{K^*}(x) \cong -\frac{n+1}{2}\log\det(x)$ and exhibits a parameter of $\nu = \frac{n(n+1)}{2}$ which is multiplicatively $\frac{n+1}{2}$ worse than the optimal barrier, the log-determined $-\log\det(x)$. However, as can be seen this barrier is quite simply a scaled version of the entropic barrier.

Bibliography

[1] R. Adamczak, A. E. Litvak, A. Pajor, and N. Tomczak-Jaegermann. Quantitative estimates of the convergence of the empirical covariance matrix in log-concave ensembles. *Journal of the American Mathematical Society*, 23:535–561, apr 2010. doi: 10.1090/S0894-0347-09-00650-X.

[2] S. Bubeck and R. Eldan. The entropic barrier: a simple and optimal universal self-concordant barrier. *arXiv preprint arXiv:1412.1587*, 2014.

[3] M. Dyer, A. Frieze, and R. Kannan. A random polynomial-time algorithm for approximating the volume of convex bodies. *J. ACM*, 38(1):1–17, Jan. 1991. ISSN 0004-5411.

[4] M. Grötschel, L. Lovász, and A. Schrijver. *Geometric algorithms and combinatorial optimization*. Algorithms and combinatorics. Springer-Verlag, 1993. ISBN 9780387136240. URL `https://books.google.com/books?id=agLvAAAAMAAJ`.

[5] O. Güler. Barrier functions in interior point methods. *Mathematics of Operations Research*, 21(4):860–885, 1996.

[6] O. Güler. On the self-concordance of the universal barrier function. *SIAM Journal on Optimization*, 7(2):295–303, 1997.

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

[8] N. Karmarkar. A new polynomial-time algorithm for linear programming. In *Proceedings of the Sixteenth Annual ACM Symposium on Theory of Computing*, STOC '84, pages 302–311, 1984.

[9] S. Kirkpatrick, M. Vecchi, et al. Optimization by simmulated annealing. *science*, 220(4598): 671–680, 1983.

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

[11] L. Lovász and S. Vempala. The geometry of logconcave functions and an $o(n^3)$ sampling algorithm. Technical report, Technical Report MSR-TR-2003-04, Microsoft Research, 2003.

[12] L. Lovász and S. Vempala. Hit-and-run from a corner. *SIAM J. Comput.*, 35(4):985–1005, Apr. 2006. ISSN 0097-5397.

[13] L. Lovász and S. Vempala. Fast algorithms for logconcave functions: Sampling, rounding, integration and optimization. In *Foundations of Computer Science, 2006. FOCS'06. 47th Annual IEEE Symposium on*, pages 57–68. IEEE, 2006.

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

[15] A. Nemirovski. Interior point polynomial time methods in convex programming. *Lecture Notes–Faculty of Industrial Engineering and Management, Technion–Israel Institute of Technology, Technion City, Haifa*, 32000, 1996.

[16] Y. Nesterov and A. Nemirovskii. *Interior-point Polynomial Algorithms in Convex Programming*, volume 13. SIAM, 1994.

[17] R. T. Rockafellar. *Convex analysis*. Princeton university press, 1970.

[18] R. L. Smith. Efficient monte carlo procedures for generating points uniformly distributed over bounded regions. *Operations Research*, 32(6):1296–1308, 1984.

[19] S. Vempala. Geometric random walks: a survey. *MSRI volume on Combinatorial and Computational Geometry*, 2005.

[20] M. J. Wainwright and M. I. Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1(1-2):1–305, 2008.

A An Explanation of Newton's Method via Reweighting

Proposition 1 brings out a strong connection between interior point techniques and the ability to sample from Boltzmann distributions. But with this stochastic viewpoint, it is not immediately clear why Newton's method is an appropriate iterative update scheme for optimization. We now provide some evidence along these lines.

Assuming we have already computed (an approximation of) $x(\theta)$, and our distribution parameter is updated to a "nearby" θ', our goal is now to compute the new mean $x(\theta')$.

$$-x(\theta') = \int_{\mathcal{K}} x\, dP_{\theta'} = \int_{\mathcal{K}} x\, \frac{dP_{\theta'}(x)}{dP_\theta(x)} dP_\theta \quad = \quad \mathop{\mathbb{E}}_{X\sim P_\theta}\Big[X \exp(X^\top(\theta-\theta') + A(\theta') - A(\theta))\Big]$$

Think of the last term as the reweighting factor. Now we are going to rewrite $A(\theta) - A(\theta') = -\nabla A(\theta)(\theta'-\theta) - D_A(\theta',\theta) = x(\theta)^\top(\theta'-\theta) - \mathrm{KL}(P_\theta, P_{\theta'})$. We shall use the following approximation of the exponential: $\exp(z) \approx 1+z$ for small values of z. Proceeding,

$$\begin{aligned}
-x(\theta') &= \mathop{\mathbb{E}}_{X\sim P_\theta}\Big[X \exp(X^\top(\theta-\theta') - x(\theta)^\top(\theta'-\theta) + \mathrm{KL}(P_\theta, P_{\theta'}))\Big] \\
&= e^{\mathrm{KL}(P_\theta,P_{\theta'})} \mathop{\mathbb{E}}_{X\sim P_\theta}\Big[X \exp((X-x(\theta))^\top(\theta'-\theta))\Big] \\
&\approx e^{\mathrm{KL}(P_\theta,P_{\theta'})} \mathop{\mathbb{E}}_{X\sim P_\theta}\Big[X (1+(X-x(\theta))^\top(\theta'-\theta))\Big] \\
&= e^{\mathrm{KL}(P_\theta,P_{\theta'})}(-x(\theta) + \Sigma_\theta(\theta'-\theta)).
\end{aligned}$$

Duality theory tells us that $\Sigma_\theta = \nabla^2 A(\theta) = \nabla^{-2} A^*(x(\theta))$ and $\theta'-\theta$ is precisely the gradient of the objective $\theta'^\top x - A^*(x)$ at the point $x(\theta)$. The $e^{\mathrm{KL}(P_\theta,P_{\theta'})}$ term is somewhat mysterious, but it can be interpreted as something of a "damping" factor akin to the Newton decrement damping of the the Newton update.

B Proof structure of the Kalai-Vempala theorem

We hereby sketch the structure of the proof of theorem 1 for completeness. Recall the statement of the theorem:

Algorithm 2 with a temperature schedule that satisfies the following condition:

The successive distributions are not "too far" in total variational distance. That is, for every j,

$$\max\left\{\left\|\frac{P_{\theta_j}}{P_{\theta_{j-1}}}\right\|_2, \left\|\frac{P_{\theta_{j-1}}}{P_{\theta_j}}\right\|_2\right\} \le 10$$

Guarantees that HitAndRun requires $N = \tilde{O}(n^3)$ steps in order to ensure mixing to the stationary distribution P_{θ_j}.

Proof sketch. The proof is based on iteratively applying the following Theorem from [12]:

Theorem 5. *Let f be a density proportional to $e^{-a^T x}$ over a convex set K such that*

[a]. the level set of probability 1/64 contains a ball of radius s

[b]. $\mathbb{E}_f(\|x-\mu_f\|^2) \le S$, where $\mu_f = \mathbb{E}_f[x]$ is the mean of f

[c]. the ℓ_2 norm of the starting distribution σ w.r.t. the stationary distribution of HitAndRun *denoted π_f, is at most M.*

Let σ^m be the distribution of the current point after m steps of HitAndRun *applied to f. Then, for any $\tau > 0$, after $m = O(\frac{n^2 S^2}{s^2} \log^5 \frac{nM}{\tau})$ steps, the total variation distance of σ^m and π_f is less than τ.*

The proof proceeds to show that conditions [a]-[c] of the theorem above are all satisfied if indeed condition (6) is satisfied, along the steps below. Once it is established that the conditions of the theorem hold, then the next HitAndRun walk mixes and computes warm start and variance estimates for the next epoch. Then again, the conditions of the theorem hold, and this whole process is repeated for the entire temperature schedule.

To show conditions [a]-[c], first notice that condition (6) is essentially equivalent to condition [c] above. Thus we only need to worry about conditions [a],[b].

[I]. For simplicity, we assumed that at the current iteration, $\Sigma_j = I$ is the identity. This can be assumed by a transformation of the space, and allows for simpler discussion of isotropy of densities (otherwise, the isotropy condition would be replaced by relative-isotropy w.r.t the current variance).

[II]. A density f is C-isotropic if for any unit vector $\|v\| = 1$,

$$\frac{1}{C} \leq \int_{\mathbb{R}^n} (v^\top (x - \mu_f))^2 f(x) dx \leq C$$

It is shown (Lemma 4.2) that if the density given by f is $O(1)$-isotropic, then conditions [a],[b] are satisfied with $\frac{S}{s} = \tilde{O}(\sqrt{n})$.

[III]. It is shown (Lemma 4.3) that if f is C-isotropic, and $\max\left\{ \left\| \frac{f}{g} \right\|_2, \left\| \frac{g}{f} \right\|_2 \right\} \leq D$, then g is CD-isotropic.

[IV]. Since condition (6) holds, together with the previous points [II,III] this implies that $f_{\theta_{j+1}}$ is isotropic for some constant. Thus, conditions [a]-[c] of Theorem 5 hold. Therefore we can sample sufficiently many samples to estimate the covariance matrix Σ_{j+1} and proceed to the next epoch.

Throughout the proof special care needs to be taken to ensure that repeated samples are nearly-independent for various concentration lemmas to apply, we omit discussion of these and the reader is referred to the original paper of [7].

□

C Interior point methods with a membership oracle

Below we sketch a universal IPM algorithm - one that applies to any convex set described by a membership oracle - that can be implemented to run in polynomial time. This algorithm is an instantiation of Algorithm 3 with the particular barrier function $A^*(x)$ as defined in section 4.1.

Without loss of generality, we can assume our goal is to (approximately) compute the update direction

$$\nabla^{-2} A^*(x)(\theta - \nabla A^*(x))$$

for some x which is already within the Dikin ellipsoid of radius $1/2$ around $x(\theta)$. First, we note that the IPM analysis of [15] allows one to replace the inverse hessian $\nabla^{-2}A^*(x)$ with the nearby $\nabla^{-2}A^*(x(\theta)) = \text{CovMtx}(P_\theta)$. Of course the latter can be estimated via sampling, in the sense that the estimate $\hat{\Sigma}$ will be "ϵ-isotropically close":

$$(1-\epsilon)v^\top \nabla^2\Psi(\theta')v \le v^\top \hat{\Sigma} v \le (1+\epsilon)v^\top \nabla^2\Psi(\theta')v$$

for any unit vector v. See, for example, [1] on the concentration of empirical covariance matrices.

It remains to compute $\nabla A^*(x)$. Define $\theta(x)$ to be

$$\theta(x) = \arg\max_\theta \theta \cdot x - \log \int_K \exp(-\theta \cdot y)dy = \nabla A^*(x) \tag{21}$$

Then $\theta(x)$ can be computed in polynomial time by another interior point algorithm – this problem, however, is much simpler to work with. Define $\Psi(\theta') := \theta \cdot x - \log \int_K \exp(-\theta \cdot y)dy$ to be the objective we want to optimize. Notice that $\nabla\Psi(\theta') = x - \mathbb{E}_{X' \sim P_{\theta'}}[X']$ and the latter can be estimated to within ϵ via SimulatedAnnealing with $\tilde{O}(n/\epsilon^2)$ samples. The hessian $\nabla^2\Psi(\theta') = -\text{CovMtx}(P_{\theta'})$ can similarly be estimated with an ϵ-isotropically close empirical covariance. Because the error gap is multiplicatively close to 1, the inverse operation on $\nabla^2\Psi(\theta')$ maintains the approximation.