Simulated Annealing for Convex Optimization

Adam Kalai
TTI-Chicago*

Santosh Vempala
MIT†

Abstract

We apply the method known as *simulated annealing* to the following problem in convex optimization: minimize a linear function over an arbitrary convex set, where the convex set is specified only by a membership oracle. Using distributions from the Boltzmann-Gibbs family leads to an algorithm that needs only $O^*(\sqrt{n})$ *phases* for instances in $\mathbb{R}^n$. This gives an optimization algorithm that makes $O^*(n^{4.5})$ calls to the membership oracle, in the worst case, compared to the previous best guarantee of $O^*(n^5)$.

The benefits of using annealing here are surprising due to the fact that such problems have no local minima that are not also global minima. Hence, we conclude that one of the advantages of simulated annealing, in addition to avoiding poor local minima, is that in these problems it converges faster to the minima that it finds. We also give a proof that under certain general conditions, the Boltzmann-Gibbs distributions are optimal for annealing on these convex problems.

1 Introduction

Simulated annealing, proposed by Kirkpatrick et al. [12], is a randomized search method for optimization. It tries to improve a solution by walking randomly in the space of possible solutions and gradually adjusting a parameter called "temperature." At high temperature, the random walk is almost unbiased and it converges to essentially the uniform distribution over the whole space of solutions; as the temperature drops, each step of the random walk is more likely to move towards solutions with a better objective value, and the distribution is more and more biased towards the optimal solutions. The sequence of temperatures and lengths of time for which they are maintained is called the *annealing schedule* in analogy with statistical mechanics.

Although notoriously difficult to analyze, the usual justification for its empirical success [17] is that "it avoids getting stuck in local optima." In this paper, we analyze it on problems where all the local optima are also global optima. Our results suggest that annealing has further advantages besides avoiding local optima. In particular, it also speeds up convergence to the optimum.

An important problem in this class is minimizing a linear function over a convex set. In the most general version, the convex set is presented only by a membership oracle [7]. In Section 4, we show that simulated annealing, which can be viewed as an interior-point algorithm, takes only $O^*(\sqrt{n})$ phases and $O^*(n^{4.5})$ oracle queries to solve this problem. It is faster than the current best algorithm by a factor of $\sqrt{n}$. (The $O^*(\cdot)$ notation hides poly-logarithmic factors in the parameters of the problem, i.e., $\log^k(nR^2/(r\varepsilon\delta))$, where k is a constant, n is the dimensionality of the problem, ε is the distance from optimality, R/r is the ratio of containing and contained balls for the convex set, and the algorithm succeeds with probability $1-\delta$.) This illustrates (and provides a rigorous guarantee for) the advantage of simulated annealing even for problems with no bad local minima.

Simulated annealing is a special case of a stochastic search method that starts with one distribution (e.g., uniform over a given convex body) and gradually changes it to another target distribution (e.g., one that is concentrated near the optimum). The intermediate distributions satisfy the following two properties:

*kalai@tti-c.org
†vempala@math.mit.edu

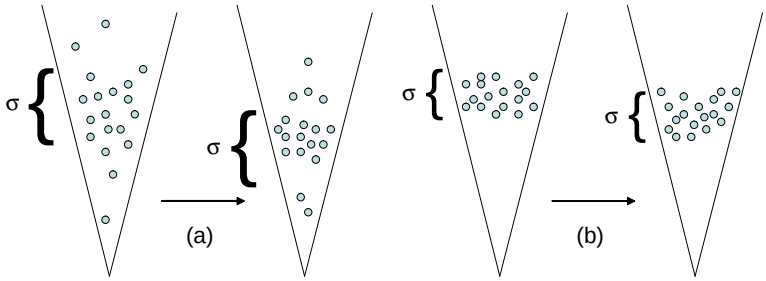


Figure 1: Optimization over a high-dimensional cone: (a) A pair of consecutive Boltzmann distributions $e^{-c\cdot x/T}$. The standard deviation σ in the direction of optimization is large. (b) The same picture for a pair of uniform distributions (used by [2]) over truncated cones $c \cdot x \leq T$. The standard deviations are much smaller, allowing less movement and requiring more phases.

1. Any two consecutive distributions must not be too different; one way to formalize this is to require that their *total variation distance* (see Section 3) is bounded away from 1.

2. All the distributions along the way must be efficiently sampleable. The most general class of distributions for which we currently have efficient sampling algorithms are the class of logconcave distributions.

In simulated annealing, the intermediate distributions are all from the exponential family (density at x is proportional to $e^{-c^T x}$ for some vector c) restricted to some domain. The random walk algorithm of Bertsimas and Vempala [2], for example, maintains a uniform distribution over a convex body. In each phase, it restricts the current convex body by a half-space. This choice requires $O^*(n)$ phases in the worst case.

From this perspective, it is natural to ask what choice of intermediate distributions would be optimal for convex problems. In Section 5, we show that, in the worst-case, the exponential family used in simulated annealing, is in fact the best possible. We give an example where any method satisfying the above two properties needs $\Omega(\sqrt{n})$ phases[1]. Thus, in this sense, simulated annealing is an optimal stochastic search method. Moreover, we have shown yet another optimality property of the Boltzmann distributions (a different, well-known property is that of maximum entropy).

Finally, in Section 6, we suggest how the algorithm might extended to the problem of minimizing a convex function over a convex set, where again the set is specified by a membership oracle. While our results show that the number of phases will be $O^*(\sqrt{n})$ for any convex function, further analysis of rapidly-mixing walks for logconcave functions is required to achieve the $O^*(n^{4.5})$ membership queries guarantee. Our approach here shows that one can move between any two logconcave densities in $O^*(\sqrt{n})$ steps where each step moves between logconcave densities whose total variation distance is bounded away from 1.

1.1 Related work

Our approach is an improvement on the algorithm of Bertsimas and Vempala [2], which introduced the analysis of an efficient stochastic search method for convex optimization but required $O^*(n)$ phases. Their method, involving a sequence of uniform distributions over sets with smaller objective function values $c \cdot x$, is illustrated in Figure 1. Lovász and Vempala used a reverse annealing technique for estimating volume [14] where they slowly *increased* the temperature. Our analysis is similar to theirs, and we will use some of the tools developed there.

Stochastic search methods have been analyzed for special cases of global optimization (see [24, 25]). It is known that an exponentially long annealing schedule can guarantee convergence to the global optimum (even for non-convex problems) [8]. Simulated annealing has also been shown to be efficient for finding planted

[1]The $\Omega(T(n))$ notation means that a function grows at a rate *at least* as fast as $kT(n)$ for every constant $k > 0$, for sufficiently large n.

bisections in a random graph [10] and for minimizing one-dimensional fractal functions [21]. Although originally proposed for discrete optimization problems, simulated annealing has been widely used for continuous optimization. The book by Spall [22] provides an introduction to both the theoretical and practical aspects of annealing.

2 Algorithm and guarantees

We apply annealing to the following linear minimization problem: for a unit vector $c \in \mathbb{R}^n$ and a convex set $K \subset \mathbb{R}^n$:

$$\min_{x \in K} c \cdot x$$

We assume only that we are given a membership oracle, that identifies whether or not a point is in the set K, as well as a starting point in it. As is standard [7], we need an upper bound R on the radius of a ball that contains K, and a lower bound r on the radius of a ball around the starting point contained in K. Approaches such as simplex and interior point methods do not seem to generalize to this problem. The ellipsoid algorithm solves the problem using $O^*(n^{10})$ membership queries. This has been improved to $O^*(n^5)$ using randomized search [2].

Our approach runs over phases $i = 0, 1, \ldots, m$:

1. Let the *temperature* $T_i = R(1 - 1/\sqrt{n})^i$.

2. Move the current point to a sample from a distribution μ_i whose density is proportional to

$$f_i(x) = e^{-c \cdot x / T_i}.$$

   Do this by executing k steps of a biased random walk, to be described later.

3. Using $O^*(n)$ samples observed during the above walk, estimate the covariance matrix V of the above distribution, i.e., $V_{ij} = \mathsf{E}[X_i X_j] - \mathsf{E}[X_i]\mathsf{E}[X_j]$.

Steps 1 and 2 are standard for simulated annealing. In particular, the exponential temperature schedule (fixed decay rate) and the Boltzmann distributions (of the form $e^{-E(x)/T}$) as is typical and was proposed with the introduction of simulated annealing [12].

For many "round" shapes, such as a ball, cube, cone or cylinder, Step 3 is unnecessary or may be implemented using a many fewer samples. However, for arbitrary convex sets K, the update can be done using $O^*(n)$ samples (in the worst case, it must be done every phase). Step 3 estimates the shape of the distribution, i.e., which dimensions are long and which are short, via its covariance matrix. (Recall that a covariance matrix is what is used to define an n-dimensional normal distribution or ellipsoid.) In order to get the best possible rates of convergence for the random walk in Step 2, we need to incorporate this shape estimate into the walk, so that we have a better idea of which way to step.

Theorem 2.1 *For any convex set $K \in \mathbb{R}^n$, with probability $1 - \delta$, the algorithm below given a membership oracle $\mathcal{O}_K$, starting point X_{init}, R, r, $I = O(\sqrt{n} \log(Rn/\varepsilon\delta))$, $k = O^*(n^3)$, and $N = O^*(n)$,[2] outputs a point $X_I \in K$ such that*

$$c \cdot X_I \leq \min_{x \in K} c \cdot x + \varepsilon.$$

The total number of calls to the membership oracle is $O^*(n^{4.5})$.

The algorithm goes through a series of temperatures, starting high and decreasing. At each temperature, it runs the *hit-and-run* random walk. The stationary density of this random walk will be proportional to $e^{-c \cdot x / T}$ for any point $x \in K$, where T is the current temperature.

[2] Again, the O^* notation hides logarithmic factors. In this case, for example, the theorem states that there exists a polynomial p such that for $N = np \log(nR^2/(r\varepsilon\delta))$, the theorem holds.

The Algorithm.

Inputs:

$n \in \mathbb{N}$	(dimensionality)
$\mathcal{O}_K : \mathbb{R}^n \to \{0,1\}$	(mebership oracle for convex set K)
$c \in \mathbb{R}^n$	(direction of minimization, $\|c\| = 1$)
$X_{\text{init}} \in K$	(starting point)
$R \in \mathbb{R}_+$	(radius of ball containing K centered at X_{init})
$r \in \mathbb{R}_+$	(radius of ball contained in K centered at $X_{mboxinit}$
$I \in \mathbb{N}$	(number of phases)
$k \in \mathbb{N}$	(number of steps per walk)
$N \in \mathbb{N}$	(number of samples for rounding)

- (X_0, V_0):=UniformSample$(X_{\text{init}}, \mathcal{O}_K, R, r)$
- For $i = 1, 2, \ldots, I$:
  - $T_i := R\left(1 - \frac{1}{\sqrt{n}}\right)^i$
  - X_i:=hit-and-run$(e^{-c\cdot x/T_i}, \mathcal{O}_K, V_{i-1}, X_{i-1}, k)$
  - Update Covariance:
    - For $j = 1$ to N: X_i^j:=hit-and-run$(e^{-c\cdot x/T_i}, \mathcal{O}_K, V_{i-1}, X_{i-1}, k)$
    - $V_i := \frac{1}{N}\sum_j X_i^j (X_i^j)^T - \frac{1}{N}\sum_j X_i^j \left(\frac{1}{N}\sum X_i^j\right)^T$
- Return X_I.

The UniformSample routine picks a uniformly random point from K. Additionally, it estimates the covariance matrix V_0 of the uniform distribution over the set K. This subroutine is the "Rounding the body" algorithm of Lovasz and Vempala [14], which uses $X_{\text{init}}, \mathcal{O}_K, R$, and r, and returns a nearly random point while making $O^*(n^4)$ membership queries.

We next describe the random walk precisely. The hit-and-run random walk takes as input a function f, a membership oracle, a covariance matrix V, a starting point $x \in K$, and a number of steps k. It then performs the following procedure k times:

- Pick a random vector v according to the n-dimensional normal distribution with mean 0 and covariance matrix V. Let ℓ be the line through the current point in the direction v.
- Move to a random point on the intersection of ℓ and K (this is a one-dimensional chord in K), where point is chosen with density proportional to the function f (restricted to the chord).

It is well-known that the stationary distribution of this random walk has density proportional to f. The rate of convergence of the walk to its stationary distribution depends on the starting point and how good an approximation V is to the true covariance matrix of the stationary distribution.

Our algorithm, like other continuous applications of annealing [22], maintains the form of annealing by performing a random walk whose stationary distributions are the Boltzmann-Gibbs distributions. Indeed, our algorithm could be further also be stated in terms of defining a neighborhood set and performing a uniform step with possible rejection, as is more familiar in annealing. However, the above random walk is slightly more efficient and easier to state.

Lastly, as mentioned before, for many common shapes the third step is unnecessary as the (normalized) covariance matrix remains more or less unchanged. In practice, the algorithm could be run with spherical steps on many shapes, though the reshaping is necessary in the worst case.

3 Preliminaries

Definition 1 *A function $f : \mathbb{R}^n \to \mathbb{R}_+$ is* logconcave *for any two points $a, b \in \mathbb{R}^n$ and any $\lambda \in (0, 1)$,*

$$f(\lambda a + (1 - \lambda)b) \geq f(a)^\lambda f(b)^{1-\lambda}.$$

In other words, a nonnegative function f is logconcave if its support is convex and $\log f$ is concave. For example, a function that is constant over a bounded convex set and zero outside the set is logconcave. Another example is a Gaussian density function. It can be easily verified from the definition above that the product of two logconcave functions is also logconcave (but not necessarily their sum). The following fundamental property of logconcave functions was proved by Dinghas [4], Leindler [13] and Prékopa [18, 19].

Theorem 3.1 *All marginals of a logconcave function are logconcave. The convolution of two logconcave functions is logconcave.*

A logconcave distribution in $\mathbb{R}^n$ is one whose density is a logconcave function. The next lemma is from [14].

Lemma 3.2 *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a logconcave function with support K. For $a > 0$, define*

$$Z(a) = \int_K f(ax)\, dx.$$

Then $a^n Z(a)$ is a logconcave function of a.

The following variant will also be useful.

Lemma 3.3 *Let $f : \mathbb{R}^n \to \mathbb{R}$ be an integrable logconcave function. For $a > 0$, define*

$$Y(a) = \int_{\mathbb{R}^n} f(x)^a\, dx.$$

Then $a^n Y(a)$ is a logconcave function of a.

Proof. Let $F : \mathbb{R}^n \times \mathbb{R}_+ \to \mathbb{R}$ be defined as

$$F(x, t) = f\left(\frac{x}{t}\right)^t.$$

Then we can verify that $F(x, t)$ is logconcave: for any $\lambda \in [0, 1]$, using the logconcavity of f,

$$\begin{aligned} F\left(\lambda(x, t) + (1 - \lambda)(x', t')\right) &= f\left(\frac{\lambda x + (1 - \lambda)x'}{\lambda t + (1 - \lambda)t'}\right)^{\lambda t + (1-\lambda)t'} \\ &= f\left(\frac{\lambda t}{\lambda t + (1 - \lambda)t'}\frac{x}{t} + \frac{(1 - \lambda)t'}{\lambda t + (1 - \lambda)t'}\frac{x'}{t'}\right)^{\lambda t + (1-\lambda)t'} \\ &\geq f\left(\frac{x}{t}\right)^{\lambda t} f\left(\frac{x'}{t'}\right)^{(1-\lambda)t'} \\ &= F(x, t)^\lambda F(x', t')^{1-\lambda}. \end{aligned}$$

Hence, the marginal of F along t is also logconcave, i.e.,

$$\int_{\mathbb{R}^n} F(x, t)\, dx = \int_{\mathbb{R}^n} f\left(\frac{x}{t}\right)^t\, dx = t^n \int_{\mathbb{R}^n} f(x)^t\, dx = Y(t)$$

is a logconcave function of t. □

To compare two distributions, we consider two measures. The *total variation distance* between ν and π is:

$$\|\nu - \pi\|_{tv} = \frac{1}{2}\int_{\mathbb{R}^n} |d\nu(x) - d\pi(x)|dx$$

The L_2 norm of a distribution ν w.r.t. a distribution π is defined as,

$$\|\nu/\pi\| = \mathsf{E}_\nu\left(\frac{d\nu(x)}{d\pi(x)}\right) = \int_K \frac{d\nu(x)}{d\pi(x)}d\nu = \int_K \left(\frac{d\nu(x)}{d\pi(x)}\right)^2 d\pi.$$

These two distance measures can be related as follows, as proven in the appendix.

Lemma 3.4

$$\|\nu - \pi\|_{tv} \leq \max\left\{\frac{1}{2}, 1 - \frac{1}{\|\nu/\pi\|}\right\}.$$

4 Analysis of the algorithm

In Section 4.1, we show that, as the temperature decreases at an exponential rate, we rapidly approach the minimum. In particular, we show that $\mathsf{E}[c \cdot x]$, for x drawn according to the stationary distribution of the random walk, approaches the minimum of the function at a rate proportional to the temperature. However, it remains to show that we take sufficiently many steps in the random walk to be quite close to the stationary distribution. In Section 4.2, we state the main theorem for proving the rapid mixing of the hit-and-run walk, i.e., that it quickly approaches the stationary distribution. To apply the theorem we show two things: first, that the covariance estimates do not change too much from one phase to the next, and second, that consecutive distributions do not change too much (Section 4.2.1). We put all of these together to prove our main theorem in Section 4.3.

In Section 5, we show that no sequence of distributions can, in general, solve the problem in less than $\Omega(\sqrt{n})$ phases.

4.1 Convergence to optimal

The following lemma shows that as the temperature decreases, the expected value of the function on the stationary distribution rapidly approaches the minimum.

Lemma 4.1 *For any unit vector $c \in \mathbb{R}^n$, temperature T, and X chosen according to distribution with density proportional to $e^{-c\cdot x/T}$,*

$$\mathsf{E}(c \cdot X) \leq nT + \min_K c \cdot x.$$

Proof. First, without loss of generality, we may assume that $c = (1, 0, 0, \ldots)$, so $c \cdot x = x_1$. Let's say the value of this expectation is $\mathsf{E}(c \cdot X) = v$. Also, let $H(y)$ be the hyperplane $x_1 = y$, namely,

$$H(y) = \{x \in \mathbb{R}^n | x_1 = y\}.$$

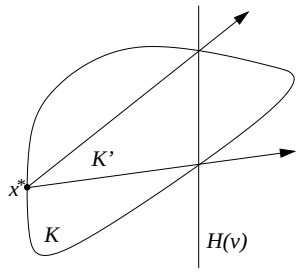
Next, consider how changing the set K affects v. Clearly, adding mass to K in points with $x_1 > v$ will increase v, as will removing mass from K where $x_1 < v$.

Let us change the set in such ways. Take some optimal point x^*, i.e., $c \cdot x^* = \min_K c \cdot x$. WLOG we assume that $c \cdot x^* = 0$. Consider the convex set

$$K' = \{x \in \mathbb{R}^n | x = x^* + \alpha(y - x^*) \text{ for some } \alpha \geq 0, y \in K \cap H(v)\}.$$

In other words, K' is the infinite extension of the cone with vertex x^* and convex base $K \cap H(v)$, as shown in the figure. It is not difficult to see, by definition of convexity:

$$\begin{aligned} K' \cap \{x | x_1 \leq v\} &\subseteq K \cap \{x | x_1 \leq v\} \\ K' \cap \{x | x_1 \geq v\} &\supseteq K \cap \{x | x_1 \geq v\} \end{aligned}$$



Thus, using K' instead of K, we will have only increased $\mathsf{E}(c \cdot X)$. So, it suffices to prove the lemma for $K = K'$.

$$\begin{aligned}
\mathsf{E}(c \cdot X) &= \frac{\int_{-\infty}^{\infty} y e^{-y/T} \mathrm{vol}_{n-1}(K' \cap H(y)) dy}{\int_{-\infty}^{\infty} e^{-y/T} \mathrm{vol}_{n-1}(K' \cap H(y)) dy} \\
&= \frac{\int_0^{\infty} \mathrm{vol}_{n-1}(K' \cap H(v)) y \left(\frac{y}{v}\right)^{n-1} e^{-y/T} dy}{\int_0^{\infty} \mathrm{vol}_{n-1}(K' \cap H(v)) \left(\frac{y}{v}\right)^{n-1} e^{-y/T} dy} \\
&= \frac{\int_0^{\infty} y^n e^{-y/T}\, dy}{\int_0^{\infty} y^{n-1} e^{-y/T}\, dy} \\
&= \frac{n! T^{n+1}}{(n-1)! T^n} = nT.
\end{aligned}$$

The last step uses the fact that

$$\int_0^{\infty} y^n e^{-y/a}\, dy = n! a^{n+1}.$$

□

4.2 Sampling at a fixed temperature

The following theorem about hit-and-run applied to an exponential density was proved in [16].

Theorem 4.2 *Let f be a density proportional to $e^{-a^T x}$ over a convex set K such that (i) the level set of probability $1/64$ contains a ball of radius s, (ii) $\mathsf{E}_f(|x - z_f|^2) \leq S^2$ and (iii) the L_2 norm of the starting distribution σ w.r.t. the stationary distribution π_f is at most M. Let σ^m be the distribution of the current point after m steps of uniform hit-and-run applied to f. Then, for any $\tau > 0$, after*

$$m = O\left(\frac{n^2 S^2}{s^2} \ln^5 \frac{nM}{\tau}\right),$$

steps, the total variation distance of σ^m and π_f is less than τ.

In the above theorem, z_f denotes the mean the density f, i.e., $z_f = \mathsf{E}_f[x]$. A level set refers to a set $\{x \in K | e^{-a^T x} \geq \theta\}$, where $\theta \in \mathbb{R}$. And the level set of probability $1/64$ corresponds to the choice of θ so that the level set has measure $1/64$.

Also in the above, uniform hit-and-run refers to hit-and-run where the direction is chosen uniformly at random. However, in our description of hit-and-run, we chose our direction according to a normal distribution with some covariance matrix V. (Thus uniform hit-and-run chooses its direction from a spherical normal distribution with identity covariance matrix.) These two approaches are basically the same. For the covariance matrix V of the sample, one can apply the affine transformation $y = V^{-\frac{1}{2}} x$ to the space. Now, if we choose a vector X at random from the n-dimensional normal distribution with covariance V and transform

it to $Y = V^{-\frac{1}{2}}X$, it will be distributed according to a normal distribution, and the new covariance matrix will be:

$$E[YY^T] = E[V^{-\frac{1}{2}}XX^T(V^{-\frac{1}{2}})^T] = V^{-\frac{1}{2}}VV^{-\frac{1}{2}} = I$$

Thus, choosing a random direction according to V in the original space is equivalent to choosing a uniformly random direction in the transformed space. For the rest of analysis, we find it easier to use the latter perspective. In other words, we imagine a rounding step that transforms the body by $V^{-\frac{1}{2}}$, making the sampler choose steps from a spherically symmetric distribution.

We define the distribution μ_i to be stationary distribution of the ith phase, i.e., proportional to $e^{-c\cdot x/T_i}$ over the convex set K, *transformed* in the above manner. In particular, if our estimate of the covariance matrix was perfectly accurate, μ_i would be perfectly isotropic.

The remainder of this section is dedicated to showing that conditions (i) and (ii) hold in the above theorem with $S/s = O(\sqrt{n})$. In Section 4.2.1, we show that $M < 5$ in each phase for (iii).

The next definition is a measure of the "roundness" of a density function.

Definition 2 *A density function f with centroid z_f is said to be* C-isotropic *if for every unit vector v,*

$$\frac{1}{C} \leq \int_{\mathbb{R}^n} \big(v \cdot (x - z_f)\big)^2 f(x)dx \leq C.$$

As the algorithm proceeds it is possible that the transformed density becomes less and less isotropic (i.e., C gets larger and larger). In terms of covariance, it is possible that our matrix V stops being an accurate estimate of the true covariance of the distribution. It is for this reason, that we re-estimate V periodically, and for the purposes of analysis, we can imagine periodically transforming the space as well.

We next show that the distribution to be sampled remains C-isotropic for some constant C throughout the course of the algorithm. Our starting point is the following lemma from [15] which shows that for a near-isotropic density, $S/s = O(\sqrt{n})$.

Lemma 4.3 *Let $d\mu$ be a C-isotropic density function. Then $\mathsf{E}_f(|x - z_f|^2) \leq Cn$ and any level set L of f contains a ball of radius $\mu(L)/e\sqrt{C}$.*

From the lemma it follows that if f is C-isotropic, then $(S/s)^2 = O(C^2 n) = O(n)$, if C is a constant.

We continue by showing that isotropy is changed by at most a fixed constant factor from one phase to the next.

Lemma 4.4 *Let f and g be logconcave densities over K with centroids z_f and z_g, respectively. Then for any $c \in \mathbb{R}^n$,*

$$\mathsf{E}_f((c \cdot (x - z_f))^2) \leq 16\mathsf{E}_f\left(\frac{f}{g}\right)\mathsf{E}_g((c \cdot (x - z_g))^2)$$

Proof. By the Cauchy-Schwartz inequality,

$$\begin{aligned}\int_K \frac{f(x)}{g(x)}f(x)dx \int_K (c \cdot (x - z_g))^2 g(x)dx &\geq \left(\int_K |c \cdot (x - z_g)|f(x)dx\right)^2 \\ \mathsf{E}_f\left(\frac{f(x)}{g(x)}\right)\mathsf{E}_g((c \cdot (x - z_g))^2) &\geq \mathsf{E}_f(|c \cdot (x - z_g)|)^2 \qquad (1)\end{aligned}$$

Now, for an arbitrary logconcave density h, Theorem 5.22 of [15] states:

$$\mathsf{E}_h(|y|)^2 \geq \mathsf{E}_h(|y|^2)/(2k)^2.$$

Let $x \in \mathbb{R}^n$ be a random variable chosen according to f, and $y = c \cdot (x - z_g)$. Then y also has a logconcave distribution. This is because any marginal of a logconcave function is also logconcave (Lemma 3.1). So

in our case the marginal $c \cdot x$ has a logconcave distribution; further, translation preserves logconcavity. So $y = c \cdot x - c \cdot z_g$ has a logconcave distribution. Thus we get,

$$\mathsf{E}_f(|c \cdot (x - z_g)|)^2 \geq \mathsf{E}_f((c \cdot (x - z_g))^2)/16 \tag{2}$$

Finally, the following inequality just says that the centroid is the point which minimizes the average squared distance to a point.

$$\mathsf{E}_f((c \cdot (x - z_g))^2) \geq \mathsf{E}_f((c \cdot (x - z_f))^2)$$

Combining (1) and (2) with the above inequality, we obtain the lemma. □

Thus, if the transformed distribution in one phase (with density μ_{i-1}) is C-isotropic, then the distribution μ_i in the next phase with the same transformation will be C'-isotropic for $C' = 16C \max(\|\mu_{i-1}/\mu_i\|, \|\mu_i/\mu_{i-1}\|)$. The main theorem of this section is the following.

Theorem 4.5 *Using $N = O^*(t^3 n)$ samples per phase in each of I phases, with probability $1 - I/2^t$, every distribution μ_i encountered by the sampling algorithm is* 160*-isotropic.*

Proof. (sketch) For the main part of the proof, we assume that the sampler gives us independent points from exactly the desired distribution. By Corollary A.2, the current distribution, say μ_{i-1}, is 2-isotropic after the rounding step with probability $1 - 1/2^{t+1}$. To bound the isotropy of the next distribution to be sampled, μ_i, we apply Lemma 4.4 to the distributions μ_{i-1}, μ_i. First, note that by Lemma 4.6 of the following section, for $n \geq 8$,

$$E_{\mu_i}\left(\frac{d\mu_i(x)}{d\mu_{i-1}(x)}\right) < 5.$$

Thus,

$$E_{\mu_i}((v^T(x - z_{\mu_i}))^2) \leq 80 E_{\mu_{i-1}}(v^T(x - z_{\mu_{i-1}}))^2) \leq 160.$$

The lower bound is obtained by switching the roles of μ_{i-1} and μ_i.

We have assumed that the samples used in each phase are independent and exactly from the target distribution. In fact, the random walk only gives us samples from (a) a distribution that is close to the right one and (b) they are only nearly independent. The first difficulty (a) can be handled by a trick sometimes known as "divine intervention" (see e.g., [14]). Theorem 4.2 guarantees samples from a distribution within variation distance τ from the target distribution, and the dependence of the number of steps on τ is polylogarithmic. Set $\tau = 2^{-t}/(2N)$. For a sample X which is drawn from a distribution ν that is within τ of the stationary distribution π, we can think of it as being drawn as follows: first we pick X according to the stationary distribution π. Then, with probability $\|\nu - \pi\|_{tv} \leq \tau$ we change the point X so that it is distributed exactly according ν. The important point is that the distribution ν is obtained by modifying X only with probability τ. We then bound the probability that the distribution of X is modified. In this lemma, since we are drawing N samples with $\tau = 2^{-t}/(2N)$, we have that the probability that the modification occurs in any phase is at most 2^{-t-1}. The earlier analysis which assumed we are sampling exactly from π can now be applied with a failure probability of at most 2^{-t-1}.

The other difficulty (b) can be handled in two different ways. The first is via a small modification in the algorithm. Suppose the covariance estimate needs m samples in any one phase. Then we start with m truly independent samples which lead to m independent threads of samples, i.e., in each phase a random walk is started from the current point in each thread to give a new point after a prescribed number of steps of the walk. The threads thus remain independent and when the covariance estimate is computed we have m truly independent samples (note that the samples within a thread are not independent). The other way to handle (b) without any change to the algorithm is by quantifying the dependence of the random variables as in [11] and [14]. Specifically, we use the notion of μ-independence. Two random variables X, Y are said to be μ-independent where

$$\mu(X, Y) = \sup_{A,B} \left|\mathsf{P}(X \in A,\ Y \in B) - \mathsf{P}(X \in A)\mathsf{P}(Y \in B)\right|$$

and A and B range over measurable subsets of the ranges of X and Y respectively. As shown in Lemma 4.3(a) of [14], consecutive samples produced by the random walk are τ-independent (and the dependence on τ is polylogarithmic). With τ set to be an inverse polynomial, we can apply the sampling bound of Theorem A.1 and Corollary A.2 with a small increase in the number of samples (for a similar analysis, see Theorem 5.11 in [11]).

□

4.2.1 Warm start

We bound the L_2 distance between two consecutive distributions.

Lemma 4.6 *For $n \geq 8$, $\quad \|\mu_i/\mu_{i+1}\| \leq 5$ and $\|\mu_{i+1}/\mu_i\| \leq 4$.*

Proof. As in the proof of Lemma 4.4 in [14] let $Z(a) = \int_K e^{-ax}\,dx$. Then,

$$\begin{aligned}\mathsf{E}_{\mu_i}\left(\frac{d\mu_i(x)}{d\mu_{i+1}(x)}\right) &= \frac{\int_K e^{-c\cdot x/T_i + c\cdot x/T_{i+1}} e^{-c\cdot x/T_i}dx}{\int_K e^{-c\cdot x/T_i}dx}\,\frac{\int_K e^{-c\cdot x/T_{i+1}}dx}{\int_K e^{-c\cdot x/T_i}dx}\\ &= \frac{Z(2/T_i - 1/T_{i+1})Z(1/T_{i+1})}{Z(1/T_i)Z(1/T_i)}.\end{aligned}$$

Now $a^n Z(a)$ is a logconcave function of a by Lemma 3.2. Thus,

$$\left(\frac{a+b}{2}\right)^{2n} Z\left(\frac{a+b}{2}\right)^2 \geq a^n Z(a) b^n Z(b) \quad \text{i.e.,} \quad \frac{Z(a)Z(b)}{Z(\frac{a+b}{2})^2} \leq \left(\frac{\left(\frac{a+b}{2}\right)^2}{ab}\right)^n.$$

Applying this with $a = 2/T_i - 1/T_{i+1}$ and $b = 1/T_{i+1}$, we get

$$\begin{aligned}\mathsf{E}_{\mu_i}\left(\frac{d\mu_i(x)}{d\mu_{i+1}(x)}\right) &\leq \left(\frac{1/T_i^2}{(2/T_i - 1/T_{i+1})(1/T_{i+1})}\right)^n\\ &= \left(\frac{(T_{i+1}/T_i)^2}{2T_{i+1}/T_i - 1}\right)^n\\ &= \left(\frac{(1-1/\sqrt{n})^2}{2(1-1/\sqrt{n}) - 1}\right)^n\\ &= \left(1 + \frac{1}{n - 2\sqrt{n}}\right)^n\\ &\leq e^{n/(n-2\sqrt{n})} < 5 \quad \text{for } n > 8.\end{aligned}$$

The analysis for $\|\mu_{i+1}/\mu_i\|$ is similar. □

4.3 Proof of Theorem 2.1

We start with temperature $T_0 = R$. This is chosen so that the uniform distribution π is a warm start for μ_0, i.e., $\|\pi/\mu_0\| \leq e$. After $\sqrt{n}\ln\frac{3Rn}{\varepsilon\delta}$ phases, by Lemma 4.1,

$$\mathsf{E}_{\mu_I}(c\cdot x) \leq \min_{x\in K} c\cdot x + \varepsilon\delta/3.$$

By Markov's inequality, this guarantee in expectation can be translated into a high-probability guarantee that, with probability at most $\delta/3$, a random x drawn from μ_I would have $c\cdot x$ larger than the minimum plus ε. In each phase, we obtain up to $N+1 = O^*(n)$ samples. Hence there are at most $(N+1)I$ samples used.

Unfortunately, X_I is not drawn exactly from the stationary distribution μ_I. However, by a choice of $k = O^*(n^3)$, we can get within $\tau = \delta/(3(N+1)I)$ to the stationary distribution (so close, that with probability $1 - \delta/3$, no divine intervention occurs and we can assume the examples were drawn from the stationary distribution – see the discussion of "divine intervention" in the proof of Lemma 4.5). This choice of k suffices, because by Theorem 4.5 and Lemma 4.6, we can apply Theorem 4.2 with $M \leq 5$ and $S/s = O(\sqrt{n})$ and $\tau = \delta/(3(N+1)I)$.

Thus, each sample takes only $k = O^*(n^3)$ steps of the random walk. Hence, the number of membership queries per phase is $(N+1)O^*(n^3) = O^*(n^4)$, which gives an overall query complexity of $O^*(n^{4.5})$.

5 The optimality of annealing for convex problems

Our analysis of annealing raises some interesting issues with regard to why simulated annealing seems to work so well in practice. A common justification of simulated annealing (and other stochastic search methods) is that it helps avoid bad local minima. This argument does not apply to convex problems where there are no such bad minima.

We propose an alternate justification for annealing: it is a type of interior point algorithm. It is well-known that trying to move directly in the direction of optimization quickly runs into problems when we hit the boundary. Decreasing the temperature too quickly may have the same effect, because it forces such steps.

This section is devoted to explaining annealing by showing that the Boltzmann distributions with a geometric temperature schedule are worst-case optimal, to within a constant factor, over a general class of stochastic search algorithms. While the Boltzmann distributions are well-known to have several nice properties [3], such as *maximum entropy*, it is not clear why these properties will be advantageous to annealing. (Instead, we are using the ratio of the standard deviation to the mean of the distributions.)

In general, the type of optimization we are considering is the following stochastic search procedure. We are trying to minimize the function $c \cdot X$ over convex set K described by a membership oracle $\mathcal{O}_K : \mathbb{R}^n \to \{0, 1\}$ that identifies whether a point is in the set K or not.

Stochastic-min$(\mathcal{O}_K, c, X_{\text{init}})$

- Let $f_0, f_1, \ldots, f_m : \mathbb{R} \to \mathbb{R}$ be a series of non-negative functions.
- X:=X_{init}
- For $i = 0, 1, \ldots, m$:

$$X := \text{Sample}(f_i(c \cdot x), \mathcal{O}_K, X)$$

- Return X

Here the Sample function samples a random point according to the density proportional to $f_i(c \cdot x)$ over the convex set K,

$$d\nu_i(x) = \frac{f_i(c \cdot x)}{\int_K f_i(c \cdot x)dx}, \text{ for } x \in K, \text{ and 0 otherwise.}$$

The sampling is typically implemented as a biased random walk and thus requires a starting point X.

We would like two things to happen. First, the means of the densities should hopefully approach the minimum quickly, i.e., $\mathsf{E}_{\nu_i}[c \cdot x] \longrightarrow \min_K c \cdot x$. Second, we would like the Sample routine to be efficient. For this we make the following assumptions:

1. **Logconcavity**: The density $d\nu_i$ to be sampled must be logconcave.

2. **Overlap**: The starting densities cannot change too quickly, which we formalize by saying that the total variation distance between two consecutive distributions must be less than $1 - 1/\text{poly}$:

$$\forall_i \|\nu_{i-1} - \nu_i\|_{tv} \leq 1 - \frac{1}{M}, \quad M \in \text{poly}(n)$$

The justification for the first assumption is that the most general known analyses of rapidly-mixing walks [1, 5, 15, 23] require the function to be logconcave (true in the statistics literature as well [6]).

As for the second, note that the variation distance of disjoint densities is 1. This assumption is much weaker than the current assumptions required by known analyses, the weakest of which [15] is a polynomial "expected warm start," $\|\nu_{i-1}/\nu_i\| < M$. Lemma 3.4 demonstrates that the variation distance condition above is weaker – it simply says that consecutive distributions must overlap on at least a $1/\text{poly}(n)$ fraction of their mass.

Based on these two assumptions, it is possible to show that functions of the form $e^{-\frac{c \cdot x}{T_i}}$ are optimal in the following sense.

Theorem 5.1 *Suppose we have a sequence of functions $f_1, f_2, \ldots, f_m : \mathbb{R} \to \mathbb{R}$ and their corresponding densities $d\nu_i(x) = f_i(x)/\int_K f_i(x)dx$, and suppose for all i: (a) f_i is logconcave, and (b) the variation distance $\|\nu_{i-1} - \nu_i\|_{tv} < 1 - \tau$, where $1/\tau \in poly(n)$. Then, for the cone optimization problem with $c = (1, 0, \ldots, 0), K = \{x \in \mathbb{R}^n : |x| \leq 2x_1 \leq 2\}$, for $m = \frac{\sqrt{n}}{2\ln(2e/\tau)} = \Omega(\sqrt{n})$ distributions,*

$$\mathsf{E}_{\nu_m}[c \cdot X] \geq \mathsf{E}_{\nu_1}[c \cdot X]/e.$$

The theorem says that after $\sqrt{n}$ phases, the mean along the axis can only drop by a factor of e towards the optimum (the origin) and thus any such stochastic search requires $\Omega(\sqrt{n})$ phases. Lemma 4.1 asserts that the Boltzmann distributions with a geometric temperature schedule achieve this lower bound using $O^*(\sqrt{n})$ phases.

Our proof proceeds by bounding the "spread" of distributions satisfying the assumptions. Let

$$\begin{aligned} \bar{\nu}_i &= \mathsf{E}_{\nu_i}[c \cdot X] \\ \sigma_i &= \sqrt{\mathsf{E}_{\nu_i}[(c \cdot X)^2] - (\mathsf{E}_{\nu_i}[c \cdot X])^2} = \sqrt{\mathsf{E}_{\nu_i}[(c \cdot X - \bar{\nu}_i)^2]} \end{aligned}$$

The next lemma shows that the rate of decrease of the means can be bounded in terms of the standard deviations and variation distances.

Lemma 5.2 *Suppose $\|\nu_{i-1} - \nu_i\|_{tv} \leq 1 - \frac{1}{M}$, for logconcave densities ν_i and ν_{i-1}. Then,*

$$\bar{\nu}_{i-1} - \bar{\nu}_i \leq (\sigma_i + \sigma_{i-1}) \ln(2eM).$$

Proof. By Lemma 5.17 of [15], for a random variable X drawn from logconcave ν_i and any $t > 1$,

$$P_{\nu_i}(c \cdot X > \bar{\nu}_i + t\sigma_i) < e^{-t+1}.$$

Let $\|\nu_{i-1} - \nu_i\|_{tv} = 1 - \frac{1}{M}$ and suppose for a contradiction that $\bar{\nu}_i < \bar{\nu}_{i-1} - (\sigma_i + \sigma_{i-1}) \ln 2eM$. Then,

$$P_{\nu_i}(c \cdot X > \bar{\nu}_i + \sigma_i \ln 2eM) < \frac{1}{2M}$$

and for Y drawn from ν_{i-1},

$$P_{\nu_{i-1}}(c \cdot Y \geq \bar{\nu}_{i-1} - \sigma_{i-1} \ln 2eM) \geq 1 - \frac{1}{2M}.$$

These two imply that the variation distance is less than $1 - \frac{1}{M}$, which is a contradiction. □

The next lemma is about the special case of optimizing over a cone along its axis. This simple case provides much intuition and is actually the worst case for annealing with logconcave functions.

Lemma 5.3 *Consider optimization over a cone, where* $c = (1, 0, 0, \ldots 0)$ *and,*

$$K = \{x \in \mathbb{R}^n : |x| \leq 2x_1 \leq 2\}$$

Then for any logconcave function $f_i : \mathbb{R} \to \mathbb{R}$ *and its corresponding normalized density* $\nu_i(x) = f_i(x_1)/\int_K f_i(x_1)dx$ *over* K*,*

$$\frac{\sigma_i}{\bar{\nu}_i} \leq \frac{1}{\sqrt{n}}$$

Proof. The statement of the lemma can be rewritten as,

$$\begin{aligned}
\frac{\mathsf{E}_{\nu_i}[x_1^2] - \bar{\nu}_i^2}{\bar{\nu}_i^2} &\leq \frac{1}{n} \\
\frac{\mathsf{E}_{\nu_i}[x_1^2]}{(\mathsf{E}_{\nu_i}[x_1])^2} &\leq 1 + \frac{1}{n} = \frac{n+1}{n} \\
\int_K x_1^2 \nu_i(x)dx \int_K \nu_i(x)dx &\leq \left(\frac{n+1}{n}\right)\left(\int_K x_1 \nu_i(x)dx\right)^2 \\
\int_0^1 x^{n+1} f_i(x)dx \int_0^1 x^{n-1} f_i(x)dx &\leq \left(\frac{n+1}{n}\right)\left(\int_0^1 x^n f_i(x)dx\right)^2
\end{aligned}$$

The last step comes from the fact that the volume of a cross section of K at x is proportional to x^{n-1}. Setting $f(x) = f_i(x)$ for $x \in [0, 1]$ and 0 elsewhere, we can rewrite this as

$$\left(\frac{1}{(n+1)!}\int_0^\infty x^{n+1} f(x)dx\right)\left(\frac{1}{(n-1)!}\int_0^\infty x^{n-1} f(x)dx\right) \leq \left(\frac{1}{n!}\int_0^\infty x^n f(x)dx\right)^2.$$

This follows from Lemma 5.3c of [15] states that the sequence

$$s_n = \frac{1}{n!}\int_0^\infty x^n f(x)dx$$

is logconcave, i.e., $s_{n+1}s_{n-1} \leq s_n^2$. □

The previous two lemmas lead to Theorem 5.1.

Proof. [of Theorem 5.1] Combining the previous two lemmas,

$$\begin{aligned}
\bar{\nu}_{i-1} - \bar{\nu}_i &\leq \frac{\bar{\nu}_{i-1} + \bar{\nu}_i}{\sqrt{n}} \ln(2eM) \\
\bar{\nu}_i\left(1 + \frac{\ln(2eM)}{\sqrt{n}}\right) &\geq \bar{\nu}_{i-1}\left(1 - \frac{\ln(2eM)}{\sqrt{n}}\right) \\
\frac{\bar{\nu}_i}{\bar{\nu}_{i-1}} &\geq 1 - 2\frac{\ln(2eM)}{\sqrt{n}}
\end{aligned}$$

Thus, after $m = 2\ln(2eM)/\sqrt{n}$ phases, the mean can have dropped by at most

$$\bar{\nu}_m \geq \bar{\nu}_1\left(1 - \frac{2\ln(2eM)}{\sqrt{n}}\right)^{\frac{\sqrt{n}}{2\ln(2eM)}} \geq \bar{\nu}_1/e$$

□

6 Extending to arbitrary convex functions

It would be nice to extend the approach to minimizing an arbitrary convex function, in the natural way. Namely, given a convex function f, the functions $e^{-f(x)/T_i}$ are logconcave, and would be a natural sequence of distributions to use for annealing. While results exist for mixing times of logconcave functions, sufficiently good results do not exist. However, in keeping with the previous analysis, the difficult point seems to be the "warm start" condition on the L_2 norms of successive distributions. This can still be bounded.

Lemma 6.1 *Let f be a convex function over convex set K with range $M = \max_K f(x) - \min_K f(x)$. Let $T_i = M\big(1 - \frac{1}{\sqrt{n}}\big)^i$, and π_i be the distribution with density proportional to $e^{-f(x)/T_i}$. Then for all $i \geq 0$,*

$$\|\pi_i/\pi_{i+1}\| \leq 5.$$

Proof. Let π_i be the stationary distribution of the ith distribution. As in Lemma 3.3,

$$Y(a) = \int_{\mathbb{R}^n} e^{-f(x)a}\, dx.$$

$$\begin{aligned} \|\pi_i/\pi_{i+1}\| &= \int \frac{e^{-f(x)/T_i}}{e^{-f(x)/T_{i+1}}} \frac{e^{-f(x)/T_i}}{Y(1/T_i)}\, dx \frac{e^{-f(x)/T_{i+1}}}{Y(1/T_i)} \\ &= \frac{Y(2/T_i - 1/T_{i+1})Y(1/T_{i+1})}{Y(1/T_i)^2} \\ &\leq 5. \end{aligned}$$

The last step uses the logconcavity of $a^n Y(a)$ by Lemma 3.3 and the computation is the same as in the proof of Lemma 4.6. □

Note (although it's not in the statement of the lemma), that the first distribution π_0 is close to the uniform distribution π over K, in that $\|\pi_0/\pi\| \leq e$, because the range of $f(x)/T_0$ is at most 1. Finally, a statement analogous to Lemma 4.1 holds also for general convex functions.

Lemma 6.1 can be interpreted as follows: any logconcave density can be morphed into a uniform density over a convex body in $O^*(\sqrt{n})$ steps, where the intermediate distributions are logconcave and have large overlap from one step to the next. In fact, we can further morph the uniform density over a convex body into an exponential density (proportional to $e^{-C|x|}$ for some C)in $O^*(\sqrt{n})$ iterations. This implies that one can go between any two logconcave densities in $O^*(\sqrt{n})$ steps.

7 Conclusions and future work

A theorem analogous to theorem 5.1, with a nearly identical proof applies to reverse annealing, where the temperature is increased. This shows that these distributions are also optimal for the volume algorithm in [14].

One criticism of our argument is that due to Step 3, the covariance estimation, our algorithm is really doing something more than annealing. However, on the n dimensional cone, where our upper and lower bounds match, Step 3 is not necessary (and so the algorithm actually has an $O(n^{3.5})$ guarantee). Thus, even without covariance updates, annealing is speeding things up.

While it seems that on other shapes, such as a cylinder, a much shorter temperature schedule (with Boltzmann distributions) may succeed, we do not have any examples where a non-Boltzmann type of distribution is strictly superior.

Finally, the original method of Bertsimas and Vempala can be used for minimizing arbitrary quasiconvex functions. It would be nice to extend annealing to this setting. We have presented some positive results in this direction.

References

[1] D. Applegate and R. Kannan: Sampling and integration of near logconcave functions, *Proceedings of the Annual Symposium on the Theory of Computing*, 156–163, 1990.

[2] D. Bertsimas and S. Vempala: Solving convex programs by random walks, *Journal of the ACM* **51**(4), 540–556, 2004.

[3] T. Cover and J. Thomas. *Elements of Information Theory.* John Wiley & Sons, New York, NY, 1991.

[4] A. Dinghas: Über eine Klasse superadditiver Mengenfunktionale von Brunn–Minkowski–Lusternikschem Typus, *Math. Zeitschr.* **68**, 111–125, 1957.

[5] A. Frieze, R. Kannan, and N. Polson: Sampling from logconcave distributions, *Annals of Applied Probability* **4**, 812–837, 1994; Correction: Sampling form logconcave distributions, ibid. **4**, 1255, 1994.

[6] W. Gilks and P. Wild: Adaptive Rejection Sampling for Gibbs Sampling. Applied Statistics, **41**(2), 337–348, 1992.

[7] L. Grotchel, L. Lovasz, and A. Schrijver, *Geometric algorithms and Combinatorial Optimization*, Springer-Verlag, Berlin, 1988.

[8] B. Hajek: Cooling schedules for optimal annealing, *Mathematics of Operations Research*, **13**, 311–329, 1988.

[9] M. Jerrum, A. Sinclair, and E. Vigoda: A polynomial-time approximation algorithm for the permanent of a matrix with non-negative entries, *Proceedings of the Annual Symposium on the Theory of Computing*, 712–721, 2001.

[10] M. Jerrum and G. Sorkin: Simulated annealing for graph bisection, *Proceedings of the Annual IEEE Symposium on Foundations of Computer Science*, 94–103, 1993.

[11] R. Kannan, L. Lovasz, and M. Simonovits: Random walks and an $O^*(n^5)$ volume algorithm for convex bodies. *Random Structures and Algorithms* **11**, 1–50, 1997.

[12] S. Kirkpatrick, C.D. Gelatt Jr, and M.P. Vecchi: Optimization by Simulated Annealing, *Science* **220**, 671—680, 1983.

[13] L. Leindler: On a certain converse of Hölder's Inequality II, *Acta Sci. Math. Szeged* **33**, 217–223, 1972.

[14] L. Lovász and S. Vempala: Simulated Annealing in Convex Bodies and an $O^*(n^4)$ Volume Algorithm, to appear in *J. Comp. Sys. Sci.*. Prelim. version in *Proceedings of the Annual IEEE Symposium on Foundations of Computer Science*, 660–669, 2003.

[15] L. Lovász and S. Vempala: The Geometry of Logconcave Functions and Sampling Algorithms, available at http://math.mit.edu/~vempala/papers/logcon.pdf. Preliminary version in *Proceedings of the Annual IEEE Symposium on Foundations of Computer Science*, 650–659, 2003.

[16] L. Lovász and S. Vempala: Hit-and-run from a corner, to apear in *SIAM J. on Computing.* Prelim. version in *Proceedings of the ACM Symposium on the Theory of Computing*, 310–314, 2004.

[17] W. Press, B. Flannery, S. Teukolsky and W. Vetterling: *Numerical Recipes in C: the Art and Science of Computing* (2nd edition). Cambridge University Press, Cambridge, 1992.

[18] A. Prékopa: Logarithmic concave measures and functions, *Acta Sci. Math. Szeged* **34**, 335–343, 1973.

[19] A. Prékopa: On logarithmic concave measures with applications to stochastic programming, *Acta Sci. Math. Szeged* **32**, 301–316, 1973.

[20] M. Rudelson: Random vectors in the isotropic position, *Journal of Functional Analysis* **164**, 60–72, 1999.

[21] G. Sorkin: Efficient simulated annealing on fractal energy landscapes, *Algorithmica*, **6**, 367–418, 1991.

[22] J. Spall: *Introduction to Stochastic Search and Optimization*, John Wiley & Sons, New Jersey, 2003.

[23] S. Vempala: Geometric Random Walks: A Survey, *MSRI volume on Combinatorial and Computational Geometry*. http://www-math.mit.edu/~vempala/survey.ps

[24] Z. B. Zabinsky, R. L. Smith, J. F. McDonald, H. E. Romeijn and D. E. Kaufman: Improving Hit-and-Run for Global Optimization, *Journal of Global Optimization* **3**, 171–192, 1993.

[25] Z. B. Zabinsky: *Stochastic Adaptive Search for Global Optimization*, Springer, 2003.

A Remaining proofs

Proof. [of Lemma 3.4] For any set S and probability distribution μ, let $\mu|_S$ be the distribution μ restricted to the set S, i.e. with density function $d\mu|_S(x) = d\mu(x)/\mu(S)$. Next, since the L_2 norm of any distribution is at least 1,

$$\begin{aligned}
\|\nu|_S/\pi|_S\| &\geq 1 \\
\mathsf{E}_{\nu|_S}\left(\frac{d\nu|_S(x)}{d\pi|_S(x)}\right) &\geq 1 \\
\mathsf{E}_{\nu|_S}\left(\frac{d\nu(x)}{d\pi(x)}\frac{\pi(S)}{\nu(S)}\right) &\geq 1 \\
\mathsf{E}_{\nu|_S}\left(\frac{d\nu(x)}{d\pi(x)}\right) &\geq \frac{\nu(S)}{\pi(S)}
\end{aligned}$$

Now consider the set $A = \{x | d\nu(x) \geq d\pi(x)\}$. Then the variation distance is $\|\nu - \pi\|_{tv} = \nu(A) - \pi(A)$. Using the above,

$$\begin{aligned}
\|\nu/\pi\| &= E_\nu\left(\frac{d\nu(x)}{d\pi(x)}\right) \\
&= \nu(A)E_{\nu|_A}\left(\frac{d\nu(x)}{d\pi(x)}\right) + (1-\nu(A))E_{\nu|_{A^c}}\left(\frac{d\nu(x)}{d\pi(x)}\right) \\
&\geq \nu(A)\frac{\nu(A)}{\pi(A)} + (1-\nu(A))\frac{1-\nu(A)}{1-\pi(A)} \\
&= 1 + \frac{(\nu(A)-\pi(A))^2}{\pi(A)(1-\pi(A))}.
\end{aligned}$$

Next, $\pi(A) = \nu(A) - \|\nu - \pi\|_{tv} \leq 1 - \|\nu - \pi\|_{tv}$. So in the case where $\|\nu - \pi\|_{tv} > 1/2$ (the lemma holds trivially in the other case), we have the better bound that $\pi(A)(1-\pi(A)) \leq (1 - \|\nu - \pi\|_{tv})\|\nu - \pi\|_{tv}$. Combining this with the previous displayed equation, gives,

$$\begin{aligned}
\|\nu/\pi\| &\geq 1 + \frac{\|\nu - \pi\|_{tv}^2}{(1 - \|\nu - \pi\|_{tv})\|\nu - \pi\|_{tv}} \\
&= \frac{1}{1 - \|\nu - \pi\|_{tv}}
\end{aligned}$$

Rearranging terms gives the lemma, which is tight for variation distances greater than $1/2$. (A tight bound for the $\leq 1/2$ case can be found by using $\pi(A)(1-\pi(A)) \leq 1/4$.) □

The next theorem, which follows from a theorem of Rudelson [20] and Theorem 5.22 from [15], is the basis of the covariance estimate. It was also used in [14, 2].

Theorem A.1 *Let f be any logconcave density in isotropic position. Let $x_1, \ldots, x_m$ be drawn independently from f and define $Y = \frac{1}{m}\sum_{i=1}^{m} xx^T$. Then, there is a constant C_1 such that for*

$$m > C_1 \frac{n}{\eta^2}(p \log \frac{n}{\eta^2})^2 \max\{p, \log n\},$$

$$\mathsf{E}\left(||Y - I||^p\right) \leq \eta^p.$$

As a consequence, we get the following guarantee about isotropy after each rounding update.

Corollary A.2 *Using $m > C_2 t^3 n \log^2 n$ samples, for $t > \log n$, the rounding step applied in phase i will put the current distribution μ_i in 2-isotropic position with probability at least $1 - \frac{1}{2^t}$.*

Proof. Without loss of generality, we can assume that μ_i is in isotropic position, and then show that for any unit vector v, the quantity $v^T Y v$ is between $1/2$ and 2 with high probability.

By Theorem A.1, for m large enough,

$$\begin{aligned} Pr(||Y - I|| > 2\eta) &= Pr(||Y - I||^p > (2\eta)^p) \\ &\leq \frac{\mathsf{E}(||Y - I||^p)}{(2\eta)^p} \\ &\leq \frac{1}{2^p}. \end{aligned}$$

Thus, if we set $\eta = 1/4$ and $p = t$, for the corresponding value of m from the theorem, we have

$$v^T Y v = v^T (I + Y - I) v \leq 1 + ||Y - I|| \leq \frac{3}{2}$$

with probability at least $1 - \frac{1}{n^t}$ as required. The lower bound is similar. □