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 $x(\theta')$.

$$\begin{aligned} -x(\theta') &= \int_{\mathcal{K}} x\, dP_{\theta'} = \int_{\mathcal{K}} x \frac{dP_{\theta'}(x)}{dP_\theta(x)} dP_\theta \\ &= \mathop{\mathbb{E}}_{X\sim P_\theta}\left[X \exp(X^\top(\theta-\theta') + A(\theta') - A(\theta))\right] \end{aligned}$$

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. We can get a more precise estimate of $-x(\theta')$ as

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

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 (Lovász & Vempala, 2006):

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 $m = O(\frac{n^2S^2}{s^2}\log^5\frac{nM}{\tau})$ 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 $\|v\| = 1$,

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

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 $\max\left\{\left\|\frac{f}{g}\right\|_2, \left\|\frac{g}{f}\right\|_2\right\} \le D$, 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 (Kalai & Vempala, 2006).

□

C. Proof of Lemma 5

Proof. We first show by elementary linear algebra that

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

Let us consider the log of the 2-norm:

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

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

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

Lemma 1 implies $\lambda(x(\theta),1+\gamma) \le (1+c)\lambda(x(\theta),1)+c = c \le \frac{1}{4}$. We start by applying Lemma 4,

$$\begin{aligned}D_A((1+\gamma)\theta,\theta) &\le 2\|x(\theta)-x((1+\gamma)\theta)\|^2_{x((1+\gamma)\theta)} \\ \text{(2.28) in (Nemirovski, 1996)} &\le 2\left(\frac{\lambda(x(\theta),1+\gamma)}{1-\lambda(x(\theta),1+\gamma)}\right)^2 \\ \text{Lemma 1} &\le 2\left(\frac{c}{1-c}\right)^2 < 2\frac{(1/3)^2}{(2/3)^2} = \frac{1}{2}\end{aligned}$$

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):

$$\begin{aligned}D_A((1-\gamma)\theta,\theta) &= D_A(\tilde\theta,\tilde\theta(1+\tilde\gamma)) \\ \text{(Lemma 4)} &\le 2\|x(\tilde\theta)-x((1+\tilde\gamma)\tilde\theta)\|^2_{x(\tilde\theta)} \\ \text{((2.27) in Nemirovski (1996))} &\le 2\left(\frac{\lambda(x(\tilde\theta),1+\tilde\gamma)}{1-\lambda(x(\tilde\theta),1+\hat\gamma)}\right)^2 \\ \text{(Lemma 1)} &\le 2\left(\frac{c}{1-c}\right)^2 < 2\frac{(1/3)^2}{(2/3)^2} = \frac{1}{2}\end{aligned}$$

It follows that

$$\begin{aligned}\|P_\theta/P_{(1+\gamma)\theta}\| &\le e^{D_A((1+\gamma)\theta,\theta)+D_A((1-\gamma)\theta,\theta)} \\ &\le e^{\frac{1}{2}+\frac{1}{2}} \le 10.\end{aligned}$$

□

D. 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 (Nemirovski, 1996) allows one to replace the inverse hessian $\nabla^{-2}A^*(x)$ with the nearby $\nabla^{-2}A^*(x(\theta)) = \mathrm{CovMtx}(P_\theta)$. 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, (Adamczak et al., 2010) on the concentration of empirical covariance matrices.

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

$$\theta(x) = \arg\max_\theta \theta \cdot x - \log \int_K \exp(-\theta \cdot y) dy = \nabla A^*(x) \tag{17}$$

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.

E. Some history on the entropic barrier and the universal barrier for cones

Let K be a cone in $\mathbb{R}^n$ and let $K^* = \{\theta : \theta^\top x \geq 0 \; \forall x \in K\}$ 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 $K^*(x) := \{y \in K^* : x^\top y \leq 1\}$. Nesterov & Nemirovskii (1994) defined the first generic self-concordant barrier function, known as the *universal barrier* in terms of these regions. Namely, they show that the function

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

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 (1996) 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 (1996) 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 (1996), $A^*_-(\cdot)$ is (up to additive constant) the Fenchel conjugate of the universal barrier u_{K^*} for K^*. It was shown by Nesterov & Nemirovskii (1994) (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 & Eldan (2014) 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* $K(U) := \{\alpha(x, 1) : x \in U, \alpha \geq 0\} \subseteq \mathbb{R}^{n+1}$. 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 (1996) 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 $F_K(\alpha(x, 1)) \cong F^*_{K^*}(\alpha(x, 1)) \cong -\frac{n+1}{2} \log(\alpha^2(1 - \|x\|^2))$. 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.