Statistics 291: Lecture 15 (March 19, 2024)

Subag's Optimization Algorithm and Ultrametricity

Instructor: Mark Sellke

Scribe: Kenny Gu

1 Introduction

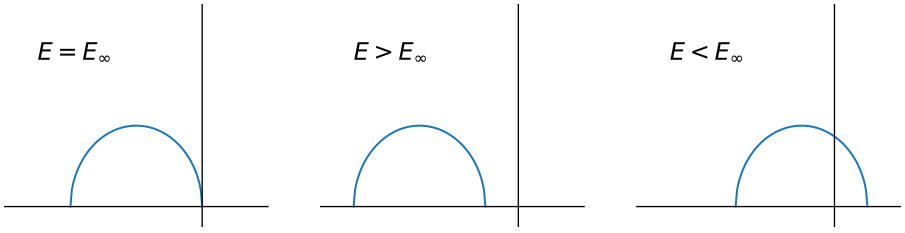
So far, we've primarily been looking at spherical spin glasses of the form $H_{N,p}(x) = N^{-(p-1)/2}\langle G_N^{(p)}, x^{\otimes p}\rangle$. We will now begin to explore algorithms for optimizing these objects. Next lecture, we will also explore optimization techniques for large independent sets on sparse random graphs.

Earlier in the course, the Kac-Rice formula suggested $E_\infty = 2\sqrt{\frac{p-1}{p}}$ as a natural threshold for optimization problems. Recall that

- If $E < E_\infty$, then $\mathbf{E}\left|\text{LocalMax}_{S_N}(H_{N,p}; (-\infty, E])\right| \le e^{-cN^2}$
- If $E > E_\infty$, then most critical points are local maxima in expectation. That is,

$$\mathbf{E}\left|\operatorname*{Crt}_{S_N}(H_{N,p}; [E,\infty))\right| = (1+o(1))\mathbf{E}\left|\operatorname*{LocalMax}_{S_N}(H_{N,p}; [E,\infty))\right|$$

To derive this threshold, we computed the following: conditioning on $\nabla_{\text{sph}} H_{N,p}(x) = 0$ and $H_{N,p}(x)/N = E$, we have $\nabla^2_{\text{sph}} H_{N,p}(x) \overset{d}{=} \sqrt{p(p-1)}\text{GOE}(N-1) - pEI_{N-1}$. By the Wigner semicircle law calculations, the distributions looked like:



In the case of $E < E_\infty$, we see that we are unlikely to have a local maximum, while for $E > E_\infty$, we probably have a local maximum.

We might guess that gradient flow reaches this threshold E_∞, which might be optimal. However it is highly unclear: one could imagine that gradient flow gets stuck in a saddle and hence fails to reach any local maxima, or it could avoid local maxima altogether and outperform E_∞. Later we will prove that the latter cannot happen, at least on dimension-free time scales. Today, we will explore a somewhat strange algorithm that achieves E_∞ without getting stuck.

2 Greedy Hessian ascent

Greedy Hessian ascent, as introduced in Subag (2018), works by (1) initializing x_0 at the origin and (2) taking orthogonal steps on the tangent space of subspheres at each step in a greedy way until we reach S_N.

More formally, we fix some $\delta > 0$ small and let $m = 1/\delta \in \mathbb{Z}_+$. For $k = 0, 1, \dots, m-1$:

(1) Let v^k be the top unit eigenvector of $\nabla^2_{\text{tan}} H_{N,p}(x^k) = P^{\perp}_{x^k} \nabla^2 H_{N,p}(x^k) P^{\perp}_{x^k}$ such that $v^k \perp x^k$.

(2) Set $x^{k+1} = x^k \pm \sqrt{N\delta}\, v^k$, choosing $\pm$ such that $\langle x^{k+1} - x^k, \nabla_{\text{sph}} H_{N,p}(x^k) \rangle \geq 0$.

Theorem 2.1. *For any $\epsilon > 0$, $\lim_{\delta \to 0} \lim_{N \to \infty} \mathbb{P}\left(H_{N,p}(x_m)/N \geq 2\sqrt{\frac{p-1}{p}} - \epsilon \right) = 1$.*

Proof. (incorrect cheating proof) In our first attempt at a proof, we'll cheat a little bit: we'll see afterward why this "proof" is not correct and propose an alternative correct proof.

Recall that for any fixed $x \in S_N$, $\nabla^2_{\text{tan}} H_{N,p}(x) \overset{d}{=} \sqrt{p(p-1)\frac{N-1}{N}}\, GOE(N-1)$. Similarly, for $\|x\| = \sqrt{Nq}$, $\nabla^2_{\text{tan}} H_{N,p}(x) \overset{d}{=} \sqrt{p(p-1)q^{p-2}}\, GOE(N-1)$, which implies $\lambda_{\max}(\nabla^2_{\text{tan}} H_{N,p}(x^k)) \approx 2\sqrt{p(p-1)(k\delta)^{p-2}}$. We can then Taylor expand

$$H_{N,p}(x^{k+1}) = H_{N,p}(x^k) + \langle \nabla_{\text{sph}} H_{N,p}(x^k), x^{k+1} - x^k \rangle + \frac{1}{2} \langle \nabla^2_{\text{sph}} H_{N,p}(x^k), (x^{k+1} - x^k)^{\otimes 2} \rangle + O\left(\|x^{k+1} - x^k\|^3 \cdot \sup_{\|y\| \leq \sqrt{N}} \|\nabla^3 H_{N,p}(y)\|_{\text{op}} \right)$$

where we have

- $\langle \nabla_{\text{sph}} H_{N,p}(x^k), x^{k+1} - x^k \rangle \geq 0$ by construction
- $\frac{1}{2} \langle \nabla^2_{\text{sph}} H_{N,p}(x^k), (x^{k+1} - x^k)^{\otimes 2} \rangle \approx \frac{N\delta}{2} \lambda_{\max}(\nabla^2_{\text{tan}} H_{N,p}(x^k)) \approx N\delta\sqrt{p(p-1)(k\delta)^{p-2}}$
- $O\left(\|x^{k+1} - x^k\|^3 \cdot \sup_{\|y\| \leq \sqrt{N}} \|\nabla^3 H_{N,p}(y)\|_{\text{op}} \right) = O(\delta^{3/2} N)$

Together, we have

$$H_{N,p}(x^{k+1}) - H_{N,p}(x^k) \geq N\delta\sqrt{p(p-1)(k\delta)^{p-2}} + O(\delta^{3/2} N)$$

so we can use a telescoping sum to write

$$\begin{aligned} H_{N,p}(x_m) &= \sum_{k=0}^{m-1} (H_{N,p}(x^{k+1}) - H_{N,p}(x^k)) \\ &\overset{\delta \to 0}{\geq} N(1-\epsilon) \int_0^1 \sqrt{p(p-1)q^{p-2}}\, dq - O(\delta^{1/2} N) \\ &= N(1-\epsilon) \cdot 2\sqrt{\frac{p-1}{p}}. \end{aligned}$$

□

The key to the above "proof" was the distribution of $\nabla^2_{\text{tan}} H_{N,p}(x)$. The proof fails because this fact only holds for fixed x independent of the disorder, but x^k necessarily depends on $H_{N,p}$. One of the main insights of Subag (2018) is that this dependency doesn't end up being too much of an issue, though we need to adapt our proof to handle this dependency.

The following lemma is the key to completing the correct proof.

Lemma 2.2. *For any $\epsilon > 0$, with probability $1 - e^{-N}$ for large N, for all $x \in \mathbf{R}^N$ with $\|x\| \in [1, \sqrt{N}]$* simultaneously, $\lambda_{\max}(\nabla^2_{tan} H_{N,p}(x^k)) \geq 2\sqrt{p(p-1)R(x,x)} - \epsilon$.

Proof. Intuitively, for $\lambda_{\max}(GOE(N)) \le 2-\epsilon$, we need a constant fraction of the eigenvalues to move. Let $A_j = \frac{1}{j}\sum_{i=1}^{j}\lambda_i(M_n)$ for $M_n \sim GOE(n)$. Recall from Lecture 7 that the Hoffman-Wielandt inequality implies that $\mathbb{P}(|A_j - \mathbf{E}A_j| \ge t) \le e^{-Njt^2/1000}$. Therefore, for small δ depending on ϵ, we have the implication

$$\mathbf{E}[A_{\delta N}] \ge 2-\epsilon \implies \mathbb{P}(A_{\delta N} \ge 2-2\epsilon) \ge 1 - e^{-\delta\epsilon^2 N^2/1000}$$

We then proceed by an ϵ-net argument. Letting $\mathcal{N}$ be a N^{-10}-net for $\{x \in \mathbf{R}^N : \|x\| \in [1,\sqrt{N}]\}$, we know $|\mathcal{N}| \le N^{100N}$ and the set of points in this ϵ-net is independent of $H_{N,p}$. Thus the GOE hessian description works for all $y \in \mathcal{N}$, so by a union bound and the above argument:

$$\begin{aligned}\mathbb{P}(\forall y \in \mathcal{N}, \lambda_{\max}(\nabla^2_{\tan} H_{N,p}(y)) \ge 2\sqrt{p(p-1)R(y,y)} - \epsilon) &\ge 1 - N^{100N}\cdot e^{-\delta\epsilon^2 N^2/1000} \\ &\ge 1 - e^{-\delta\epsilon^2 N^2/2000}.\end{aligned}$$

We can conclude the proof by extending this event to all $\|x\| \in [1,\sqrt{N}]$ by rounding to the nearest $y \in \mathcal{N}$. Here we use the high-probability bounds on $\sup_{\|x\|\le\sqrt{N}} \|\nabla^3 H_{N,p}(x)\|$ to control the rounding error. The requirement that $\|x\| \ge 1$ ensures that rounding does affect the orthogonal projection matrix $P_x^{\perp}$ too much (which appears in the definition of tangential Hessian). □

3 Mixed models and ultrametricity

For a mixed model $H_N = \sum_p \gamma_p H_{N,p}$, recall that we have $\xi(q) = \sum_p \gamma_p^2 q^p$. In fact, the same algorithm as above results in performance $\mathrm{ALG}(\xi) = \int_0^1 \sqrt{\xi''(q)}\,dq$.

Theorem 3.1 (Chen-Sen)**.** *The following are equivalent:*

- $\lim_{N\to\infty} \mathbf{E}\left[\max_{x\in S_N} \frac{H_N(x)}{N}\right] = ALG(\xi)$.
- $\gamma_1 = 0$ *and* $\gamma''(q)^{-1/2}$ *is concave on* $[0,1]$ *— i.e.,* ξ *is mostly quadratic.*
- ξ *is full RSB at* 0 *temperature.*

If we were to use the top two unit eigenvectors during each iteration of Subag's algorithm, we end up with an orthogonally branching tree. In full RSB, for low temperature, μ_β looks like this complete binary tree in the sphere in some sense. On the other hand, from k-RSB, if we think of the measure as being generated by such a tree, we are limited to k "levels of branching."

The following theorem helps formalize this connection between RSB and trees.

Theorem 3.2. *(Chen-Panchenko). Suppose* $\gamma_1 = 0$ *and* $\gamma_p > 0$ *for* $p \ge 2$*. Fix* $k > 0$ *and let* $x^1,\dots,x^k \overset{iid}{\sim} \mu_\beta$*. Consider the* $k\times k$ *array* $M_{i,j}^{N,k} = \|x^i - x^k\|/\sqrt{N} \in [0,2]$*. As* $N\to\infty$*, Law*$(M_{i,j}^{N,k})$ *converges to* $M^{\infty,k}$ *which is almost surely ultrametric (i.e.,* $M_{i,j}^{\infty,k} \le \max(M_{i,\ell}^{\infty,k}, M_{\ell,k}^{\infty,k})$ *for any* ℓ *a.s.).*

Above, we required $\gamma_p > 0$ for $p \ge 2$ (we call such a ξ "generic"). We need this condition to break symmetry, though even very small $\gamma_p \sim N^{-0.1}$ suffices.

Ultrametrics and trees have a two-sided connection. On one end, for T a finite tree with positive edge weights, the distance between leaves is an ultrametric. On the other end, all finite ultrametrics correspond to a tree structure: for all $t > 0$, $x \sim y$ if $\mathrm{dist}(x,y) \le t$ is an equivalence relation. Furthermore, smaller values of t correspond to refined partitions. Therefore, we can directly build a tree according to these partitions.

For large β (i.e., low temperatures), Law$(M_{1,2}^{\infty})$ has full RSB behavior at low temperatures; it will depend on ξ. In this case, one has "full" support $[0,1-\epsilon]$ for the overlap, with $\epsilon\to 0$ as $\beta\to\infty$.

Next time, we will use the overlap gap property to show the failure of algorithms.