Statistics 291: Lecture 14 (March 7, 2024)

Shattering III

Instructor: Mark Sellke

Scribe: Alan Chung

1 Introduction

Today, we show that shattering implies that there cannot exist a stable sampling algorithm. A sampling algorithm A is defined as a function

$$A(H_{N,p}, \omega) \to \sigma \in \mathbb{R}^N \text{ (or } S^N),$$

where ω is some randomness independent from the Hamiltonian $H_{N,p}$. In the case of Langevin Dynamics, this randomness would be the initialization and Brownian motion $(x_0, B_{[0,T]})$. Denote the distribution of samples that A outputs, conditional on the Hamiltonian, as $\text{Law(A;H}_{\text{N,p}}) = \text{Law(A(H}_{\text{N,p}}, \omega))|\text{H}_{\text{N,p}})$.

Definition 1.1 (δ-approximate sampler)**.** We say that A is a δ-approximate sample for μ_β (which, in our contexts, will be a gibbs measure), if

$$\mathbf{E}_{H_{N,p}}\left[\frac{1}{\sqrt{N}} W_2(\text{Law(A;H}_{\text{N,p}}), \mu_\beta)\right] \le \sqrt{\delta}.$$

Here, the expectation is taken over the randomness in the hamiltonian.

Essentially, this is saying is that the Wasserstein distance is small on average for most hamiltonians, though potentially is large for some set of "bad" hamiltonians. Note that this is a relatively weak requirement for a sampling algorithm, i.e. it is much less stringent than demanding small total variation error.

2 Stability

In this section, we discuss what it means for an algorithm to be stable and consider a few examples.

Let $H_{N,p}$ and $\widetilde{H}_{N,p}$ be two i.i.d. p-spin Hamiltonians. Then, define the perturbation

$$H^\epsilon_{N,p} := \sqrt{1-\epsilon} H_{N,p} + \sqrt{\epsilon}\widetilde{H}_{N,p}.$$

We let μ_β be the Gibbs measure for $H_{N,p}$ and μ^ϵ_β be the Gibbs measure for $H^\epsilon_{N,p}$.

Definition 2.1 (Stable)**.** We say that a sequence of algorithms (A_N) is **stable** if

$$\lim_{\epsilon \downarrow 0} \limsup_{N\to\infty} \mathbf{E}_{H_{N,p}}\left[\frac{1}{\sqrt{N}} W_2(\text{Law(A;H}_{\text{N,p}}), \text{Law(A, H}_{\text{N,p}})\right] = 0.$$

Definition 2.2 (Strongly Stable)**.** We say that a sequence of algorithms (A_N) is **strongly stable** if

$$\lim_{\epsilon\downarrow 0}\limsup_{N\to\infty}\mathbf{E}_{H_{N,p},\omega}\left[\frac{1}{\sqrt{N}}\left\|A(H_{N,p},\omega)-A(H_{N,p},\omega)\right\|\right]=0$$

Proposition 2.3. *Let x_T be the output of spherical Langevin Dynamics. Then x_T is strongly stable for any fixed T.*

Proof. We have

$$d(x_t-x_t^\epsilon)=\beta\left(\nabla_{sph}H_{N,p}(x_t)-\nabla_{sph}H_{N,p}^\epsilon(x_t^\epsilon)\right)dt-\frac{N-1}{N}\left(x_t-x_t^\epsilon\right)dt+\sqrt{2}\left(P_{x_t}^\perp-P_{x_t^\epsilon}^\perp\right)dB_t$$

If $H_{N,p}$ and $\widetilde{H}_{N,p}$ are C-bounded, then Ito's formula yields

$$\frac{d}{dt}\underbrace{\mathbf{E}\left[\frac{1}{N}\|x_t-x_t^\epsilon\|_2^2\right]}_{f(t)}\le C'\mathbf{E}\left[\frac{1}{N}\|x_t-x_t^\epsilon\|_2^2+\epsilon\right].$$

For this f defined above, note $f(0)=0$. Then, Gronwell's lemma yields

$$f'(t)\le C'(f(t)+\epsilon)\Rightarrow f(t)\le C'\epsilon te^{C't}.$$

This suffices for the proof, since this bound is N-independent. □

Corollary 2.4. *For $\beta\le\beta_0(p)$, we have that*

$$\lim_{\epsilon\downarrow 0}\limsup_{N\to\infty}\frac{1}{\sqrt{N}}W_2(\mu_\beta,\mu_\beta^\epsilon)=0.$$

In light of this, we say that μ_β is stable.

Proof. Consider

$$W_2(\mu_\beta,\mu_\beta^\epsilon)\le W_2(\mu_\beta),\mathrm{Law(x_T)})+\mathrm{W_2(Law(x_T),Law(x_T^\epsilon))}+\mathrm{W_2(Law(x_T^\epsilon),}\mu_\beta^\epsilon).$$

We have seen that the first and third terms decay at the rate $Ce^{-T/c}\sqrt{N}$ with high probability. Then, we've bounded the second term by $F(t)\sqrt{\epsilon N}$. Then, we can conclude the proof by taking T large, and ϵ small (depending on T). (Informally, $\epsilon\ll\frac{1}{T}\ll 1$.) □

In fact, what we've shown is that if there exists a stable δ-approximate sampler for every $\delta>0$, then μ_β is stable (by taking $\epsilon\ll\delta\ll 1$). We can also consider the contrapositive, that is, if μ_β is unstable, then there does not exist a stable δ-approximation for some $\delta>0$. We will show that shattering implies that μ_β is unstable, which will then imply that there does not exist a stable sampler.

Theorem 2.5. *Let $\beta\in\left[C,\sqrt{\left(\frac{1}{2}-o(1)\right)\log(p)}\right]$ (so that shattering occurs), and let p be odd (this is just for convenience; we show how to extend this to all p below). Then,*

$$\liminf_{\epsilon\downarrow 0}\liminf_{N\to\infty}\mathbf{E}\left[\frac{1}{\sqrt{N}}\mathbf{E}[W_2(\mu_\beta,\mu_\beta^\epsilon)]\right]>0.$$

Let $\{\mathscr{C}_m\}_{m=1}^M$ denote the set of shattering clusters for $H_{N,p}$. The idea of the proof is the following two steps:

A) If ϵ is small, then $\{\mathscr{C}_m\}_{m=1}^M$ is also a valid shattering decomposition for μ_β^ϵ.

B) The weights of the clusters change drastically, in that $\mu_\beta(\mathscr{C}_m)$ and $\mu_\beta^\epsilon(\mathscr{C}_m)$ are completely different.

2.1 Proving A)

Suppose that $H_{N,p}$ and $\widetilde{H}_{N,p}$ are C-bounded. Then $\|H_{N,p} - H^{\epsilon}_{N,p}\| \le O(CN\sqrt{\epsilon})$. This implies that for every $S \subseteq S_N$, we have

$$e^{-O(\beta CN\sqrt{\epsilon})} \le \frac{\mu_\beta(S)}{\mu^{\epsilon}_\beta(S)} \le e^{O(\beta CN\sqrt{\epsilon})}.$$

This implies that if $\epsilon \ll \frac{c}{\beta C}$, the same shattering decomposition is valid for μ^{ϵ}_β. (Just check the defining properties.)

2.2 Proving B)

Next, suppose we view the clusters as point masses, so that we are looking at a distribution over a discrete set of objects. In particular, supposing that there are M clusters, and that we label the clusters $\{1,2,\dots,M\}$, define

$$\hat{\mu}_\beta(m) = \frac{\mu_\beta(\mathscr{C}_m)}{\sum_{m=1}^M \mu_\beta(\mathscr{C}_m)}, \quad \hat{\mu}^{\epsilon}_\beta(m) = \frac{\mu^{\epsilon}_\beta(\mathscr{C}_m)}{\sum_{m=1}^M \mu^{\epsilon}_\beta(\mathscr{C}_m)}. \tag{1}$$

We now state the following lemma. Here s is the separation parameter between clusters in the shattering decomposition, and the last exponential term captures the total mass not included in the clusters.

Lemma 2.6.

$$\mathbf{E}\left[\frac{1}{N} W_2(\mu_\beta, \mu^{\epsilon}_\beta)^2\right] \ge s^2 \mathbf{E}\left[d_{\mathrm{TV}}(\hat{\mu}_\beta, \hat{\mu}^{\epsilon}_\beta)\right] - e^{-cN/10}.$$

Proof Outline. Note that in (1), the denominators are each at least $1 - e^{-cN/5}$ by construction. Aside from this tiny error, any coupling between μ_β and μ^{ϵ}_β directly induces a coupling between $\hat{\mu}_\beta$ and $\hat{\mu}^{\epsilon}_\beta$, and thinking about this coupling gives the bound. □

We now try to lower bound $\mathbf{E}\left[d_{\mathrm{TV}}(\hat{\mu}_\beta, \hat{\mu}^{\epsilon}_\beta)\right]$. We would like to show something of the form

$$\mathbb{P}\left(d_{\mathrm{TV}}(\hat{\mu}_\beta, \hat{\mu}^{\epsilon}_\beta) \ge 0.001)\right) \overset{?}{\ge} 0.01.$$

If this is indeed true, then observe that

$$d_{\mathrm{TV}}(\hat{\mu}_\beta, \hat{\mu}^{\epsilon}_\beta) = \frac{1}{2}\sum_m \left|\hat{\mu}_\beta(m) - \hat{\mu}^{\epsilon}_\beta(m)\right| = \frac{1}{2}\mathbf{E}_{m\sim\hat{\mu}_\beta}\left[\left|\frac{\hat{\mu}^{\epsilon}_\beta(m)}{\hat{\mu}_\beta(m)} - 1\right|\right].$$

By the Markov inequality, we conclude that if $d_{\mathrm{TV}}(\hat{\mu}_\beta, \hat{\mu}^{\epsilon}_\beta) \le 0.001$, then

$$\mathbb{P}_{m\sim\hat{\mu}_\beta}\left(\frac{\hat{\mu}^{\epsilon}_\beta(m)}{\hat{\mu}_\beta(m)} \in [0.9, 1.09]\right) \ge 0.9 \Rightarrow \mathbb{P}_{m_1,m_2\sim\hat{\mu}_\beta}\left(\frac{\hat{\mu}^{\epsilon(m_1)}_\beta}{\hat{\mu}^{\epsilon}_\beta(m_2)} \in [0.8, 1.2]\frac{\hat{\mu}_\beta(m_1)}{\hat{\mu}_\beta(m_2)}\right) \ge 0.8$$

It suffices to argue that this above probability inequality is false, because then $d_{\mathrm{TV}}(\hat{\mu}_\beta, \hat{\mu}^{\epsilon}_\beta) > 0.001$, which would suffice for the proof.

The final remaining point is the following anti-concentration statement for probability ratios under $\hat{\mu}^{\epsilon}_\beta$.

Lemma 2.7. *Given $H_{N,p}$ (with a valid shattering decomposition) and distinct clusters $m_1 \ne m_2$, and any $a, b \in \mathbb{R}$, we have*

$$\mathbb{P}_{\widetilde{H}_{N,p}}\left(\log\left(\frac{\hat{\mu}^{\epsilon}_\beta(m_1)}{\hat{\mu}^{\epsilon}_\beta(m_2)}\right) \in [a,b] \middle| H_{N,p}, \widetilde{G}^{(p)}_N\right) \le O\left(\frac{b-a}{\sqrt{N\epsilon}}\right)$$

Noting that the clusters have a separation of $s\sqrt{N}$ and are spherical caps, we can find a vector v with norm $\|v\| = \sqrt{N}$ such that for any $x_1 \in \mathscr{C}_{m_1}$ and $x_2 \in \mathscr{C}_{m_2}$:

$$\langle x_1, v\rangle \geq \langle x_2, v\rangle + s\sqrt{N}^{\,p \text{ odd}} \Rightarrow \langle x_1, v\rangle^p \geq \langle x_2, v\rangle^p + \left(\frac{sN}{2}\right)^p. \tag{2}$$

(Recall that p is odd in the statement of Theorem 2.5.)

The goal is to show that $\log\left(\hat{\mu}^\epsilon_\beta(m_1)/\hat{\mu}^\epsilon_\beta(m_2)\right)$ anti-concentrates. The main idea is to show that its derivative with respect to the underlying (gaussian) randomness is very large. This will imply that the above quantity is a function of a gaussian whose derivative grows very large, which would imply anticoncentration.

Let $\widetilde{G}^{(p)}_N$ be the p-tensor corresponding to $\widetilde{H}^p_N$. We can write

$$\widetilde{G}^{(p)}_N = Z v^{\otimes p} + \check{G}^{(p)}_N$$

such that $\langle v^{\otimes p}, \check{G}^{(p)}_N\rangle = 0$, i.e., we consider the decomposition of the tensor $\widetilde{G}^{(p)}_N$ in the direction v. Hence $\widetilde{H}_{N,p}(x) = \widetilde{Z}\langle v, x\rangle^p + H_{N,p}(x)$ for some Gaussian random variable $\widetilde{Z}$. Rescaling, we can write $g = \widetilde{Z}\cdot N^{p-1/2} \sim \mathscr{N}(0,1)$, noting that $\widetilde{H}_{N,p}(v) \sim \mathscr{N}(0,N) \Rightarrow \widetilde{Z} \sim \mathscr{N}(0, N^{1-2p})$. Now defining,

$$Y(\widetilde{Z}) = \log\left(\frac{\hat{\mu}^\epsilon_\beta(m_1)}{\hat{\mu}^\epsilon_\beta(m_2)}\right),$$

it follows from (2) that

$$Y'(\widetilde{Z}) \geq \frac{\beta s^p N^p \sqrt{\epsilon}}{2^p}, \quad \forall \widetilde{Z} \in \mathbb{R}.$$

This implies that $\frac{dY}{dg} \geq \frac{\beta s^p}{2^p}\sqrt{N\epsilon}$. Hence, this function Y is a function of a standard Gaussian, but the derivative of Y grows arbitrarily large, which implies anticoncentration.

In the above, we assumed that p is odd so that an inequality was preserved by monotonicity. When p is even, one can slightly "redefine" the shattering by pairing up antipodal clusters on the sphere, since the Gibbs measure μ_β is invariant under this symmetry. Then in Lemma 2.7, one requires that $m_1 \neq m_2$ and that the clusters are not antipodal, and needs to be slightly more ad-hoc to construct a good direction v.

To conclude these notes, we relate the content in this paper to some notions from other papers. In particular, [AMS23] calls the condition

$$\liminf_{\epsilon\downarrow 0}\liminf_{N\to\infty} W_2(\mu_\beta, \mu^\epsilon_\beta) > 0$$

transport disorder chaos. A theorem by Chatterjee proves that for general mixed p-spin models without external field (at low or high temperature), there is "disorder chaos". Namely if $x \sim \mu_\beta$ and $x^\epsilon \sim \mu^\epsilon_\beta$ are independent Gibbs samples, then

$$\text{plim}_{N\to\infty} \mathrm{R(x, x^\epsilon)} = 0.$$

In the RSB setting, when $\beta > \beta_c$, then it is not true: for $x, x' \overset{IID}{\sim} \mu_\beta$, we have $\mathbf{E}[R(x,x')^2] \not\to 0$. In the RSB phase of low temperature, disorder chaos implies transport chaos, which implies stable sampling is impossible under RSB. However in the RS phase, transport disorder chaos is a more informative notion (and might or might not hold depending on β).

References

[AMS23] Ahmed El Alaoui, Andrea Montanari, and Mark Sellke. Shattering in pure spherical spin glasses, 2023. 4