# Statistics 291: Lecture 4 (February 1, 2024) Geometric and statistical consequences of annealed free energy

Instructor: Mark Sellke

Scribe: Zad Chin

## 1 A Recall on Main Formulas

From the last few classes, we defined the following:

(a)  $H_{N,p}(x) = N^{-(p-1)/2} \sum_{i_1,\dots,i_p=1}^N g_{i_1\dots i_p} x_{i_1} \cdots x_{i_p} = N^{-(p-1)/2} \langle G_N^{(p)}, x^{\otimes p} \rangle$ 

(b) 
$$Z_N(\beta) = \int_{S_N} e^{\beta H_{N,p}(x)} dx$$

(c) 
$$F_N(\beta) = \frac{1}{N} \log Z_N(\beta)$$

(d) 
$$d\mu_{\beta}(x) = \frac{e^{\beta H_{N,p}(x)}}{Z_N(\beta)}, \mu \sim \text{Unif}(S_N)$$

### **2** Concentration of Measure (Continued)

Recall that in last lecture, we had the following lemma.

**Lemma 2.1.** If  $F : \mathbb{R}^d \to \mathbb{R}$  is *L*-Lipshitz and  $G \sim \mathcal{N}(0, \mathbb{I}_d)$ , then

$$\mathbb{P}[|F(G) - \mathbb{E}F(G)| \ge \lambda] \le 2e^{-\lambda^2/8L^2}.$$

In particular, the above probability is small once  $\lambda \gg L$ .

*Proof.* By smoothing, e.g. using convolution, we may assume  $F \in C^1(\mathbb{R}^d)$ . Now, we will use the interpolation method: let

$$G_0, G_{\pi/2} \sim \mathcal{N}(0, \mathbb{I}_d)$$

be i.i.d. standard gaussians. Consider the "path" from  $G_0$  to  $G_{\pi/2}$  given by

$$G_{\theta} = \cos(\theta)G_0 + \sin(\theta)G_{\pi/2}.$$

Let  $\tilde{G}_{\theta} := \frac{d}{d\theta} G_{\theta} = -\sin(\theta)G_0 + \cos(\theta)G_{\pi/2}$ . By the fundamental theorem of calculus and the chain rule,

$$F(G_{\pi/2}) - F(G_0) = \int_0^{\pi/2} \frac{d}{d\theta} F(G_\theta) \, d\theta = \int_0^{\pi/2} \langle \nabla F(G_\theta), \tilde{G}_\theta \rangle \, d\theta. \tag{1}$$

Here, we will finish the proof by applying Jensen's inequality on (1) to bound  $\mathbb{E}[\exp(t(F(G_{\pi/2}) - F(G_0)))]$ . Observe that for  $t \ge 0$ ,

$$\mathbb{E}[\exp(t(F(G_{\pi/2}) - F(G_{1})))] \leq \frac{2}{\pi} \int_{0}^{\pi/2} \mathbb{E}\left[\exp\left(\frac{\pi t}{2} \langle \nabla F(G_{\theta}), \tilde{G}_{\theta} \rangle\right)\right] d\theta$$
$$\leq e^{t^{2}L^{2}(\pi^{2}/8)}$$
$$\leq e^{2t^{2}L^{2}},$$

where the first inequality follows from applying Jensen on  $u \mapsto e^{tu}$  and the second inequality follows from the fact that for any  $\theta$ ,  $G_{\theta}$ ,  $\tilde{G}_{\theta} \sim \mathcal{N}(0, \mathbb{I}_d)$  are i.i.d., which implies that for all  $v, w \in \mathbb{R}^d$ ,

$$\mathbb{E}[\langle G_{\theta}, v \rangle \langle \tilde{G}_{\theta}, w \rangle] = 0.$$

By Jensen again, observe that

$$\mathbb{E}[e^{-tF(G_0)}] \ge e^{-t\mathbb{E}[f(G)]}$$

which implies that

$$\mathbb{E}[\exp(t(F(G_{\pi/2}) - F(G)))] \le e^{2t^2 L^2}.$$

2 2

By Markov's inequality, we observe that

$$\mathbb{P}[F(G) - \mathbb{E}F(G) \ge \lambda] \le \min_{t \ge 0} (e^{2t^2 L^2 - t\lambda}) = e^{-\lambda^2/8L^2}, \text{ with } t = \lambda/4L^2.$$

By symmetry,  $\mathbb{P}[F(G) - \mathbb{E}F(G) \le -\lambda] \le e^{-\lambda^2/8L^2}$ . Thus, adding this up we get that

$$\mathbb{P}[|F(G) - \mathbb{E}F(G)| \ge \lambda] \le 2e^{-\lambda^2/8L^2}$$

which is what we needed to show

*Remark.* This inequality holds more generally for (1)  $\text{Unif}(S_{\sqrt{d}})$  r.v.s, or (2) r.v.s with log-concave density. For the gaussian density  $\exp(-\|x\|^2/2)$ , the Hessian of the log-density  $= -\mathbb{I}_d$ . If this Hessian has maximum eigenvalues  $\leq -c \leq 0$  for some constant *c*, uniformly in  $\mathbb{R}^d$ , we will get a similar result to the one above.

# **3** Application (Borell-TIS inequality)

The Borell–TIS inequality is a result that bounds the probability of a deviation of the uniform norm of a centered gaussian stochastic process above its expected value. It is named after Christer Borell and its independent discovers, Boris Tsirelson, Ildar Ibragimov, and Vladimir Sudakov.

**Theorem 3.1** (Borell-TIS inequality). Suppose  $(g_1, \dots, g_d)$  is a (possibly non-centered) Gaussian vector and

$$\max_{1 \le k \le d} \operatorname{Var}[g_k] \le 1.$$

Then,

$$\mathbb{P}\left[\left|\max_{k} g_{k} - \mathbb{E}\max_{k} g_{k}\right| \geq \lambda\right] \leq 2e^{-\lambda^{2}/8}$$

*Proof.* Let  $\tilde{g}_k = g_k - \mathbb{E}g_k$ . By general principles (or if we follow Harvard's introductory probability class, by definition!), there is a linear function  $\phi : \mathbb{R}^d \to \mathbb{R}^d$  so that if  $\hat{G} \sim \mathcal{N}(0, \mathbb{I}_d)$ , then

 $\phi(\hat{G}) \stackrel{d}{=} (\tilde{g_1}, \cdots, \tilde{g_s}).$ 

Each  $\tilde{g}_i = \langle \hat{G}, v_i \rangle$  and  $||v_i|| = \sqrt{\operatorname{Var}[\tilde{g}_i]} = \sqrt{\operatorname{Var}[g_i]} \le 1$ . Hence,

$$\hat{G} \mapsto \max_{k} g_{k}$$

is 1-Lipshitz. Now we may simply apply concentration of measure and conclude.

In particular, with  $d = \infty$ , this implies that

$$\mathbb{P}\left[\left|\max_{x\in S_N}H_{N,p}(x)-\mathbb{E}\max_{x\in S_N}H_{N,p}(x)\right|\geq \lambda N\right]\leq 2e^{-\lambda^2 N/8}$$

## **4** Geometric Information

**Theorem 4.1.** Let  $x, \tilde{x} \sim \mu_{\beta}$  be *i.i.d.*; then:

(a) If  $\beta \leq \beta_0$ , then

$$\lim_{N \to \infty} R(x, \tilde{x}) \to 0 \text{ in probability.}$$

(b) If  $\beta \ge \beta_1$ , then the above limit is false.

Remark. We will see later on that this transition corresponds to RSB.

*Proof of Theorem* 4.1(*a*). First, we prove (a). Conditioning on  $H_{N,p}$ , we see that

$$\mathbb{P}\left[|R(x,\tilde{x})| \ge \epsilon \mid H_{N,p}\right] = \frac{\int_{S_N} \int_{S_N} \exp\left(\beta H_{N,p}(x) + \beta H_{N,p}(\tilde{x})\right) \mathbf{1}_{|R(x,\tilde{x})| \ge \epsilon} \, dx d\tilde{x}}{Z_N(\beta)^2} \tag{2}$$

Note that for (4), the numerator is small while the denominator is large.

· For the numerator, we observe that

$$\frac{1}{N}\log\mathbb{E}[\text{numerator}] = \max_{-1 \le R \le 1, |R| \ge \epsilon} \left\{ \beta^2 (1+R^p) + \frac{1}{2}\log(1-R^2) \right\} \le \beta^2 - \eta, \text{ where } \eta = \eta(\beta, \epsilon) > 0.$$

• For the denominator, observe that  $Z_N(\beta)^2 = \exp(\beta^2 N + o(N))$ , and if we fix  $\epsilon$  and  $\beta$ , we can, for instance, get that  $Z_N(\beta)^2 = \exp(\beta^2 N + \eta/10)$ , where  $\eta$  is defined above.

By Markov's inequality on the numerator, with high probability, we conclude that

$$\mathbb{P}\left[|R(x,\tilde{x})| \ge \epsilon \mid H_{N,p}\right] \le e^{-\eta N/2}.$$

This proof of (a) works for all *p*, but it lacks motivation and requires the second moment method to work. As such, we will now work our way towards an alternative proof.

**Theorem 4.2.** For any N,  $\beta$ , let x,  $\tilde{x} \sim \mu_{\beta}$  be *i.i.d.*; then,

$$\frac{d}{d\beta} \mathbb{E} F_N(\beta) = \beta \left( 1 - \mathbb{E} R(x, \tilde{x})^p \right)$$

Note that this requires  $G_N^{(p)}$  to be gaussian and the expectation to be taken over all the randomness.

**Proposition 4.3.** Note that with Theorem 4.2, and also assuming we can commute  $N \to \infty$  and  $\frac{d}{d\beta}$ , we can informally prove Theorem 4.1 as follows:

(a) To prove (a), note that when  $N \to \infty$ ,

$$\beta(1 - \mathbb{E}R(x, \tilde{x})^p) = \frac{d}{d\beta} \mathbb{E}F_N(\beta) \approx \frac{d}{d\beta} (\beta^2/2) = \beta.$$

Thus  $\mathbb{E}R(x, \tilde{x})^p \to 0$ . This will work if p is even.

(b) To prove (b), observe that  $\mathbb{E}F_N(\beta) \le c\beta$ , so

$$\frac{d}{d\beta}\mathbb{E}F_N(\beta) \le c,$$

 $so \mathbb{E}R(x, \tilde{x})^p = 1 - \mathcal{O}(1/\beta).$ 

*Note that if*  $N \gg \beta \gg 1$ *, then*  $R(x, \tilde{x}) \approx 1$  *if* p *is odd and*  $|R(x, \tilde{x})| \approx 1$  *if* p *is even.* 

*Proof of Theorem* **4.2***.* For fixed  $H_{N,p}$ , observe that

$$\frac{d}{d\beta}F_{N}(\beta) = \frac{1}{N}\frac{d}{d\beta}\log\int_{S_{N}}e^{\beta H_{N,p}(x)}dx 
= \frac{\int_{S_{N}}H_{N,p}(x)e^{\beta H_{N,p}(x)}dx}{NZ_{N}(\beta)} 
= \sum_{i_{1},\cdots,i_{p}=1}^{N}g_{i_{1}}\cdots_{i_{p}}\left(\frac{\int_{S_{N}}x_{i_{1}}\cdots x_{i_{p}}e^{\beta H_{N,p}(x)}dx}{N^{(p+1)/2}Z_{N}(\beta)}\right)$$
(4)

We use gaussian integration by parts now.

**Lemma 4.4.** Let  $G \sim \mathcal{N}(0, \mathbb{I}_d)$  and  $f : \mathbb{R}^d \to \mathbb{R}$  have 2 bounded derivatives. Then, for  $1 \le j \le d$ ,

$$\mathbb{E}[g_j f(G)] = \mathbb{E}[\delta_{g_i} f(G)]$$

Proof. Observe that

$$\mathbb{E}[g_j f(G)] = \int_{\mathbb{R}^d} f(G) g_j e^{-\sum_k g_k^2/2} dG$$
$$= \int_{\mathbb{R}^d} \delta g_j f(G) e^{\sum_k g_k^2/2} dG$$
$$= \mathbb{E}[\delta_{g_i} f(G)],$$

where the last line comes from integration by parts.

Substituting Lemma 4.4 and adding expectations to (4) above, we have that

$$\mathbb{E}\frac{d}{d\beta}F_{N}(\beta) = \mathbb{E}N^{-(p+1)/2} \sum_{i_{1},\cdots,i_{p}=1}^{N} \frac{d}{dg_{i_{1}\cdots i_{p}}} \left(\frac{\int_{S_{N}} x_{i_{1}}\cdots x_{i_{p}} e^{\beta H_{N,p}(x)} dx}{\int_{S_{N}} e^{\beta H_{N,p}(x)} dx}\right)$$
$$= \frac{\beta}{N^{p}} \mathbb{E}\left[\sum_{i_{1},\cdots,i_{p}=1}^{N} \frac{\int_{S_{N}} x_{i_{1}}^{2}\cdots x_{i_{p}}^{2} e^{\beta H_{N,p}(x)} dx}{Z_{N}(\beta)} - \left(\frac{\int_{S_{N}} x_{i_{1}}\cdots x_{i_{p}} e^{\beta H_{N,p}(x)} dx}{\int_{S_{N}} e^{\beta H_{N,p}(x)} dx}\right)^{2}\right]$$
$$= \beta(1 - \mathbb{E}R(x, \tilde{x})^{p}),$$

where the last line is from the fact that in the second inequality, the first team evaluates to

$$\sum_{i_1,\cdots,i_p=1}^N \frac{\int_{S_N} x_{i_1}^2 \cdots x_{i_p}^2 e^{\beta H_{N,p}(x)} dx}{Z_N(\beta)} = \sum_{i_1,\cdots,i_p=1}^N x_{i_1}^2 \cdots x_{i_p}^2 = \|x\|^{2p} = N^p$$

and the second term evaluates to

$$\left(\frac{\int_{S_N} x_{i_1} \cdots x_{i_p} e^{\beta H_{N,p}(x)} \, dx}{\int_{S_N} e^{\beta H_{N,p}(x)} \, dx}\right)^2 = \frac{\int_{S_N} \int_{S_N} x_{i_1} \tilde{x_{i_1}} \cdots x_{i_p} \tilde{x_{i_p}} e^{\beta H_{N,p}(x) + H_{N,p}(\tilde{x})} \, dx d\tilde{x}}{Z_N(\beta)^2} = \mathbb{E}[\langle x, \tilde{x} \rangle^p \mid H_{N,p}].$$

Combining both terms yields  $\mathbb{E} \frac{d}{d\beta} F_N(\beta) = \beta (1 - \mathbb{E} R(x, \tilde{x})^p).$ 

To make Proposition 4.3 more formal, we will need to use the two following results.

**Proposition 4.5.**  $F_N(\beta)$  is convex in  $\beta$  (for any  $H_{N,p}$ ).

Proof. This is a homework problem. A hint is to use Holder's inequality.

**Proposition 4.6.** Suppose  $f_N(\beta) \to f(\beta)$  is a pointwise convergence of convex functions, and that  $f_N(\beta)$  is smooth for all N. If  $f'(\hat{\beta})$  exists (and it is continuous in  $\hat{\beta}$ ), then

$$\lim_{N\to\infty} f'_N(\hat{\beta}) = f'(\hat{\beta})$$

For our earlier proof, we can thus take  $f_N = \mathbb{E}F_N$ .

*Proof.* This is just a sketch of the proof. For fixed  $\epsilon$ , take N large enough so that

$$|f_N(\beta) - f(\beta)| \le \epsilon^2 \text{ for } \beta \in \{\hat{\beta} - \epsilon, \hat{\beta}, \hat{\beta} + \epsilon\}$$

By convexity,

$$\frac{f(\hat{\beta}) - f(\hat{\beta} - \epsilon)}{\epsilon} - \epsilon \leq \frac{f_N(\hat{\beta}) - f_N(\hat{\beta} - \epsilon)}{\epsilon} \leq f'_N(\hat{\beta}) \leq \frac{f_N(\hat{\beta} + \epsilon) - f_N(\hat{\beta})}{\epsilon} \leq \frac{f(\hat{\beta} + \epsilon) - f(\hat{\beta})}{\epsilon} + \epsilon$$

For  $\epsilon \to 0$  (and for correspondingly large  $N \ge N_0(\epsilon)$ ), these upper and lower bounds converge to  $f'(\hat{\beta})$ .  $\Box$ In the small  $\beta$  case ( $\beta \le \beta_0$ ),

$$f(\beta) = \beta^2/2.$$

For large  $\beta$ , by convexity,

$$\frac{d}{d\beta} \mathbb{E}F_N(\beta) \le \left(\frac{\mathbb{E}F_N(2\beta) - \mathbb{E}F_N(\beta)}{\beta}\right) = 3c \le O(1), \text{ where } c \text{ is a constant}$$

#### 5 On Tensor PCA with Weak Signal

For tensor PCA, the hyperparameter is the signal strength  $\lambda$  instead of  $\beta$ , and we will consider

$$\hat{H}_N(x) = H_N(x) + \lambda N R(x,\sigma)^p$$

and the posterior  $\hat{\mu}_{\lambda}$  for  $\hat{H}_N(x)$ . Here, the free energy is still  $\lambda^2/2$  for small  $\lambda$ . For the lower bound, we can restrict to the nearly "orthogonal band"

$$\left\{ x \mid |R(x,\sigma) \leq \frac{1}{N^{10}} \right\}$$

and re-run the second moment method. For the upper bound, integrating with respect to the overlap with the signal gives a first moment

$$\lim_{N\to\infty}\frac{1}{N}\log\mathbb{E}\hat{Z}_N(\lambda) = \max_{-1\leq R\leq 1}\left(\frac{\lambda^2}{2} + \lambda^2 R^p + \frac{1}{2}\log(1-R^2)\right),$$

maximized at R = 0 for small  $\lambda$ . The conclusion is that  $\hat{\mu}_{\lambda}$  concentrates near  $\sigma^{\perp}$ , since by Markov the free energy on  $\{x \mid |R(x, \sigma)| \ge \epsilon\}$  is at most

$$\max_{|R| \ge c} \left( \frac{\lambda^2}{2} + \lambda^2 R^p + \frac{1}{2} \log(1 - R^2) \right)$$

which is strictly smaller than the free energy on the entire sphere.