CPSC 664, Spring 2023 Scribe: Abhinav Bhardwaj
February 28 Lecturer: Tim Kunisky

Lecture 13: Free Probability and Kac-Rice

1 Introduction

This lecture covers applying free probability to landscape complexity. First, we revisit the Kac-Rice analysis of the "noisy well" and solve for asymptotics of critical points using free convolution. We will discuss the role of large deviations reasoning and discuss a model used to describe the behavior of simple neural networks.

2 Noisy Well Model

Details can be found in [ABM21].

The main idea of this section is to obtain a closed form approximation of the total number of critical points of f in al of $\mathbb{R}^n$.

Recall the noisy well model.

$$f(x) = \frac{\alpha}{2}\|x\|_2^2 + g(x) \tag{1}$$

Where $g(x)$ is the random function from Lecture #8, Continuous Counting with the Kac-Rice Formula. We do not reproduce g here, but we established that for a set A,

$$\mathbb{E}\mathsf{Crit}(f, A) = \frac{1}{\alpha^n}\mathbb{E}[|\mathsf{det}(H)|]\mathbb{P}[\mathcal{N} \in A], \tag{2}$$

where $\mathcal{N}$ was some normal random variable whose parameters do not matter. For details, see Lecture #8. Taking $A := \mathbb{R}^n$, the probability on the right hand side is 1 and we are left analyzing the expected (absolute) determinant.

H was the matrix

$$H := \alpha I_n + \sqrt{d(d-1)N^{d-2}}hI_N + \sqrt{d(d-1)N^{d-1}}W + \sqrt{dN^{d-1}}D. \tag{3}$$

where $h \sim \mathcal{N}(0,1)$ is independent of D and W. D is a random diagonal matrix with independent standard normal random variables on the diagonal. W is a matrix drawn from the *Gaussian orthogonal ensemble* (GOE), normalized to have spectral norm at most 2. For the precise definition of the GOE, see Lecture 8. The most important observation is that the eigenvalue distribution of W follows the semicircle law and that D and W are asymptotically free by Voiculescu's theorem, since the GOE is rotationally invariant.

Define $\widehat{\alpha} := \alpha/N^{\frac{(d-1)}{2}}$, and $\widehat{H} := H/N^{(d-1)/2}$. Set $\widehat{H}^{(0)} := \sqrt{d(d-1)}W + \sqrt{d}D$, the last two terms in $\widehat{H}$. We have, for some constant C,

$$\mathbb{E}[|\mathsf{det}(H)|] = \frac{C}{\widehat{\alpha}^n}\int e^{-N\frac{h^2}{2}}\mathbb{E}[|\mathsf{det}(hI + \widehat{\alpha}I + \widehat{H}^{(0)})|]dh \tag{4}$$

We can rewrite the integral as

$$C \int \exp\left(N \left[-\log \widehat{\alpha} - \frac{h^2}{2} + \frac{1}{N} \log \mathbb{E}[|\det(hI + \widehat{\alpha} I + \widehat{H}^{(0)})|] \right] \right) dh \tag{5}$$

This form motivates us to use the Laplace approximation, which gives

$$\begin{aligned} \mathbb{E}[|\det(H)|] &\approx C \exp\left(N \sup_h \left\{ -\log \widehat{\alpha} - \frac{h^2}{2} + \frac{1}{N} \log \mathbb{E}[|\det(hI + \widehat{\alpha} I + \widehat{H}^{(0)})|] \right\} \right) \\ &= C \exp\left(N \sup_h \left\{ -\log \widehat{\alpha} - \frac{h^2}{2} + \frac{1}{N} \log \mathbb{E}\left[\exp\left(\sum_{i=1}^{N} \log |h + \widehat{\alpha} + \lambda_i| \right) \right] \right\} \right) \end{aligned} \tag{6}$$

where λ_i are the eigenvalues of $\widehat{H}^{(0)}$.

It turns out that by concentration/large deviation bounds, we can very closely approximate this last term by commuting log and the expectation:

$$\mathbb{E}[|\det(H)|] \approx C \exp\left(N \sup_h \left\{ -\log \widehat{\alpha} - \frac{h^2}{2} + \frac{1}{N} \mathbb{E}\left[\sum_{i=1}^{N} \log |h + \widehat{\alpha} + \lambda_i| \right] \right\} \right)$$

and we recognize this last term as an integral with respect to the spectral measure of $\widehat{H}^{(0)}$,

$$= C \exp\left(N \sup_h \left\{ -\log \widehat{\alpha} - \frac{h^2}{2} + \int \log |h + \widehat{\alpha} + t| d\widehat{\mu}_N(t) \right\} \right). \tag{7}$$

Because D and W are asymptotically free, the spectral measure $\hat{\mu_N}$ is approximately the additive free convolution of the spectral measures of D and W,

$$\hat{\mu_N} \approx \mu_{sc} \boxplus \mu_D, \tag{8}$$

where μ_{sc} is the semicircle distribution (the spectral measure of W, see Lecture 10) and μ_D is the spectral measure given by the entries along the diagonal of D (which means it is a Gaussian measure). Therefore, we have a closed form approximation for $\mathbb{E}[|\det(H)|]$, which allows us to approximate the number of critical points of f:

$$\mathbb{E}[|\det(H)|] \approx C \exp\left(N \sup_h \left\{ -\log \widehat{\alpha} - \frac{h^2}{2} + \int \log |h + \widehat{\alpha} + t| d(\mu_{sc} \boxplus \mu_D)(t) \right\} \right). \tag{9}$$

We note that, while not quite a closed form in terms of $\widehat{\alpha}$, this is at least a function that we can approximate numerically, by computing an approximation to the integral term as a function of h and $\widehat{\alpha}$ and then solving an optimization problem over h.

3 A Basic Neural Network Model

The next model we consider is the following.

$$L(x) = \frac{1}{m} \sum_{i=1}^{m} \phi(\langle \zeta_i, x \rangle). \tag{10}$$

We will only sketch how to apply the Kac-Rice formula to this model.

Think of ϕ as an activation function, and assume it is smooth. $x \in \mathbb{S}^{n-1}$ is a unit vector and ζ_i are random vectors whose components are independent, identically distributed Gaussians.

Under some technical assumptions, one can establish a limiting formula for the number of critical points of L in $B \subset \mathbb{R}$.

The main result is that the expected number of critical points has a (convoluted) closed form in terms of the multiplicative free convolution. Another main idea is that here, we need the theory of large deviations in order to get the result. In particular, concentration of the expected determinant will not hold sufficiently to exchange the log and expectation as we did in the previous argument.

3.1 Step 1: Kac-Rice Formula

The Kac-Rice formula on the sphere for this case gives, for a set $B \subset \mathbb{R}$,

$$\mathbb{E}\mathsf{Crit}(L, B) = \int_{\mathbb{S}^{n-1}} \rho_{\nabla L(x)}(0)\mathbb{E}\left[\mathbb{1}\{L(x) \in B\}|\mathsf{det}\,\nabla^2 L(x)|\middle|\nabla L(x) = 0\right] d\mu_{\mathbb{S}}(x), \tag{11}$$

where $\mu_{\mathbb{S}}$ is the uniform measure on the sphere, $\rho_{\nabla L(x)}$ is a density, and the gradients and Hessians must be taken as their Riemannian versions on the sphere, which amounts to introducing some projections to the direction orthogonal to x.

We may take $B = \mathbb{S}^{n-1}$ for the sake of simplicity. In this case, as usual, the main difficulty is controlling the expectation of the absolute value of the determinant. This will involve the corrections for the spherical setting mentioned above, but essentially will boil down to a computation like

$$\mathbb{E}\left|\mathsf{det}\left(\sum_{i=1}^{m} \phi''(\langle\zeta_i, x\rangle)\frac{(P_{x^\perp}\zeta_i)(P_{x^\perp}\zeta_i)^\top}{n}\right)\right|, \tag{12}$$

where $P_{x^\perp} = I - xx^\top$ is the projection to the direction orthogonal to x. Conveniently, if we set $u_i = \langle\zeta_i, x\rangle$ and v_i to be an $(n-1)$-dimensional embedding of $P_{x^\perp}\zeta_i$, then u_i and all coordinates of v_i are independent standard Gaussians, and the above is

$$\mathbb{E}\left|\mathsf{det}\left(\sum_{i=1}^{m} \phi''(u_i)\frac{v_i v_i^\top}{n}\right)\right|, \tag{13}$$

a much nicer random matrix than it initially appears.

3.2 Step 2: Large Deviation Bound

We can write the matrix above as $\frac{1}{n}VDV^\top$ for V an i.i.d. Gaussian matrix and D a diagonal matrix with entries i.i.d. distributed as $\phi''(\mathcal{N}(0,1))$ (i.e., ϕ'' applied to i.i.d. standard Gaussians). If $\frac{m}{n} \to \lambda > 0$ for some parameter λ, then the e.s.d. of $\frac{1}{n}VV^\top$ converges to ρ_λ the Marcenko-Pastur distribution with parameter λ. Thus, the spectrum of $\frac{1}{n}VDV^\top$ converges to $\rho_\lambda \boxtimes \phi''(\mathcal{N}(0,1))$, the multiplicative free convolution. We might then expect

$$\mathbb{E}\left|\mathsf{det}\left(\sum_{i=1}^{m} \phi''(u_i)\frac{v_i v_i^\top}{n}\right)\right| \overset{?}{\approx} \mathsf{exp}\left(m\int \mathsf{log}\,|t| d(\rho_\lambda \boxtimes \phi''(\mathcal{N}(0,1)))(t)\right). \tag{14}$$

As before, this would require sufficiently strong concentration of a particular function of the spectrum of the left-hand side around its typical value. But, in this case, this does *not* hold. Instead, we need to consider large deviations probabilities: by *Sanov's theorem*, the probability that $u_1, \dots, u_m$ have an empirical distribution close to some measure ν is roughly $\mathsf{exp}(-mH(\nu|\mathcal{N}(0,1)))$, where $H(\cdot|\cdot)$ is the relative entropy. Applying a Laplace method approximation, we get instead

$$\mathbb{E}\left|\mathsf{det}\left(\sum_{i=1}^{m}\phi''(u_i)\frac{v_i v_i^\top}{n}\right)\right| \approx \mathsf{exp}\left(m\sup_{\nu}\left\{\int \mathsf{log}\,|t|d(\rho_\lambda \boxtimes \phi''(\nu))(t) - H(\nu|\mathcal{N}(0,1))\right\}\right). \tag{15}$$

Most of the work in [MAB23] is dedicated to showing that a version of this holds for a slightly different random matrix. They then arrive at a similar variational problem over measures that controls the expected number of critical points of L.

References

[ABM21] Gérard Ben Arous, Paul Bourgade, and Benjamin McKenna. Landscape complexity beyond invariance and the elastic manifold, 2021.

[MAB23] Antoine Maillard, Gérard Ben Arous, and Giulio Biroli. Landscape complexity for the empirical risk of generalized linear models, 2023.