# The Kac-Rice Formula for Landscape Complexity and Calculations Applied to Phase Retrieval

Notes, Presentation, & Application by Ian Joffe

Based on *Landscape Complexity for the Empirical Risk of Generalized Linear Models* by Antoine Maillard, Gerard Ben Arous, and Giulio Biroli, 2019

Presented February 2024 for Stat 37797

## 1 Setup

Maillard, Ben Arous, and Biroli's paper considers models of the form

$$y_\mu = \phi(\xi_\mu \cdot x^*) \tag{1}$$

where $x^*$ is a point in the $n$-dimensional unit sphere, $\xi_\mu$ is one of $m$ random measurement vectors sampled i.i.d. from a Gaussian distribution, and $\phi$ is a nonlinearity. I refer to this as a random generalized linear model (GLM), and the formulation is useful across statistics as well as in applications we've seen in class, such as phase retrieval.

One may aim to estimate $x^*$ with $x$. The accuracy of the estimation is quantified by mean square loss

$$L_2(x) = \frac{1}{2m}\sum_{\mu=1}^{m}[\phi(\xi_\mu \cdot x) - y_\mu]^2 \tag{2}$$

To perform the estimation, one may try to find the minima of $L$. One factor that contributes to the difficulty of this optimization is the number of critical points that exist - a higher number of critical points indicates a "rougher" loss landscape. The nature of this property is described by the *landscape complexity*, which Maillard, Ben Arous, and Biroli's paper gives an analytical form of.

## 2 The Kac-Rice Formula

The central tool in calculating landscape complexity is the Kac-Rice formula. My understanding of the formula, summarized here, comes from the notes for Tim Kunisky's course CPSC 664.

Let $N$ be the number of critical points, i.e. the number of zeros of $\nabla L$, within an interval $(a, b)$. The Dirac delta function $\delta(\cdot)$ is defined such that $\int_{-\infty}^{\infty} f(x)\delta(x)dx = f(0)$. In the **one-dimensional case**, where the gradient of $L(x)$ is just a scalar $L'(x)$, consider the line integral over the curve $\gamma$ that is characterized by $L'$ on $(a, b)$. We can see that

$$N = \int_\gamma 1 \cdot \delta(\omega)d\omega \tag{3}$$

because the dirac delta "picks out" a 1 from the 1-function whenever $\gamma$ has a zero. Translating this into an integral over $x$,

$$N = \int_a^b \delta(L'(x))|L''(x)|dx\,. \tag{4}$$

Since the measurement vectors are random, we can consider the loss $L$ and its derivative $L'$ to also be random. With that in mind, we get this into a usable form by taking the expectation of $N$ conditional on $L'(x)$, using its density $p_{L'(x)}$:

$$
\begin{aligned}
\mathbb{E}(N) &= \int_{x=a}^{b} \int_{y=-\infty}^{\infty} \mathbb{E}\left[\delta(L'(x))|L''(x)|\middle| L'(x)=y\right] p_{L'(x)}(y)dydx \\
&= \int_{x=a}^{b} \int_{y=-\infty}^{\infty} \delta(L'(y))\mathbb{E}\left[|L''(x)|\middle| L'(x)=y\right] p_{L'(x)}(y)dydx \\
&= \int_{a}^{b} \mathbb{E}\left[|L''(x)|\middle| L'(x)=0\right] p_{L'(x)}(0)dx\,. \qquad (5)
\end{aligned}
$$

In **multiple dimensions**, $\gamma$ from the line integral becomes a manifold, and the infinitesimals of distance along it become $|\det \nabla(\nabla L)| = |\det \nabla^2 L|$, the determinant of the Hessian. Vector calculus analogous to the derivation in Equation (5) gives the final (non-rigorous) Kac-Rice formula

$$
\mathbb{E}(N) = \int_{A} \mathbb{E}\left[|\det(\nabla^2 L(x))|\middle| \nabla L(x) = \mathbf{0}\right] p_{\nabla L(x)}(\mathbf{0})dx \qquad (6)
$$

where $A$ is some set in $L$'s parameter space and $\mathbf{0}$ is the 0-vector.

# 3 Results for GLMs

M,A,& B consider two expressions of landscape complexity, borrowed from statistical physics. The *annealed complexity* is the log of the expected number of critical points, while the *quenched complexity* is the expectation of the log of the number of critical points. Generally, the quenched complexity is more descriptive of the landscape's nature since it is robust to rare or pathological cases. While the paper does derive impressive expressions for the quenched complexity, they're rather unsightly and difficult to work with, and their derivation is through the replica method, which is out of this report's scope. I'll focus on the annealed case, and encourage those interested to see the quenched results in the paper.

Further, we will begin by computing the annealed complexity of $L(x) = \frac{1}{m}\sum_{\mu=1}^{m}\phi(\xi_u \cdot x)$. This doesn't make much sense as a loss function, but it's a simpler problem and it's results generalize to those for $L_2$, which I'll present at the end. The Kac-Rice formula essentially translates the geometric problem of finding zeros into a random matrix problem. For select toy models, the gradient and Hessian in (6) become independent under fixed $x$, so we can take the expectation unconditionally. Unfortunately, random GLMs are not such a case.

Recall $n$ is the dimensionality of $x$, $m$ is the number of samples (which we index by $\mu$), and $\xi_\mu \sim N_n(0, I_n)$. Let $y_\mu \sim N(0, I_m)$ and $z \sim N(0, I_{n-1})$ be independent random vectors. Then,

$$
L(x) \overset{d}{=} \frac{1}{m}\sum_{\mu=1}^{m} \phi(y_\mu) \qquad (7)
$$

$$
\nabla L(x) \overset{d}{=} \frac{1}{m}\sum_{\mu=1}^{m} \phi'(y_\mu) z_\mu \qquad (8)
$$

$$
\nabla^2 L(x) \overset{d}{=} \left[\frac{1}{m}\sum_{\mu=1}^{m} \phi''(y_\mu) z_\mu z_\mu^T\right] - \left[\left(\frac{1}{m}\sum_{\mu=1}^{m} y_\mu \phi'(y_\mu)\right) I_{n-1}\right] \qquad (9)
$$

The derivations of the above are standard and follow from general formulas given in the paper.

While we don't get independence between the first- and second-order gradients, we do have a stroke of luck in that these distributions are independent of $x$. The first advantage of this is that that the integral over $x$ in Kac-Rice will just be the value at any $x$ times size of the total space we are integrate over. Second, we are now allowed to let $x$ be anything we want. Recall that $x$ is on the unit sphere. Let $\omega_n$ be the volume of the n-dimensional unit sphere. M,A,&B select $x$ to be a the North pole $[1, 0, 0, ..., 0]^T$, denoted $x = e_n$. Therefore applying Kac-Rice, and separating the expectations,

$$
\begin{aligned}
\mathbb{E}(N) &= \omega_n \cdot p_{\nabla L(e_n)}(\mathbf{0}) \cdot \mathbb{E}_{y,z}\left[|\det(\nabla^2 L(e_n))|\middle| \nabla L(e_n) = \mathbf{0}\right] \\
&= \omega_n \cdot \mathbb{E}_y\left[p_{\nabla L(e_n)}(\mathbf{0}) \cdot \mathbb{E}_z\left[|\det(\nabla^2 L(e_n))|\middle| \nabla L(e_n) = \mathbf{0},\ y\right]\right]\,. \qquad (10)
\end{aligned}
$$

The paper has particular interest in the case of the "thermodynamic limit," with high dimensionality and sample size $(n, m \to \infty)$ but a constant underparameterized parameterization ratio $\alpha$ $(m/n \to \alpha > 1)$. Since $y$ is Gaussian, it is known that

$$\omega_n p_{\nabla L(e_n)|y}(\mathbf{0}) = C \exp\left(\frac{n}{2} + \frac{n}{2}\log\frac{m}{2} - \frac{n-1}{2}\log\left(\frac{1}{m}\sum_{\mu=1}^{m}\phi'(y_\mu)^2\right)\right) \tag{11}$$

where $\ln C$ is $\mathcal{O}(n)$. Define $\Lambda := (I_m - \frac{\phi'(y)\phi'(y)^T}{\|\phi'(y)\|^2})D(y)(I_m - \frac{\phi'(y)\phi'(y)^T}{\|\phi`(y)\|^2})$, where $D(y)$ is defined as the diagonal $mxm$ matrix with $\mu$th diagonal element $\frac{n}{m}\phi''(y_\mu)$ . Then, extended calculations with the Gaussian reveal

$$\mathbb{E}_z\left[|\det \nabla^2 L(e_n)|\middle|\nabla L(e_n) = \mathbf{0},\ y\right] = \mathbb{E}_z\left[|\det H_n^\Lambda(y)|\right] \tag{12}$$

where $H_n^\Lambda(y)$ is defined be equivalent in distribution to $\frac{1}{n}z\Gamma(y)z^T - \left(\frac{1}{m}\sum_{\mu=1}^m y_\mu\phi'(y_\mu)\right)I_n$. Putting (10), (11), and (12) together, the expected number of critical points at finite $N$ is

$$\mathbb{E}(N) = Ce^{n\frac{1+\log\alpha}{2}}\mathbb{E}_y\left[e^{-\frac{n-1}{2}\log(\frac{1}{m}\sum_\mu \phi'(y_\mu)^2)}\mathbb{E}_z\left[|\det H_n^\Lambda(y)|\right]\right]\ . \tag{13}$$

Now we move towards the thermodynamic limit of equation (13). The details of the proof are a serveral-page journey into large deviation theory, which is out of scope, but the core is an application of Varadhan's lemma, which is used to express limits of the annealed form $\lim_{n\to\infty}\frac{1}{n}\log\int\exp(\cdot)$. To satisfy the lemma's antecedents, the authors exhaustively demonstrate the boundedness of different parts of equation (13) based on the concentration of $H_n^\Lambda(y)$. The consequence is the annealed complexity of $L$

$$\lim_{n\to\infty}\frac{1}{n}\log\mathbb{E}(N) = \sup_{\nu\in\mathcal{M}_\phi(A)}\left[\frac{1+\log\alpha}{2} - \frac{\mathcal{E}_\phi(\nu)}{2} + \kappa_{\alpha,\phi}(\nu, t_\phi(\nu)) - \alpha H(\nu|\mu_G)\right] \tag{14}$$

- $A$ is the set of parameters which we are searching for $x$ over
- $\nu$ is some probability measure
- $\mathcal{M}_\phi(A)$ is set the set of probability measures on $x$ for which the expectation of $\phi$ is bounded
- $\mathcal{E}_\phi(\nu) := \log\int\nu(dx)\phi'(x)^2$, the log of the second moment of $\phi(x)'$
- $t_\phi(\nu) := \int\nu(dx)x\phi'(x)$, the expectation of $x\phi'(x)$
- $\mu_G$ is the Gaussian measure
- $H(\nu|\mu_G)$ is the conditional entropy
- $\kappa_{\alpha,\phi}(\nu, C) := \int\mu_{\alpha,\phi}[\nu](dx)\log(x - C)$
- $\mu_{\alpha,\phi}[\nu]$ is the asymptotic spectral measure of a matrix constructed using $\phi''$ and two Gaussian vectors, a concept from large deviation theory for random matrices.

The authors then sketch how their derivation can be applied to $L_2$. The main difference is that $L_2$ includes $x^*$. Instead of requiring $x^*$ to be known, which is unrealistic, this is dealt with by requiring the specification of how much signal is present, quantified by $q = x^* \cdot x$. One may specify $q$ exactly, or in the more general case, can specify an interval $Q$ for $q$ that is integrated over in Kac-Rice. We know that $Q \subseteq [-1, 1]$ since $\|x\|_2^2 = 1$.

Analogous steps in Gaussian arithmetic followed by large deviation theory gives the formula for the annealed complexity of $L_2$ as

$$\begin{aligned}\lim_{n\to\infty}\frac{1}{n}\log\mathbb{E}(N) = \frac{1+\log\alpha}{2} + \sup_{q\in Q}\ \sup_{\nu\in M_\phi(B,q)}\Bigg[&\frac{1}{2}\log(1-q^2)\\ &- \frac{1}{2}\mathcal{E}_\phi(q,\nu) + \kappa_{\alpha,\phi}(q,\nu) - \alpha H(\nu|\mu_G)\Bigg]\end{aligned} \tag{15}$$

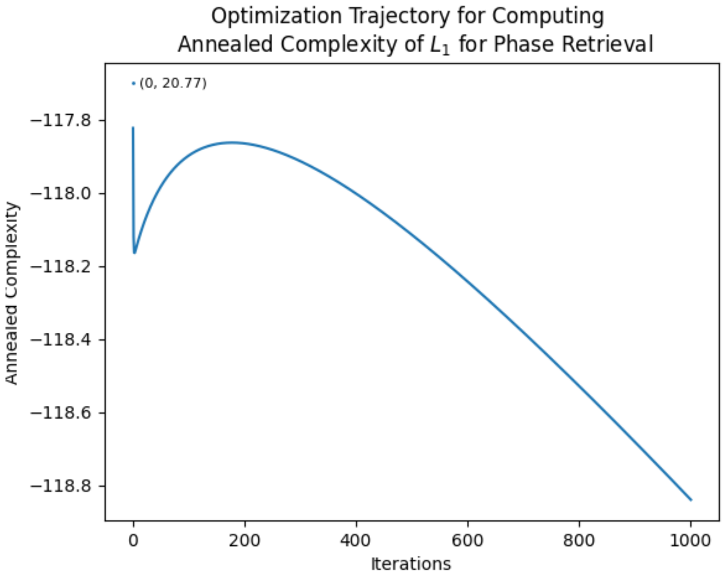
with similar definitions of the terms, but adjusted to include $q$.

# 4 Application to Phase Retrieval

Phase retrieval is a data science problem that we've studied as an example of a random GLM in class. It uses the quadratic as its nonlinearity, i.e. $\phi(x) = x^2$, and is commonly used in experimental scientific domains where noisy measurements can only be made of the square of a variable of interest. Ultimately, I'd like to calculate the annealed or even quenched complexity of $L_2$ for phase retrieval to better understand its optimization, and relate the results to our studies on the problem. But for the sake of this report, I'm going to compute the annealed complexity of $L_1$ for the quadratic nonlinearity, as it's a more straightforward application of the information included in the original paper. Dr. Maillard was also kind enough to respond to an email of mine, and give additional tips about that particular calculation.

The variational expression in equation (14) is nice, but it's difficult to make sense of in terms of actual numbers. So, M,A,&B go through impressive calculations, which also lie a bit out of the scope of this review, to come up with a fixed point iteration algorithm. The solutions found by the algorithm can be plugged into an alternative form of (14) to get a number for the annealed complexity at the thermodynamic limit.

The authors report a null annealed complexity for $L_1$ with the quadratic, and add that the number of critical points is linear in $n$. I've written the attached code based on their fixed point algorithm to verify this computationally, and have indeed come to the same conclusion. My graph below shows the algorithm quickly hitting an annealed complexity such that its exponential is easily zero at machine precision. For iterations 1000 through 10000 (which is as much compute as I wanted to put on this right now), the curve continues to move towards $-\infty$ at a linear pace. This is in line with what we learned in class (although using this machinery to directly prove the algorithms we learned about exist is a ways off). Since the number of critical points of $L_1$ does not blow up exponentially as we collect data, it's plausible that $N$ may not blow up for $L_2$ either, and therefore that algorithms exist to find its global minimum as we've been taught.


Optimization Trajectory for Computing
Annealed Complexity of $L_1$ for Phase Retrieval
(0, 20.77)
Annealed Complexity
Iterations