

Landscape complexity for the empirical risk of generalized linear models: Discrimination between structured data

Theodoros G. Tsironis and Aris L. Moustakas 

*Department of Physics, National Kapodistrian University of Athens, Athens, Greece
and Athena Research Center / Archimedes Research Unit, Athens, Greece*



(Received 16 March 2025; accepted 31 October 2025; published 8 December 2025)

We use the Kac-Rice formula and results from random matrix theory to obtain the average number of critical points of a family of high-dimensional empirical loss functions, where the data are correlated d -dimensional Gaussian vectors, whose number has a fixed ratio with their dimension. The correlations are introduced to model the existence of structure in the data, as is common in current machine learning systems. Under a technical hypothesis, our results are exact in the large- d limit, and characterize the annealed landscape complexity, namely the logarithm of the expected number of critical points at a given value of the loss. We first address in detail the landscape of the loss function of a single perceptron and then generalize it to the case where two competing datasets with different covariance matrices are present, with the perceptron seeking to discriminate between them. The latter model can be applied to understand the interplay between adversity and nontrivial data structures. For completeness, we also treat the case of a loss function used in training generalized linear models in the presence of correlated input data.

DOI: [10.1103/3mbj-xkgk](https://doi.org/10.1103/3mbj-xkgk)

I. INTRODUCTION

The celebrated success of machine learning (ML) is based on the effective use of mostly local iterative algorithms that are designed to “learn” features of data that are fed to them. Underlying the learning process is the efficient search of the local minima of cost (or loss) functions of the weights of non-linear representations (such as perceptrons) of vast amounts of input data. Once the weights are determined, they can be used, e.g., for discrimination purposes on datasets stemming from the same distribution. However, understanding the behavior of these algorithms is limited by the unsuitability of traditional statistical approaches in dealing with the fact that the size of the training sample, the data dimension, and the weight dimensions are typically quite large and often of the same order [1,2]. Furthermore, the performance of local algorithms, like gradient descent (or its variants), depends strongly on the geometry of the loss landscape they operate on, which is typically highly nonconvex [3]. Therefore, the success of modern neural networks would suggest that although their loss function might have many critical points (local minima and saddle points), these points are typically rich in generalization. Nevertheless, on a fundamental level, the loss landscape of such systems is still poorly understood, with only a handful of available results in this direction [4–9].

To provide more clarity on this issue, in this paper, we will address the geometry of the loss function of a single perceptron, a simple model of a neuron, which was first introduced

as a trainable classifier [10]. It is also the building block of the first models used in statistical physics for devising a theory of learning [11–13] and has since been studied as a prototypical spin-glass model. In its simplest form, the loss function takes the form

$$L(w) = \frac{1}{m} \sum_{\mu=1}^m \sigma(w^\top \xi_\mu), \quad (1)$$

where $w \in \mathbb{S}^{d-1}$ is the weight vector and $\{\xi_\mu\}_{\mu=1}^m$ are a set of random data in $\mathbb{R}^d$. In its traditional formulation, the activation function $\sigma(\cdot)$ is a step function with a specific threshold, but choosing a smooth activation function brings the model closer to modern machine learning applications. Furthermore, as a statistical physics model, it has many interesting properties such as jamming and glassy behavior [14].

Our aim is to provide qualitative information on these random loss landscapes by estimating the number of their critical points for each of their level sets. Specifically, we are interested in the quantity

$$\mathcal{N}(B; L) = \#\{L(w) \in B \mid \nabla L(w) = 0\}, \quad (2)$$

in the nontrivial scaling limit where $d \rightarrow \infty$ and $m/d \rightarrow \beta > 1$. This last constraint is required for $L(w)$ to (almost surely) have isolated saddle points that can be counted as such. In this paper, we will focus on the so-called annealed critical point complexity, whose asymptotic form in the large- d limit may be defined as

$$\Psi(B, L) = \lim_{d \rightarrow \infty} \frac{1}{d} \log \mathbb{E} \mathcal{N}(B; L). \quad (3)$$

The expectation of $\mathcal{N}(B, L)$ over the random data vectors $\{\xi_\mu\}_{\mu=1}^m$ above can be obtained by applying the Kac-Rice

Published by the American Physical Society under the terms of the Creative Commons Attribution 4.0 International license. Further distribution of this work must maintain attribution to the author(s) and the published article's title, journal citation, and DOI.

formula [15]

$$\mathbb{E}\mathcal{N}(B, L) = \int_{\mathbb{S}^{d-1}} dw \mathbb{E}[p_{\nabla L(w)}(0) \mathbb{1}_{\{L(w) \in B\}} \times |\det \nabla^2 L(w)| | \nabla L(w) = 0], \quad (4)$$

where $p_{\nabla L(w)}$ is the probability density (with respect to the distribution of the $\{\xi_\mu\}_\mu$) of $\nabla L(w)$. In the above formula and for what follows, $\nabla L(w)$ should be understood as the derivative on the sphere, which is a vector in the tangent space of $\mathbb{S}^{d-1}$ at w and $\nabla^2 L(w)$ should analogously be considered as a matrix operating on that space. Although we limit our investigation to the expected number of critical points, we expect this number to be dominated by local minima at lower loss levels, and by saddle points with a macroscopic number of descent directions at higher loss levels. Indeed, recent related work has shown that, when conditioned on low enough values of $L(w)$, it becomes exponentially rare for the bulk of the Hessian $\nabla^2 L(w)$ spectrum to be nonpositive definite [4,5,16].

Recent advances in random matrix theory have been applied to understand the expectation in (4), resulting in the first picture of the complex loss landscapes emerging from the interplay between randomness and high dimensionality [4,6,16–18]. But in practice, sources of complexity are not exhausted there. For example, further richness of the resulting loss landscapes stems from the structure of the data distribution, which is commonly far from white noise, contrary to the common theoretical assumption. This structure is exactly what algorithms should seek to identify and leverage for discrimination or generation tasks. Each data vector can be generated from one of several distributions, corresponding to the different structure classes to be identified [1]. Alternatively, it can be generated from a distribution whose structure is to be learned and replicated [19,20]. This structure is expressed, among other properties, in the correlations of the data distribution. Specifically, even Gaussian distributions, when high dimensional, can express surprisingly rich structure and capture highly complicated datasets [21]. Furthermore, because of universality and concentration of measure, high-dimensional Gaussians are often not far in terms of properties from Lipschitz functions of high-dimensional Gaussians [21,22,23], which in turn empirically seem to be able to approximate common datasets with high precision, as is evident from the success of modern generative models for images such as generative adversarial networks (GANs) and diffusion models [24–26]. A trivial example of a property, which only depends on the covariance for data that are Lipschitz functions of Gaussian random variables, is the empirical loss value, $L(w)$ for a given w .

Another relevant source of richness of loss landscapes, which is encountered in practice, is the existence of competition, i.e., the case where the models learn not only to accept data from a specific distribution, but also to reject data from other, related distributions. In discriminating tasks, this process is evident, as common datasets contain many different classes, some of which are similar and could confuse the model [1,2]. In this context, an important example is the case of GANs [20], where the discriminating model is trained to antagonize a generator, which learns to approximate the target distribution with increasing accuracy. The success of such

models stems from the increasing competition between the two adversaries.

II. RELATED WORKS

The study of the convergence properties of descent methods has a long history since their introduction [27]. A significant achievement has been the proof of convergence of stochastic gradient descent (SGD) in the convex setting [28]. Later investigations have focused on the characterization of asymptotic trajectories and the limit points of SGD and narrowed down the locations of these limit points to a subset of the critical points [29]. Generally, there has been given particular emphasis in the exclusion of saddle points as limiting points of local methods [30–32], since the loss landscape is designed with its minima in mind. These results are greatly illuminating of the mechanisms of SGD and its variants, but practical needs increasingly move away from the scenarios where they apply. This is partly due to their inherent lack of any sort of nonasymptotic quantification of the rate of convergence, which becomes increasingly important as the relevant loss landscape becomes increasingly complex. Unfortunately, only a few results go beyond convexity and low dimensionality in a meaningful way for modern machine learning. These aspects have been studied in limited but relevant models, in the physics literature [33] and have recently also been rigorously proven [34]. Furthermore, the dynamical mean-field theory has been employed for capturing the performance of SGD [35,36]. Of course, such analyses can only be carried out when the loss landscape is fairly simple, yet they provide significant intuition to the problem.

Moving to the analysis of local algorithms on more relevant problems also requires understanding the geometry of increasingly complex landscapes. A list of initial questions to be addressed in the investigation of any landscape includes the calculation of the number of saddle points and the local minima. In the physics literature, there has already been an interest in answering these kinds of questions for random functions [37–40]. As a result, a toolkit drawn from statistical physics, random matrix theory, and large deviations theory has been developed for computing the expected number of saddle points with specific properties [6,16]. A crucial part of this toolkit relies on unproven assumptions regarding the Hessian of the loss functions and can thus only provide conjectures, but significant progress has been made in developing rigorous methods and proving these results [41]. In particular, this approach has been significantly enriched in the process of its application to understand complex landscapes that arise in the context of spin glasses [5,16,17,42–49].

The landscape analysis of spin glasses has provided insight that can be applied to machine learning problems [17,50–52]. A number of more recent works have made more quantitative connections between spin glasses and machine learning, investigating the consequences that landscape complexity theory would imply for machine learning [4,5]. These works are significant in that there is no other known approach for understanding the loss landscape of modern models in such a global manner, but they are also limited in their domain of applicability. The limitations fall in one of two categories: either the models are taken directly from the field of spin

glasses and are thus unconvincing in their correspondence with modern machine learning models [53,54], or they only model a handful of neurons [6] and so distance themselves from the emergent properties of more contemporary models. Nevertheless, both cases contribute significantly to the understanding of the properties of loss landscapes by modeling multiple of their features, such as the randomness of the data sample, the scaling between data and parameters, the nonlinearity, and complex interactions between the weights.

The proposed models also suffer from limitations that can be somewhat mitigated by appropriate generalizations. One such limitation stems from ignoring the existence of a data structure, under the common assumption that the data are (potentially labeled) white noise [6]. Overcoming this limitation could be achieved by the introduction of correlation for each datapoint. Another direction for generalization is the introduction of adversarial data, as was done, for example, in [5], in an attempt to model GAN loss-gain landscapes through spin glasses. In the context of perceptron-based models, the adversarial data can be introduced as a second set of data, which the perceptron is motivated to suppress.

From an analytical point of view, the connection of these methods with random matrix theory emerges through a random matrix related to the Hessian of the loss function at each point of the parameter manifold. Understanding the spectra of these random matrices is important to evaluate the number of critical points, whereas the analysis of the deviations of their largest eigenvalues from the bulk leads to estimations of the indices of critical points.

III. MAIN RESULTS

In this work, we consider the expected number of critical points of the loss function $L(w)$, $\mathbb{E}\mathcal{N}(B, L)$, as well as the expected number of critical points of a closely related family of loss functions, which appear in the training of generalized linear models (GLMs). These models, in their planted version, aim to infer a vector w_* from nonlinear observations of the form $\sigma(w_*^\top \xi_\mu)$ over a dataset $\{\xi_\mu\}_{\mu=1}^m$. Such a loss function typically has the form

$$L_1(w) = \frac{1}{2m} \sum_{\mu=1}^m (\sigma(w^\top \xi_\mu) - \sigma(w_*^\top \xi_\mu))^2. \quad (5)$$

This task is closer to those usually considered in machine learning when one is trying to train a classifier on data $\{\xi_\mu\}_{\mu=1}^m$ to fit the labels $\{\sigma(w_*^\top \xi_\mu)\}_{\mu=1}^m$.

The landscapes of these models have been studied in the case where the data are sampled from white noise, in the limit where the number of data scales with the data dimension so that $m/d \rightarrow \beta$ [6]. In our attempt to understand how the structure of the data affects the loss landscape, we will remain in this asymptotic regime, but study the models in the case where the data distribution has a nontrivial correlation structure, expressed through an arbitrary correlation matrix $\Sigma = R^\top R$, i.e., $\xi_\mu = Rz_\mu$ with z_μ being d -dimensional white noise, i.e., $z_\mu \sim N(0, I_d)$. In our study, we will not fully investigate the necessary conditions on Σ but instead limit our discussion to matrices with bounded spectral norm, well-defined limiting spectral distribution, and at most finitely many outlier eigenvalues, converging to specific values.

Finally, we formulate a model based on the perceptron that aims to optimize a loss function that also takes into consideration a dataset to be avoided. Suppose, for example, that $\sigma: \mathbb{R} \rightarrow [0, 1]$, and that we want to have an output as close to 0 when the input comes from one dataset $\{\xi_\mu\}_{\mu=1}^{m_1}$ and as close to 1 as possible when the input originates from another dataset, namely $\{\xi_\mu\}_{\mu=m_1+1}^{m_1+m_2}$, thus discriminating between the landscapes. Then, we might use the loss function

$$\begin{aligned} L_2(w) &= \frac{1}{m_1} \sum_{\mu=1}^{m_1} \sigma(w^\top \xi_\mu^1)^2 + \frac{1}{m_2} \sum_{\mu=m_1+1}^{m_1+m_2} (1 - \sigma(w^\top \xi_\mu^2))^2 \\ &= \frac{1}{m_1} \sum_{\mu=1}^{m_1} \sigma_1(w^\top R_1 z_\mu) - \frac{1}{m_2} \sum_{\mu=m_1+1}^{m_1+m_2} \sigma_2(w^\top R_2 z_\mu), \end{aligned} \quad (6)$$

where, for reasons of notational convenience, we have introduced $\sigma_1 = \sigma^2$ and $\sigma_2 = (1 - \sigma)^2$, but we might as well consider general σ_1 and σ_2 . As a result, we have two competing datasets, differentiated through the structure of their covariance matrices $G = R_1 R_1^\top$ and $\Sigma = R_2 R_2^\top$. It is important to keep in mind that the perceptron cannot meaningfully generalize in the discrimination task, since it only checks a one-dimensional projection of high-dimensional data, yet we study the resulting landscape as an illuminating scenario for the landscapes that result in the interplay between structure and competition.

A. Complexity Formula for $L(w)$

Since $\mathcal{N}(B, L)$ is typically an exponentially large quantity, the calculation of its mean will require the usage of large deviation principles (LDPs). In our results, we will express $\mathbb{E}\mathcal{N}(B, L)$ as an extremum of the free energy over an $\mathcal{O}(1)$ number of macroscopic order parameters. Starting with the loss function L , we introduce a number of order parameters and their associated “free energy”-type functions in order to state our results.

Consider the “geometric” order parameters ρ and q representing the values of $w^\top \Sigma w$ and $w^\top \Sigma^{-1} w$, respectively. Accordingly, their common domain is the image of $\mathbb{S}^{d-1}$ under the mapping $w \rightarrow (w^\top \Sigma w, w^\top \Sigma^{-1} w)$. It is worth pointing out that, whereas every component of the resulting formula (14) will depend on Σ only through its limiting spectral distribution, the domain of ρ and q depends on the whole spectrum, including any outlier eigenvalues.

It is convenient to define the vector y here, with elements $y_\mu = w^\top \xi_\mu$, which are the arguments of the activation function $\sigma(\cdot)$ in (1). Using this notation, we will need to analyze the “statistical” order parameters f and κ , representing the averages over the dataset $\{\xi_\mu\}$ of $\sigma'(y_\mu)^2$ and $y_\mu \sigma'(y_\mu)$, respectively. In addition, of course, ℓ represents the value of the loss, which can be expressed as the average of the quantities $\sigma(y_\mu)$. These parameters have also been defined and appear in [6]. Accordingly, it is clear that the domain of f is $[\min_y \sigma'(y)^2, \max_y \sigma'(y)^2]$, while the domain of κ is $[\min_y y \sigma'(y), \max_y y \sigma'(y)]$ and the domain of ℓ is B .

Furthermore, consider the “random matrix” order parameters ϕ and $\bar{\phi}$, representing the Stieljes transforms related to a doubly-correlated random matrix model which will be discussed in Sec. IV B. Equivalently, they represent the

minimum mean square error (MMSE) of inferring a correlated Gaussian signal through the matrix defined by the data under additive Gaussian corruption, $\{\xi_\mu\}_{\mu=1}^m$ [55,56]. Since these are Stieltjes transforms, their extremization is unconstrained over the upper half complex plane.

Finally, consider $\nu(d\zeta)$, a probability measure representing the limiting distribution of y_μ over the dataset, conditionally on the values of κ , f , and ℓ . This measure can be parametrized by three auxiliary, real, unconstrained parameters, $\bar{\kappa}$, $\bar{f}$, and $\bar{\ell}$. The explicit dependence on these parameters can be seen in the expression of ν :

$$\nu(d\zeta) = \frac{1}{\mathcal{Z}(\bar{\ell}, \bar{\kappa}, \bar{f}; \rho)} e^{-\frac{1}{2}\zeta^2 - \frac{1}{\beta}F_v(\zeta; \bar{\ell}, \bar{\kappa}, \bar{f}, \rho)} d\zeta, \quad (7)$$

where

$$F_v(\zeta; \bar{\ell}, \bar{\kappa}, \bar{f}, \rho) = \bar{\ell}\sigma(\sqrt{\rho}\zeta) + \bar{f}\sigma'(\sqrt{\rho}\zeta)^2 + \bar{\kappa}\sqrt{\rho}\zeta\sigma'(\sqrt{\rho}\zeta), \quad (8)$$

and

$$\mathcal{Z}(\bar{\ell}, \bar{\kappa}, \bar{f}; \rho) = \int d\zeta e^{-\frac{1}{2}\zeta^2 - \frac{1}{\beta}F_v(\zeta; \bar{\ell}, \bar{\kappa}, \bar{f}, \rho)}. \quad (9)$$

The normalization $\mathcal{Z}$ is closely related to F^S , the “statistical” free energy that controls the deviation of the empirical distribution of the input of the perceptron function from the normal distribution. In particular, we have

$$F^S(\ell, \kappa, f, \nu; \rho) = \bar{\ell}\ell + \bar{\kappa}\kappa + \bar{f}f + \beta \log \mathcal{Z}(\bar{\ell}, \bar{\kappa}, \bar{f}; \rho). \quad (10)$$

Taking the derivative with respect to the auxiliary barred parameters in the above equation leads to a fluctuation-dissipation type relation [57]. This reveals that the role of these auxiliary parameters is to fix the values of the means of the random variables they multiply in (8) with respect to the distribution $\nu(\cdot)$.

The “geometric” free energy, F^{SP} , contains the probability that a generic point on the sphere is a saddle point and controls the deviations of the geometric parameters of the system, ρ and q ,

$$F^{SP}(\rho, q; \kappa, f) = \frac{1}{2} \log(\beta e) - \frac{1}{2} \log(f) - \frac{1}{2} \mathbb{E}_{\lambda_\Sigma} \log(\lambda_\Sigma) + \frac{1}{2} \beta \frac{\kappa^2}{f\rho} (1 - \rho q) - I(\rho, q), \quad (11)$$

where λ_Σ is a random variable distributed according to the limiting spectral distribution of Σ and the rate function $I(\rho, q)$ is given by

$$I(\rho, q) = \sup_{\bar{\rho}, \bar{q}} \frac{1}{2} \mathbb{E}_{\lambda_\Sigma} \log [(1 - \bar{\rho}\rho - \bar{q}q)I + \bar{\rho}\lambda_\Sigma + \bar{q}\lambda_\Sigma^{-1}]. \quad (12)$$

Lastly, F^{RM} controls the large deviations of the determinant of the Hessian of the loss function at each saddle point

$$\begin{aligned} F^{RM}(\phi, \bar{\phi}; \rho, \kappa + i\epsilon; \nu) \\ = \beta \log \mathbb{E}_\zeta \left| 1 + \frac{1}{\beta} \sigma''(\sqrt{\rho}\zeta) \phi \right| \\ + \mathbb{E}_{\lambda_\Sigma} \log |(\kappa + i\epsilon)(1 + \bar{\phi}\lambda_\Sigma)| + \text{Re}((\kappa + i\epsilon)\bar{\phi}\phi), \end{aligned} \quad (13)$$

where we have made the dependence of F^{RM} on the density $\nu(\zeta)$ explicit. From the above definitions we can state our main result:

Result III.1. The critical point complexity of L can be obtained as the solution of the following extremization problem:

$$\begin{aligned} \Psi(B; L) = \sup_{\rho, q, f, \kappa, \ell \in B} (F^{SP}(\rho, q; \kappa, f) \\ + \inf_{\nu} \lim_{\epsilon \rightarrow 0^+} \text{extr}_{\phi, \bar{\phi}} (F^{RM}(\phi, \bar{\phi}; \rho, \kappa + i\epsilon; \nu) \\ + F^S(\ell, \kappa, f, \nu; \rho))), \end{aligned} \quad (14)$$

where F^S is defined in Eq. (10), F^{SP} is defined in (11), and F^{RM} is defined in (13). The domain over which the order parameters should be extremized is as described in the discussion above.

B. Complexity formula for $L_1(w)$

Moving on to the loss function related to GLMs, we have a similar formula for the complexity.

Result III.2. The critical point complexity of L_1 can be computed as the solution of the following extremization problem:

$$\begin{aligned} \Psi(B; L_1) = \sup_{\rho, q, p, r, f, \kappa, \kappa', \ell \in B} (F_1^{SP}(\rho, q, r, p; \kappa, \kappa', f) \\ + \inf_{\nu} \lim_{\epsilon \rightarrow 0^+} \text{extr}_{\phi, \bar{\phi}} (F_1^{RM}(\phi, \bar{\phi}; \rho, p, \kappa + i\epsilon; \nu) \\ + F_1^S(\ell, \kappa, \kappa', f, \nu; \rho, p))), \end{aligned} \quad (15)$$

where F_1^{SP} is defined in (62), F_1^{RM} is defined in (70), and F_1^S is defined in (68). The domains of the order parameters are analogous to the one of (14), with the differences described in Sec. V.

There are some notable changes with respect to Eq. (14), which we should mention here. First, there are new functions of the input of the perceptron that must be controlled, changing the meaning of κ , f , and ℓ , and introducing κ' , which will be defined in Sec. V. Additionally, we need to control a new pair of geometric variables, given by $p = w^\top \Sigma w_*/\sqrt{\rho}$ and $r = w^\top w_*$.

C. Complexity Formula for $L_2(w)$

Finally, the expected number of critical points and hence the complexity for the loss function in (6) can be expressed compactly in the same form as (14), but the functions and the meaning of the underlying parameters change significantly.

Result III.3. The critical point complexity of L_2 can be computed as the solution of the following extremization problem:

$$\begin{aligned} \Psi(B; L_2) = \sup_{\rho, q, f, \kappa, \ell \in B} (F_2^{SP}(\rho, q; \kappa, f) \\ + \inf_{\nu} \lim_{\epsilon \rightarrow 0^+} \text{extr}_{\phi, \bar{\phi}} (F_2^{RM}(\phi, \bar{\phi}; \rho, \kappa, \nu, \epsilon) \\ + F_2^S(\ell, \kappa, f, \nu; \rho))), \end{aligned} \quad (16)$$

where F_2^{SP} is defined in (88), F_2^S is defined in (97), and F_2^{RM} is defined in (93). The domains of the order parameters are described in Sec. VI.

The most important difference with Eq. (14) of Result III.1 is that the underlying variables κ , f , ρ , ϕ , as well as their barred counterparts, are now two-dimensional, e.g., $\kappa = (\kappa_1, \kappa_2)$ etc., containing one value for each dataset. Furthermore, q is now a six-dimensional variable, controlling various geometric properties of the system, mixing Σ and G . More details can be found in Sec. VI.

D. Notation

In the following (Secs. IV–VI), we discuss how one can arrive at Results III.1, III.2, and III.3. Before we proceed with that discussion, a comment on our notation is necessary. Given a function $\sigma : \mathbb{R} \rightarrow \mathbb{R}$, by abuse of notation, we also extend $\sigma : \mathbb{R}^m \rightarrow \mathbb{R}^m$ through $\sigma(y)_\mu = \sigma(y_\mu)$ and similarly for arguments of different dimensions. We denote I_d the $d \times d$ identity matrix, and we might omit the subscript when it is obvious. Furthermore, given a vector, say $w \in \mathbb{R}^d$, let $P_w^\perp : \mathbb{R}^d \rightarrow \mathbb{R}^{d-1}$ be an operator such that $P_w^\perp w = 0$ and $\|P_w^\perp x\| = \|x\|$ for any $x \perp w$. The discussion that follows is independent of the exact choice of $P_w^\perp$. We will also denote by $P_w^\perp$ the associated $(d-1) \times d$ matrix. We denote by ∂ the usual Euclidean gradient in the space $\mathbb{R}^d$ and by $\partial \partial^\top$ the Euclidean Hessian, while we reserve the symbols ∇ for the gradient induced to a manifold by its Riemannian structure and ∇^2 for the induced Hessian. In many cases, given the self-averaging property of order parameters, we will treat them both as random variables and as “thermodynamic” state-defining variables. Accordingly, we will overload symbols like q , ρ , ℓ , f , κ , etc., to denote both random variables and parameters of a free energy (which should then be extremized over). Finally, in certain cases where compactness of expressions is necessary, we will use the symbol 0^+ , by which we mean that 0^+ should be replaced by $\epsilon > 0$ and this ϵ should be taken to 0 after all extr operations in the expression have been performed.

IV. PERCEPTRON WITH CORRELATED DATA: $L(w)$

We will start with details of the calculation for $\mathbb{E}\mathcal{N}(B; L)$, and then we will explain how they generalize to the other models. Our analysis follows in general terms the recipe presented in [6] for the perceptron with uncorrelated data, although the existence of the covariance matrices will result in significant changes in all the steps. For clarity, we will break up the derivation into four steps. The first, in Sec. IV A, will result in an expression for the Hessian of the loss function and the corresponding Kac-Rice formula, while the second step appearing in Sec. IV B will provide details of the calculation of the determinant of the Hessian matrix using random matrix theory. Section IV C discusses a key and third part of the analysis regarding the constrained integration of w over the hypersphere, with the effect of the large deviations of the empirical measure left for Sec. IV D.

A. Distributions of the gradient and the Hessian of the loss function

It will be useful to first define a number of quantities. Let $Z \in \mathbb{R}^{d \times m}$ be the random matrix such that $Z_{i\mu} = (z_\mu)_i$. Also, recalling that $\rho = w^\top R^\top R w$, we may express the normalized

input empirical vector as $y = \frac{1}{\sqrt{\rho}} Z^\top R w$. Conditional on w , y is a Gaussian vector with $y \sim N(0, I_m)$. Also, we introduce the vector $u = \frac{1}{\sqrt{\rho}} P_w^\perp R^\top R w$. We then have the following proposition:

Proposition IV.1. Let $\kappa = \frac{1}{m} \sqrt{\rho} y^\top \sigma'(\sqrt{\rho} y)$ and $f = \frac{1}{m} \|\sigma'(\sqrt{\rho} y)\|^2$.

(1a): There exists a normal random vector $z' \in \mathbb{R}^{d-1}$ independent of y , such that the gradient of the loss function $L(w)$ has the following representation:

$$\nabla L(w) = \frac{\sqrt{f}}{\sqrt{m}} ((P_w^\perp R^\top)(P_w^\perp R^\top)^\top - uu^\top)^{1/2} z' + \frac{\kappa}{\sqrt{\rho}} u. \quad (17)$$

(1b): There exists a $(d-1) \times m$ random matrix Z' , with independent and identically distributed (i.i.d.) normal elements, independent from y , such that the Hessian of the loss function $L(w)$ on the sphere, conditionally on the constraint $\{\nabla L(w) = 0\}$, can be expressed as

$$\nabla^2 L(w) \Big|_{\nabla L(w)=0} = \frac{1}{m} (P_w^\perp R^\top Z' + \Pi) \Lambda (P_w^\perp R^\top Z' + \Pi)^\top - \kappa I_{d-1}, \quad (18)$$

where Λ is the $m \times m$ diagonal matrix with $\sigma''(\sqrt{\rho} y_\mu)$ as entries, i.e., $\Lambda = \text{diag}(\sigma''(\sqrt{\rho} y))$, and Π is a matrix of rank at most 3.

Proof. Note that the gradient ∇ of a function on the sphere $w \in \mathbb{S}^{d-1}$ is orthogonal to w . Hence, we may express $\nabla L(w)$ in terms of $P_w^\perp$ and the ambient derivative on $\mathbb{R}^d$, ∂_w , as $\nabla L(w) = P_w^\perp \partial_w L(w)$, which can be further expanded as

$$\nabla L(w) = \frac{1}{m} P_w^\perp R^\top Z \sigma'(\sqrt{\rho} y). \quad (19)$$

Since $P_w^\perp R^\top Z$ and y are jointly Gaussian, we have that $P_w^\perp R^\top Z - P_w^\perp R^\top \mathbb{E}[Z|y]$ is independent from y . From the rules about conditioning Gaussian random variables, we have that

$$\begin{aligned} \mathbb{E}[z_\mu | y_{\mu'}] &= \delta_{\mu, \mu'} \mathbb{E}[z_\mu y_\mu] \mathbb{E}[(y_\mu)^2]^{-1} y_\mu \\ &= \delta_{\mu, \mu'} \frac{1}{\sqrt{\rho}} R w y_\mu. \end{aligned} \quad (20)$$

Thus,

$$P_w^\perp R^\top \mathbb{E}[Z|y] = u y^\top. \quad (21)$$

With a straightforward calculation, we can see that

$$\text{Cov}[P_w^\perp R^\top z_\mu - u y_\mu] = (P_w^\perp R)(P_w^\perp R)^\top - uu^\top. \quad (22)$$

Combining Eqs. (21) and (22) together with the fact that $P_w^\perp R^\top Z - P_w^\perp R^\top \mathbb{E}[Z|y]$ is independent from y , we see that there exists a $(d-1) \times m$ random matrix Z' , with i.i.d. normal elements, independent from y such that

$$P_w^\perp R^\top Z = ((P_w^\perp R)(P_w^\perp R)^\top - uu^\top)^{1/2} Z' + u y^\top. \quad (23)$$

The part (1a) follows directly from the above representation and Eq. (19).

Now, consider the expectation of Z conditionally on the value of $\nabla L(w)$. We can calculate this expectation as above to

get

$$\mathbb{E}[Z|\nabla L(w)] = \sigma'(\sqrt{\rho}y)(P_w^\perp R^\top)^\top \times ((P_w^\perp R^\top)(P_w^\perp R^\top)^\top)^{-1} \nabla L(w)^\top, \quad (24)$$

which is a rank 1 matrix. Once again, due to the Gaussianity of Z and $\nabla L(w)$, $P_w^\perp R^\top(Z - \mathbb{E}[Z|\nabla L(w)])$ is independent of $\nabla L(w)$. Thus, $P_w^\perp R^\top Z$ differs from $P_w^\perp R^\top Z|_{\nabla L(w)}$ only by a matrix of rank 1. However, by Eq. (23), $P_w^\perp R^\top Z$ differs from $P_w^\perp R^\top Z'$ by a matrix of rank at most 2. As a result, we have

$$P_w^\perp R^\top Z|_{\nabla L(w)} = P_w^\perp R^\top Z' + \Pi, \quad (25)$$

with $\text{rank}(\Pi) \leq 3$.

The Hessian on $\mathbb{S}^{d-1}$, $\nabla^2 L(w)$, is connected with the Hessian on $\mathbb{R}^d$, $\partial_w \partial_w^\top L(w)$, through $\nabla^2 L(w) = P_w^\perp \partial_w \partial_w^\top L(w) P_w^\perp - w^\top \partial_w L(w) P_w^\perp$. Thus, after some algebra, the Hessian of $L(w)$ may be written as

$$\nabla^2 L(w) = \frac{1}{m} (P_w^\perp R^\top Z) \Lambda (P_w^\perp R^\top Z)^\top - \kappa I_{d-1}. \quad (26)$$

Applying (25) here, we arrive at (1b). $\blacksquare$

It is important to stress that the difference between the Hessian matrix conditioned on the vectors y and $\nabla L(w) = 0$ (in (18) and the unconditioned matrix Hessian in (25) are low rank perturbations, which will be insignificant in this analysis, where only bulk random-matrix properties are important. Considerations of outlier eigenvalues carry information on the nature of the dominating critical points with a given loss, for example, if they are local minima or saddle points, which will not be investigated here.

Now, with knowledge of the distribution of the derivative and the Hessian, we can make use of the Kac-Rice formula (4), which we rewrite in the following form [6,15] by disintegrating to condition the determinant of the Hessian on the value of y :

$$\mathbb{E}\mathcal{N}(B, L) = \int_{\mathbb{S}^{d-1}} dw \mathbb{E}_y [p_{\nabla L(w)|y}(0) \times \mathbb{E}[\det \nabla^2 L(w) | \nabla L(w) = 0, y]], \quad (27)$$

where now the probability density $p_{\nabla L(w)|y}$ that appears is that of $\nabla L(w)$ conditioned on the values of y .

B. Derivation of F^{RM} : Expectation of the determinant

To calculate $\mathbb{E}[\det \nabla^2 L(w) | \nabla L(w) = 0, y]$, we will make use of the assumption that typically the logarithm of the determinant of a random matrix of the form $H = \frac{1}{m} (P_w^\perp R^\top Z' + \Pi) \Lambda (P_w^\perp R^\top Z' + \Pi)^\top$ is dominated by its mean value to an exponential scale:

$$\log \mathbb{E}_Z[\det(H - \kappa I)] = \mathbb{E}_Z[\log |\det(H - \kappa I)|] + \mathcal{O}(1), \quad (28)$$

and that the eigenvalues of $H - \kappa$ rarely get close enough to zero to obstruct the evaluation of $\mathbb{E}[\log |\det(H - \kappa I)|]$ by treating it as a spectral statistic through a random matrix limit theorem:

$$\frac{1}{d} \mathbb{E}_Z[\log |\det(H - \kappa I)|] = \mathbb{E}_\lambda \log |\lambda - \kappa| + \mathcal{O}\left(\frac{1}{d}\right), \quad (29)$$

where λ is distributed according to the (random) limiting spectral measure of $H|_y$. Probabilistic control of the error terms in (28) and (29) to superexponential rate is expected, but remains highly nontrivial from a technical perspective. Such control is a highly motivated and common assumption, necessary to proceed with the study of these models [6]. In specific models, these results can be obtained directly [5,16,42]. In general, attempts at rigorous proofs have been fruitful, yet still unable to tackle a model like the one considered here [41]. Nevertheless, this assumption was proven in the case where H is a sample covariance matrix formed from samples of a distribution with i.i.d. elements [41], which is quite close to our case. In the model under consideration, as in [6], the main technical difficulty of rigorously proving the desired results is the presence of the low rank perturbation Π . We expect this difficulty to be only of a technical nature, since the spikes are not expected to drastically affect the spacing of the eigenvalues, which is the important statistic here. Notably, the “mutual information” $\text{tr} \log(H - \kappa I)$ is an extensively studied object with a well-known Gaussian limiting law with finite variance [55,58,59].

Given the above conjecture, we now only need the spectral measure of H for fixed y . By the Cauchy interlacing lemma [60], a fixed rank perturbation like the matrix Π does not affect the logarithmic potential to first order, and so we will ignore it. We will work with the potential defined on $\mathbb{C}^+$ by $G(z) = \mathbb{E} \log(z - \lambda)$, which is connected to the logarithmic potential through $\text{Re}[G(\kappa + i\epsilon)] \xrightarrow{\epsilon \rightarrow 0^+} \mathbb{E} \log |\lambda - \kappa|$. We know from [61], that the potential $G(z)$ can be understood through the random matrix free energy $\tilde{F}^{RM}$, whose dependence on v will be kept implicit as follows:

$$\tilde{F}^{RM}(\phi, \bar{\phi}; \rho, z) = \beta \mathbb{E}_\zeta \log \left(1 + \frac{1}{\beta} \sigma''(\sqrt{\rho}\zeta) \phi \right) + \mathbb{E}_{\lambda_\Sigma} \log(z + z\lambda_\Sigma \bar{\phi}) + z\phi \bar{\phi}, \quad (30)$$

where ζ is distributed according to the random measure $\nu_y = \frac{1}{m} \sum_\mu \delta_{y_\mu}$ and, as mentioned in Sec. III, λ_Σ is distributed according to the limiting spectral measure of Σ . Specifically, we have

$$G(z) = \text{extr}_{\phi, \bar{\phi}} \tilde{F}^{RM}(\phi, \bar{\phi}, \rho, z). \quad (31)$$

Combining Proposition IV.1 and (31) with (28) and (29), we obtain the following result:

Result IV.2. Consider $\tilde{F}^{RM}$ as defined in Eq. (30), then we have

$$\begin{aligned} \lim_{d \rightarrow \infty} \frac{1}{d} \log \mathbb{E}[\det \nabla^2 L(w) | y, \nabla L(w) = 0] \\ = \lim_{\epsilon \rightarrow 0^+} \text{extr}_{(\phi, \bar{\phi}) \in \mathbb{C}^2} \text{Re}[\tilde{F}^{RM}(\phi, \bar{\phi}; \rho, \kappa + i\epsilon)]. \end{aligned} \quad (32)$$

The extremal point of $\tilde{F}^{RM}$ is the unique solution of the fixed point equations obtained from differentiating (30)

$$-z\bar{\phi} = \mathbb{E} \frac{\sigma''(\sqrt{\rho}\zeta)}{1 + \frac{1}{\beta} \sigma''(\sqrt{\rho}\zeta) \phi}, \quad (33)$$

$$-z\phi = \mathbb{E} \frac{\lambda_\Sigma}{1 + \lambda_\Sigma \bar{\phi}}. \quad (34)$$

In the interest of obtaining a closed form for $\mathbb{E} \log |\lambda - \kappa|$, we can expand the real part of $\tilde{F}^{RM}$ for small values of the imaginary part of the argument of $G(z)$, namely $\kappa + i\epsilon$, thus obtaining

$$\begin{aligned} \text{Re}[\tilde{F}^{RM}(\phi, \bar{\phi}; \rho, \kappa + i\epsilon)] \\ = \beta \mathbb{E}_\zeta \log \left| 1 + \frac{1}{\beta} \sigma''(\sqrt{\rho}\zeta)\phi \right| \\ + \mathbb{E}_{\lambda_\Sigma} \log |((\kappa + i\epsilon)(1 + \bar{\phi}\lambda_\Sigma))| + \text{Re}[(\kappa + i\epsilon)\bar{\phi}\phi]. \end{aligned} \quad (35)$$

C. Derivation of F^{SP} : Integration around the sphere

Next, we perform the integration with respect to w . The following proposition describes how the integral can be computed using a large deviation principle. It is based on the observation that $p_{\nabla L(w)|y}(0)$ depends only on the statistical and the geometric order parameters.

Proposition IV.3. Assume that the matrix Σ has a well-defined limiting spectral distribution and let λ_Σ be a random variable distributed according to that distribution. Further assume that the eigenvalues of Σ are bounded away from 0 and ∞ , independently of d . Consider the set $\mathcal{Q} \subset \mathbb{R}^2$ to be the image of $\mathbb{S}^{d-1}$ under the mapping $w \rightarrow (w^\top \Sigma w, w^\top \Sigma^{-1} w)$ and the set

$$\bar{\mathcal{Q}} = \{(\bar{\rho}, \bar{q}) \in \mathbb{R}^2 | (1 - \bar{\rho}\rho - \bar{q}q)I + \bar{\rho}\Sigma + \bar{q}\Sigma^{-1} > 0\}. \quad (36)$$

Then we can define the rate function $I : \mathcal{Q} \rightarrow \mathbb{R}$

$$I(\rho, q) = \sup_{\bar{\rho}, \bar{q} \in \bar{\mathcal{Q}}} \frac{1}{2} \mathbb{E}_{\lambda_\Sigma} \log \left(1 - \bar{\rho}\rho - \bar{q}q + \bar{\rho}\lambda_\Sigma + \bar{q}\frac{1}{\lambda_\Sigma} \right), \quad (37)$$

and the free energy

$$\begin{aligned} F^{SP}(\rho, q; \kappa, f) &= \frac{1}{2} \log(\beta e) - \frac{1}{2} \log(f) - \frac{1}{2d} \log \det(\Sigma) \\ &+ \frac{1}{2} \beta \frac{\kappa^2}{f} \left(\frac{1}{\rho} - q \right) - I(\rho, q). \end{aligned} \quad (38)$$

Given these definitions, for $F : \mathcal{Q} \rightarrow \mathbb{R}$ bounded by an $\mathcal{O}(1)$ constant, we have the following asymptotic result:

$$\begin{aligned} \lim_{d \rightarrow \infty} \frac{1}{d} \log \int_{\mathbb{S}^{d-1}} dw p_{\nabla L(w)|y}(0) e^{dF(w^\top \Sigma w, w^\top \Sigma^{-1} w)} \\ = \sup_{\rho, q \in \mathcal{Q}} (F^{SP}(\rho, q; \kappa, f) + F(\rho, q)), \end{aligned} \quad (39)$$

with κ and f defined in Proposition IV.1.

Proof. To work with the expression of $p_{\nabla L(w)|y}(0)$, we begin with the following consequence of the Woodbury matrix formula:

$$((P_w^\perp R^\top)(P_w^\perp R^\top)^\top - uu^\top)^{-1} = P_w^\perp \Sigma^{-1} P_w^\perp{}^\top. \quad (40)$$

As a result, we have up to leading order in the exponent

$$p_{\nabla L(w)|y}(0) = (2\pi)^{-d/2} \det\left(\frac{1}{m} f \Sigma\right)^{-1/2} e^{-\frac{\kappa^2}{2\rho} u^\top (\frac{1}{m} f \Sigma)^{-1} u}. \quad (41)$$

Recalling that $u = \frac{1}{\sqrt{\rho}} P_w^\perp \Sigma w$, Eq. (41) can be written as

$$\begin{aligned} \frac{1}{d} \log p_{\nabla L(w)|y}(0) &= \frac{1}{2} \log(m) - \frac{1}{2} \log(2\pi f) - \frac{1}{2d} \log \det(\Sigma) \\ &- \frac{\beta \kappa^2}{2\rho^2} w^\top \Sigma P_w^\perp (f \Sigma)^{-1} P_w^\perp \Sigma w, \end{aligned} \quad (42)$$

thus getting

$$\begin{aligned} \frac{1}{d} \log p_{\nabla L(w)|y}(0) \\ = \frac{1}{2} \log(m) - \frac{1}{2} \log(2\pi f) - \frac{1}{2d} \log \det(\Sigma) \\ + \frac{1}{2} \beta \frac{\kappa^2}{f \rho} (1 - \rho w^\top \Sigma^{-1} w). \end{aligned} \quad (43)$$

Since we need to integrate the w dependence over the sphere, we need the joint LDP of $w^\top \Sigma w$ and $w^\top \Sigma^{-1} w$, when w is seen as a uniform random variable on the sphere. The Gaussian analog of this problem is well studied [62], and in the spherical case it reduces to determining the leading order of the volume of the high-dimensional ellipse that is defined by fixing ρ and q . This volume can be computed utilizing the spherical symmetry of Gaussian random variables. For the reader's convenience, we present this computation in Appendix A. Applying Proposition A.1 and using the limiting spectral distribution of Σ , we see that the rate function of $(w^\top \Sigma w, w^\top \Sigma^{-1} w)$ is given by $I(\rho, q)$.

Then, what is left to do is to apply Varadhan's lemma [63] for the function $\log p_{\nabla L(w)|y}(0) + F(w^\top \Sigma w, w^\top \Sigma^{-1} w)$. The conditions of Varadhan's lemma are satisfied because of the boundedness assumption on F and because of assumed bounds on the eigenvalues of Σ . The statement of the proposition thus follows from (43). The constant term comes from the volume of the d -dimensional sphere. ■

Notably, the function inside the supremum in (37) is jointly concave with respect to $\bar{\rho}$ and $\bar{q}$.

D. Derivation of F^S : Large deviations of empirical measures

Finally, we are left with the expectation over y , which appears only as statistics of ζ in the order parameters ℓ , κ , and f and in $\tilde{F}^{RM}$. Since ζ is distributed according to the empirical measure ν_y , we can use Sanov's LDP for empirical measures. According to Sanov's LDP [63], the rate function for ν_y is $-\beta \mathcal{H}(\nu_y | N)$, where N is the one-dimensional normal distribution (since y when unconstrained is Gaussian) and $\mathcal{H}$ is the relative entropy [63]. To find the infimum over the measure ν_y , we need to take into account that the order parameters f , κ , and ℓ are random variables that depend on ν_y through the relations $f = \mathbb{E}_\zeta \sigma'(\zeta)^2$, $\kappa = \mathbb{E}_\zeta \zeta \sigma'(\zeta)$ and $\ell = \mathbb{E}_\zeta \sigma(\zeta)$. In addition, $\tilde{F}^{RM}$ includes the expectation $\mathbb{E}_\zeta \log |1 + \frac{1}{\beta} \sigma''(\sqrt{\rho}\zeta)\phi|$. Hence, to constrain the measure to fixed values of ℓ , κ , and f we introduce auxiliary parameters $\bar{\ell}$, $\bar{\kappa}$ and $\bar{f}$ as Lagrange multipliers, and then minimize ν_y for these parameters fixed, using Sanov's theorem. Subsequently, we can obtain the infimum with respect to these parameters. However, care should be taken, since the values of ϕ , $\bar{\phi}$ depend on these parameters themselves through (33). These considerations lead to the following result, which remains heuristic, because of the nonrigorous exchange between the extremization over ϕ and $\bar{\phi}$ and the minimization over ν .

Result IV.4. Let $\mathcal{K}_{\rho,\beta}$ be the (three-dimensional) space of all measures of the form

$$\nu(d\zeta) = \frac{1}{\mathcal{Z}(\bar{\ell}, \bar{\kappa}, \bar{f}; \rho)} e^{-\frac{1}{2}\zeta^2 - \frac{1}{\beta}F_\nu(\zeta; \bar{\ell}, \bar{\kappa}, \bar{f}, \rho)} d\zeta, \quad (44)$$

where F_ν is given by

$$F_\nu(\zeta; \bar{\ell}, \bar{\kappa}, \bar{f}, \rho) = \bar{\ell}\sigma(\sqrt{\rho}\zeta) + \bar{\kappa}\sqrt{\rho}\zeta\sigma'(\sqrt{\rho}\zeta) + \bar{f}\sigma'(\sqrt{\rho}\zeta)^2, \quad (45)$$

and $\mathcal{Z}$ is defined for each ν by normalization. Furthermore, consider the free energy F^S as defined in (10)

$$F^S(\nu; \ell, \kappa, f, \rho) = \bar{\ell}\ell + \bar{\kappa}\kappa + \bar{f}f + \beta \log \mathcal{Z}(\bar{\ell}, \bar{\kappa}, \bar{f}; \rho). \quad (46)$$

Finally, let $p_S : \mathbb{R}^3 \rightarrow \mathbb{R}$ be the joint probability density of the random variables $\mathbb{E}_\zeta \sigma'(\zeta)^2$, $\mathbb{E}_\zeta \zeta \sigma'(\zeta)$ and $\mathbb{E}_\zeta \sigma(\zeta)$, with these quantities being random due to the underlying randomness of the measure $\nu(d\zeta)$. Then, in the large m limit, we can use

$$\begin{aligned} \log \mathbb{E} \left[\exp \left(\text{extr}_{(\phi, \bar{\phi}) \in \mathbb{C}^2} \text{Re} \left[\tilde{F}^{RM}(\phi, \bar{\phi}; \rho, \kappa + i0^+) \right] \right) | f, \kappa, \ell \right] + \log p_S(f, \kappa, \ell) \\ = \beta \inf_{\nu \in \mathcal{K}_{\rho,\beta}} \left(\text{extr}_{(\phi, \bar{\phi}) \in \mathbb{C}^2} F^{RM}(\phi, \bar{\phi}; \rho, \kappa + i0^+; \nu) + F^S(\ell, \kappa, f, \nu; \rho) \right), \end{aligned} \quad (47)$$

where $\mathbb{E}[\cdot | f, \kappa, \ell]$ denotes the expectation conditional on the equations $f = \mathbb{E}_\zeta \sigma'(\zeta)^2$, $\kappa = \mathbb{E}_\zeta \zeta \sigma'(\zeta)$, and $\ell = \mathbb{E}_\zeta \sigma(\zeta)$. where by $\kappa + i0^+$ we mean that the limit of the imaginary part $\epsilon \rightarrow 0^+$ should be taken after the extr operation. F^{RM} is as defined in Eq. (13).

The change from $\tilde{F}^{RM}$ to F^{RM} presented in (47) results from taking the infimum over ν within the extremization over ϕ and $\bar{\phi}$, in order to calculate the expected value $\mathbb{E} \log |1 + \frac{1}{\beta} \sigma''(\sqrt{\rho}\zeta)\phi|$ in (35). This amounts to replacing that expectation with $\log \mathbb{E} |1 + \frac{1}{\beta} \sigma''(\sqrt{\rho}\zeta)\phi|$ and adding the rate function F^S to the total LDP.

Finally, applying Result IV.2 and then Result IV.4, we can compute the expectation of the determinant conditionally on f, κ, ℓ and $\{\nabla L(w) = 0\}$ through the right-hand side of (47). Then we perform the integration over $\mathbb{S}^{d-1}$ by applying Proposition IV.3. Finally, the result is obtained by taking the expectation over κ, f and ℓ , which amounts to maximizing over their values on the exponent, by a standard application of Laplace's method. Thus, we arrive at Result III.1 and its discussion as presented in Sec. III.

V. GENERALIZED LINEAR MODELS: $L_1(w)$

In this section, we will explain the main changes in the above procedure when calculating $\mathbb{E}\mathcal{N}(B; L_1)$ instead of $\mathbb{E}\mathcal{N}(B; L)$. In this case, in addition to $z_\mu^\top R w$, a second set of Gaussian random variables appears, namely $z_\mu^\top R w_*$. The correlations between these two sets of random variables are nonzero only on the same data index

$$\mathbb{E}[(w^\top R z_\mu)(w_*^\top R z_{\mu'})] = \delta_{\mu\mu'} w_*^\top R R^\top w. \quad (48)$$

Now, the loss can be written in terms of two independent $\mathcal{N}(0, I_m)$ vectors, y and y' , as follows:

$$L_1(w) = \frac{1}{2m} \|\sigma(\sqrt{\rho}y) - \sigma(\sqrt{\rho}y + \sqrt{\rho_* - \rho}y')\|^2. \quad (49)$$

We thus see the appearance of a new order parameter, $\sqrt{\rho} = w_*^\top R R^\top w / \sqrt{\rho}$, as well as the (fixed) variable $\sqrt{\rho_*} = \|R^\top w_*\|$. Some definitions will be useful to proceed. For this section only, let $P_w^\perp$ be the $(d-2) \times d$ matrix projecting to the space spanned by w and w_* . Define then $u = \frac{1}{\sqrt{\rho}} P_w^\perp R^\top R w$ and $v = \frac{1}{\sqrt{\rho_* - \rho}} P_w^\perp R^\top R (w_* - \frac{\sqrt{\rho}}{\sqrt{\rho_* - \rho}} w)$. Finally, we will make use of

the following auxiliary functions:

$$\theta_\sigma(z, z') = (\sigma(z) - \sigma(z'))\sigma'(z), \quad (50)$$

and

$$\psi_\sigma(z, z') = (\sigma(z) - \sigma(z'))\sigma''(z) + \sigma'(z)^2. \quad (51)$$

Then, we have the following analog of Proposition IV.1:

Proposition V.1. Let

$$\kappa = \frac{1}{m} \sum_{\mu=1}^m \theta_\sigma(\sqrt{\rho}y_\mu, \sqrt{\rho}y_\mu + \sqrt{\rho_* - \rho}y'_\mu) \sqrt{\rho}y_\mu, \quad (52)$$

$$\kappa' = \frac{1}{m} \sum_{\mu=1}^m \theta_\sigma(\sqrt{\rho}y_\mu, \sqrt{\rho}y_\mu + \sqrt{\rho_* - \rho}y'_\mu) \sqrt{\rho}y'_\mu, \quad (53)$$

and

$$f = \frac{1}{m} \|\theta_\sigma(\sqrt{\rho}y, \sqrt{\rho}y + \sqrt{\rho_* - \rho}y')\|^2. \quad (54)$$

(Ia): There exists a normal random vector $z' \in \mathbb{R}^{d-1}$ independent of y , such that the gradient of the loss function $L_1(w)$ has the following representation:

$$\begin{aligned} \nabla L_1(w) &= \frac{\sqrt{f}}{\sqrt{m}} (P_w^\perp R^\top R P_w^\perp - uu^\top - vv^\top)^{1/2} z' \\ &\quad + \frac{\kappa}{\sqrt{\rho}} u + \frac{\kappa'}{\sqrt{\rho}} v. \end{aligned} \quad (55)$$

(Ib): There exists a $(d-1) \times m$ random matrix Z' , with i.i.d. normal elements, independent of y , such that the Hessian of the loss function $L_1(w)$ on the sphere, conditionally on the constraint $\{\nabla L_1(w) = 0\}$, can be expressed as

$$\begin{aligned} \nabla^2 L_1(w) \Big|_{\nabla L_1(w)=0} \\ = \frac{1}{m} (P_w^\perp R^\top Z' + \Pi) \Lambda (P_w^\perp R^\top Z' + \Pi)^\top - \kappa I_{d-1}, \end{aligned} \quad (56)$$

where Λ is the $m \times m$ diagonal matrix with $\psi_\sigma(\sqrt{\rho}y_\mu, \sqrt{p}y_\mu + \sqrt{\rho_* - p}y'_\mu)$ as entries, i.e.,

$$\Lambda_{\mu,\mu'} := \delta_{\mu,\mu'} \psi_\sigma(\sqrt{\rho}y_\mu, \sqrt{p}y_\mu + \sqrt{\rho_* - p}y'_\mu), \quad (57)$$

and Π is a matrix of $\mathcal{O}(1)$ rank.

Proof. The proof follows the proof of Proposition IV.1 without any noteworthy deviation. ■

Under this change of Λ , the same random matrix theory result as above is still applicable [61]. Applying it, we obtain the following free energy:

$$\begin{aligned} \tilde{F}_1^{RM}(\phi, \bar{\phi}, \rho, p, \kappa + i\epsilon;) \\ = \mathbb{E}_{\lambda_\Sigma} \log(1 + \bar{\phi}\lambda_\Sigma) \\ + \beta \mathbb{E}_{\zeta, \zeta'} \log \left(1 + \frac{1}{\beta} \psi_\sigma(\sqrt{\rho}\zeta, \sqrt{p}\zeta + \sqrt{\rho_* - p}\zeta') \phi \right) \\ + (\kappa + i\epsilon) \bar{\phi}\phi + \log(\kappa + i\epsilon), \end{aligned} \quad (58)$$

where now the random variable ζ' also appears, which is distributed according to the empirical distribution of the coordinates of y' . In conjunction with similar assumptions on the Hessian and Proposition V.1, we obtain, similarly to Sec. IV B, the following result:

Result V.2.

$$\begin{aligned} \lim_{d \rightarrow \infty} \frac{1}{d} \log \mathbb{E} [|\nabla^2 L_1(w)| | y, \nabla L_1(w) = 0] \\ = \lim_{\epsilon \rightarrow 0^+} \text{extr}_{(\phi, \bar{\phi}) \in \mathbb{C}^2} \text{Re} [\tilde{F}_1^{RM}(\kappa + i\epsilon; \phi, \bar{\phi})]. \end{aligned} \quad (59)$$

As new geometric parameters have appeared in the distribution of ∇L_1 , to perform the integration over w , we need a joint LDP for ρ, q, p , as well as $r = w_*^\top w$. This LDP can be characterized through Proposition A.1. Specifically, we obtain the following proposition:

Proposition V.3. Assume that the matrix Σ has a well-defined limiting spectral distribution, and let λ_Σ be a random variable distributed according to that distribution. Further, assume that the eigenvalues of Σ are bounded away from 0 and ∞ , independently of d . Consider the set $\mathcal{Q}_1 \subset \mathbb{R}^4$ to be the image of $\mathbb{S}^{d-1}$ under the mapping $w \rightarrow (w^\top \Sigma w, w^\top \Sigma^{-1} w, w_*^\top R R^\top w / \|R^\top w\|, w_*^\top w)$ and the set

$$\begin{aligned} \bar{\mathcal{Q}}_1 = \left\{ (\bar{\rho}, \bar{q}, \bar{p}, \bar{r}) \in \mathbb{R}^4 \mid (1 - \bar{\rho}\rho - \bar{q}q - \bar{r}r^2 - \bar{p}p)I + \bar{\rho}\Sigma \right. \\ \left. + \bar{q}\Sigma^{-1} - \bar{r}w_* w_*^\top - \frac{\bar{p}}{\rho}(\Sigma w_* w_*^\top \Sigma) \succ 0 \right\}. \end{aligned} \quad (60)$$

Then, we can define the rate function $I : \mathcal{Q}_1 \rightarrow \mathbb{R}$

$$\begin{aligned} I(\rho, q, p, r) \\ = \sup_{\bar{\rho}, \bar{q}, \bar{p}, \bar{r} \in \bar{\mathcal{Q}}_1} \frac{1}{2d} \text{tr} \log \left[(1 - \bar{\rho}\rho - \bar{q}q - \bar{r}r^2 - \bar{p}p)I \right. \\ \left. + \bar{\rho}\Sigma + \bar{q}\Sigma^{-1} - \bar{r}w_* w_*^\top - \frac{\bar{p}}{\rho}(\Sigma w_* w_*^\top \Sigma) \right], \end{aligned} \quad (61)$$

and the free energy

$$\begin{aligned} F_1^{SP}(\rho, q, r, p; \kappa, \kappa', f) \\ = \frac{1}{2} \log(\beta e) - \frac{1}{2d} \log \det(\Sigma) \end{aligned}$$

$$\begin{aligned} + \frac{1}{2} \beta \frac{1}{f\rho} \left((1 - q\rho)\kappa^2 + \kappa'^2 - 2 \frac{r\sqrt{\rho} - \sqrt{p}}{\sqrt{\rho_* - p}} \kappa\kappa' \right) \\ - I(\rho, q, p, r). \end{aligned} \quad (62)$$

Given these definitions, for a function F bounded by an $\mathcal{O}(1)$ constant with $F : \mathcal{Q}_1 \rightarrow \mathbb{R}$, we have the following asymptotic result:

$$\begin{aligned} \lim_{d \rightarrow \infty} \frac{1}{d} \log \int_{\mathbb{S}^{d-1}} dw p_{\nabla L_1(w)|y}(0) \\ \times e^{dF(w^\top \Sigma w, w^\top \Sigma^{-1} w, w_*^\top R R^\top w / \|R^\top w\|, w_*^\top w)} \\ = \sup_{\rho, q, p, r \in \mathcal{Q}_1} (F_1^{SP}(\rho, q, r, p; \kappa, \kappa', f) + F(\rho, q, p, r)). \end{aligned} \quad (63)$$

Proof. The proof begins with the observation that by the Woodbury formula,

$$(P_w^\perp R^\top (P_w^\perp R^\top)^\top - uu^\top - vv^\top)^{-1} = P_w^\perp \Sigma^{-1} P_w^{\perp\top}. \quad (64)$$

Plugging this into the density that corresponds to the representation of $\nabla L_1(w)$ provided by Proposition V.1, we obtain, to leading order in d ,

$$\begin{aligned} \frac{1}{d} \log p_{\nabla L_1(w)|y}(0) \\ = \frac{1}{2} \log(m) - \frac{1}{2} \log(2\pi f) - \frac{1}{2d} \log \det(\Sigma) + \frac{1}{2} \beta \frac{1}{f\rho} \\ \times \left((1 - q\rho)\kappa^2 + \kappa'^2 - 2 \frac{(w_*^\top w)\sqrt{\rho} - \sqrt{p}}{\sqrt{\rho_* - p}} \kappa\kappa' \right). \end{aligned} \quad (65)$$

From here, the proof proceeds with application of Proposition A.1 without significant differences from the proof of Proposition V.3. ■

Finally, to complete the large deviation analysis, one has to take into account the joint deviation of the densities of ζ and ζ' , and specifically with respect to the loss, ℓ , as well as κ, κ', f and the expectation value appearing in (58). Utilizing Sanov's theorem [63] as we did in Sec. IV D, the random matrix free energy can be expressed through the following Result:

Result V.4. Let $\mathcal{K}_{\rho,p}$ be the (four-dimensional) space of all measures of the form

$$\begin{aligned} \nu(d\zeta d\zeta') = \frac{1}{\mathcal{Z}_1(\bar{\ell}, \bar{\kappa}, \bar{\kappa}', \bar{f}; \rho, p)} \exp \left\{ -\frac{1}{2} \zeta^2 - \frac{1}{2} \zeta'^2 \right. \\ \left. - \frac{1}{\beta} F_{1\nu}(\zeta, \zeta'; \bar{\ell}, \bar{\kappa}, \bar{\kappa}', \bar{f}, \rho, p) \right\} d\zeta d\zeta', \end{aligned} \quad (66)$$

where $F_{1\nu}$ is given by

$$F_{1\nu}(\zeta, \zeta'; \bar{\ell}, \bar{\kappa}, \bar{\kappa}', \bar{f}, p, \rho) = \bar{\ell}(\sigma(\sqrt{\rho}\zeta) - \sigma(\sqrt{p}\zeta + \sqrt{\rho_* - p}\zeta'))^2 + \bar{f}(\theta_\sigma(\sqrt{\rho}\zeta, \sqrt{p}\zeta + \sqrt{\rho_* - p}\zeta'))^2 \\ + (\bar{\kappa}\sqrt{\rho}\zeta + \bar{\kappa}'\sqrt{\rho}\zeta')\theta_\sigma(\sqrt{\rho}\zeta, \sqrt{p}\zeta + \sqrt{\rho_* - p}\zeta'), \quad (67)$$

and $\mathcal{Z}_1$ is defined for each ν by normalization. Furthermore, consider the free energy F_1^S defined through

$$F_1^S(\ell, \kappa, \kappa', f, \nu; \rho, p) = \bar{f}f + \bar{\kappa}\kappa + \bar{\kappa}'\kappa' + \bar{\ell}\ell + \beta \log \mathcal{Z}_1(\bar{\ell}, \bar{\kappa}, \bar{\kappa}', \bar{f}, p, \rho). \quad (68)$$

Finally, let $p_S: \mathbb{R}^4 \rightarrow \mathbb{R}$ be the joint probability density of the random variables $\mathbb{E}_{\zeta, \zeta'} \theta_\sigma(\sqrt{\rho}\zeta, \sqrt{p}\zeta + \sqrt{\rho_* - p}\zeta')^2$, $\mathbb{E}_{\zeta, \zeta'} \theta_\sigma(\sqrt{\rho}\zeta, \sqrt{p}\zeta + \sqrt{\rho_* - p}\zeta')\sqrt{\rho}\zeta$, $\mathbb{E}_{\zeta, \zeta'} \theta_\sigma(\sqrt{\rho}\zeta, \sqrt{p}\zeta + \sqrt{\rho_* - p}\zeta')\sqrt{\rho}\zeta'$, and $\mathbb{E}_{\zeta, \zeta'} (\sigma(\sqrt{\rho}\zeta) - \sigma(\sqrt{p}\zeta + \sqrt{\rho_* - p}\zeta'))^2$. Then, we have

$$\log \mathbb{E}[\exp(\text{extr}_{(\phi, \bar{\phi}) \in \mathbb{C}^2} \text{Re}[\tilde{F}_1^{RM}(\phi, \bar{\phi}; \rho, p, \kappa + i0^+)]) | f, \kappa, \kappa', \ell] + \log p_S(f, \kappa, \kappa', \ell) \\ = \beta \inf_{\nu \in \mathcal{K}_{\rho, p}} (\text{extr}_{(\phi, \bar{\phi}) \in \mathbb{C}^2} F_1^{RM}(\phi, \bar{\phi}; \rho, p, \kappa + i0^+; \nu) + F_1^S(\ell, \kappa, \kappa', f, \nu; \rho, p)), \quad (69)$$

where F_1^{RM} is defined by

$$F_1^{RM}(\phi, \bar{\phi}; \rho, p, \kappa + i\epsilon; \nu) \\ = \mathbb{E}_{\lambda_\Sigma} \log[(\kappa + i\epsilon)(1 + \bar{\phi}\lambda_\Sigma)] \\ + \beta \log \mathbb{E}_{\zeta, \zeta'} \left[1 + \frac{1}{\beta} \psi_\sigma(\sqrt{\rho}\zeta, \sqrt{p}\zeta + \sqrt{\rho_* - p}\zeta')\phi \right] \\ + (\kappa + i\epsilon)\bar{\phi}\phi, \quad (70)$$

and $\mathbb{E}[\cdot | f, \kappa, \kappa', \ell]$ denotes the expectation conditional on the equations $f = \mathbb{E}_{\zeta, \zeta'} \theta_\sigma(\sqrt{\rho}\zeta, \sqrt{p}\zeta + \sqrt{\rho_* - p}\zeta')^2$, $\kappa = \mathbb{E}_{\zeta, \zeta'} \theta_\sigma(\sqrt{\rho}\zeta, \sqrt{p}\zeta + \sqrt{\rho_* - p}\zeta')\sqrt{\rho}\zeta$, $\kappa' = \mathbb{E}_{\zeta, \zeta'} \theta_\sigma(\sqrt{\rho}\zeta, \sqrt{p}\zeta + \sqrt{\rho_* - p}\zeta')\sqrt{\rho}\zeta'$ and $\ell = \mathbb{E}_{\zeta, \zeta'} (\sigma(\sqrt{\rho}\zeta) - \sigma(\sqrt{p}\zeta + \sqrt{\rho_* - p}\zeta'))^2$.

Combining Proposition V.3 with Results V.2 and V.4, as we did in Sec. IV D, we obtain Result III.2.

VI. DISCRIMINATING PERCEPTRON: $L_2(w)$

We will now briefly present the significant steps of the calculation of $\mathbb{E}\mathcal{N}(U, B; L_2)$. Here, the distribution of L_2 is of the form

$$L_2(w) = \frac{1}{m_1} \sum_{\mu=1}^{m_1} \sigma_1(\sqrt{\rho_1}y_{1\mu}) - \frac{1}{m_2} \sum_{\mu=1}^{m_2} \sigma_2(\sqrt{\rho_2}y_{2\mu}), \quad (71)$$

where now for $j = 1, 2$, y_j is an m_j -dimensional normally distributed vector and $\sqrt{\rho_j} := \|R_j^\top w\|$. It should be pointed out that in this section, all variables that were defined and appearing in the analysis of Sec. IV (with the exception of ϵ and q) appear in pairs, such as ρ_1 and ρ_2 . When possible, we will refer to them jointly, dropping the index 1,2, e.g., $\rho = (\rho_1, \rho_2)$. Let $P_w^\perp$ be the $(d-1) \times d$ matrix that projects orthogonally to the space spanned by w . For $j = 1, 2$, let $u_j = \frac{1}{\sqrt{\rho_j}} P_w^\perp R_j^\top R_j w$ and let $Z_j \in \mathbb{R}^{d \times m_j}$ be the random matrices formed by stacking the z_μ vectors similarly to the previous section. Now, the distribution of the gradient and the conditioned Hessian is given by the following Proposition:

Proposition VI.1. For $j = 1, 2$, let $\kappa_j = \frac{1}{m_j} \sqrt{\rho_j} y_j^\top \sigma_j'(\sqrt{\rho_j} y_j)$ and $f_j = \frac{1}{m_j} \|\sigma_j'(\sqrt{\rho_j} y_j)\|^2$.

(1a): For $j = 1, 2$, there exists normal random vectors $z_j' \in \mathbb{R}^{d-1}$, independent of each other and of y_1 and y_2 , such that the gradient of the loss function $L_2(w)$ has the following

representation:

$$\nabla L_2(w) = \frac{1}{\sqrt{m_1}} \sqrt{f_1} ((P_w^\perp R_1^\top) (P_w^\perp R_1^\top)^\top - u_1 u_1^\top)^{1/2} z_1' \\ - \frac{1}{\sqrt{m_2}} \sqrt{f_2} ((P_w^\perp R_2^\top) (P_w^\perp R_2^\top)^\top - u_2 u_2^\top)^{1/2} z_2' \\ + \frac{1}{m_1} y_1^\top \sigma_1'(\sqrt{\rho_1} y_1) u_1 - \frac{1}{m_2} y_2^\top \sigma_2'(\sqrt{\rho_2} y_2) u_2. \quad (72)$$

(1b): For $j = 1, 2$, there exist $(d-1) \times m$ random matrices Z_j' , with i.i.d. normal elements, independent of each other and of y_1 and y_2 , such that the Hessian of the loss function $L_2(w)$ on the sphere, conditionally on the constraint $\{\nabla L(w) = 0\}$, can be expressed as

$$\nabla^2 L_2(w) \Big|_{\nabla L_2(w)=0} \\ = \frac{1}{m_1} (P_w^\perp R_1^\top Z_1' + \Pi_1) \Lambda_1 (P_w^\perp R_1^\top Z_1' + \Pi_1)^\top - \frac{1}{m_2} (P_w^\perp R_2^\top Z_2' \\ + \Pi_2) \Lambda_2 (P_w^\perp R_2^\top Z_2' + \Pi_2)^\top - (\kappa_1 - \kappa_2) I_{d-1}, \quad (73)$$

where for $j = 1, 2$, Λ_j is the $m \times m$ diagonal matrix with $\sigma_j''(\sqrt{\rho_j} y_{j\mu})$ as entries, i.e., $\Lambda_{j, \mu, \mu'} := \delta_{\mu, \mu'} \sigma_j''(\sqrt{\rho_j} y_{j\mu})$, and Π_j is a matrix of rank at most $\mathcal{O}(1)$.

Proof. The proof follows the proof of Proposition IV.1. ■

This competitive structure in the Hessian necessitates a more general random matrix theory result, which can be found in [56,64]. Applying this result, as we did in Sec. IV B, we obtain the following auxiliary free energy:

$$\tilde{F}_2^{RM}(\phi, \bar{\phi}; \rho, \kappa, \epsilon) \\ = \frac{\beta}{\alpha_1} \mathbb{E}_{\zeta_1} \log \left(1 + \frac{\alpha_1}{\beta} \sigma_1''(\sqrt{\rho_1} \zeta_1) \phi_1 \right) \\ + \frac{\beta}{\alpha_2} \mathbb{E}_{\zeta_2} \log \left(1 - \frac{\alpha_2}{\beta} \sigma_2''(\sqrt{\rho_2} \zeta_2) \phi_2 \right) \\ + \frac{1}{d} \log \det((\kappa_1 - \kappa_2 + i\epsilon)(I + \bar{\phi}_1 G - \bar{\phi}_2 \Sigma)) \\ + (\kappa_1 - \kappa_2 + i\epsilon)(\bar{\phi}_1 \phi_1 - \bar{\phi}_2 \phi_2), \quad (74)$$

where the vectors $\phi, \bar{\phi}, \kappa, \rho$ represent the pair variables $\phi = (\phi_1, \phi_2)$, $\bar{\phi} = (\bar{\phi}_1, \bar{\phi}_2)$, $\kappa = (\kappa_1, \kappa_2)$, $\rho = (\rho_1, \rho_2)$, respectively, for compactness, and where we introduced the random variables ζ_1, ζ_2 distributed, respectively, according to $\nu_1(y_1) = \sum_{\mu=1}^{m_1} \delta_{y_{1\mu}}$ and $\nu_2(y_2) = \sum_{\mu=1}^{m_2} \delta_{y_{2\mu}}$. Proceeding with similar assumptions as we did in the previous sections and applying the capacity calculation described above, we obtain the following result:

Result VI.2.

$$\lim_{d \rightarrow \infty} \frac{1}{d} \log \mathbb{E} [|\nabla^2 L_2(w)| | y, \nabla L_2(w) = 0] \\ = \lim_{\epsilon \rightarrow 0^+} \text{extr}_{\phi, \bar{\phi} \in \mathbb{C}^2} \text{Re} [\tilde{F}_2^{RM}(\phi, \bar{\phi}; \rho, \kappa, \epsilon)]. \quad (75)$$

We now move to the geometric order parameters. Consider the following matrices:

$$S = \alpha_1 f_1 G + \alpha_2 f_2 \Sigma, \quad (76)$$

$$D = \frac{\kappa_1}{\rho_1} G - \frac{\kappa_2}{\rho_2} \Sigma, \quad (77)$$

$$B = \frac{\sqrt{\rho_2}}{\sqrt{\rho_1}} G - \frac{\sqrt{\rho_1}}{\sqrt{\rho_2}} \Sigma. \quad (78)$$

We need the joint rate function I_2 for the variables ρ and q , which has elements

$$q_1 = w^\top D S^{-1} D w, \quad (79)$$

$$q_2 = w^\top S^{-1} w, \quad (80)$$

$$q_3 = w^\top S^{-1} D w, \quad (81)$$

$$q_4 = w^\top B S^{-1} B w, \quad (82)$$

$$q_5 = w^\top D S^{-1} B w, \quad (83)$$

$$q_6 = w^\top S^{-1} B w. \quad (84)$$

The rate function for the joint LDP of q and ρ can be captured by Proposition A.1, resulting in the following:

Proposition VI.3. Assume that the matrices Σ and G have well-defined limiting spectral distributions, and let λ_Σ and λ_G be random variables distributed according to these distributions, respectively. Further assume that the eigenvalues of Σ and G are bounded away from 0 and ∞ , independently of d . Consider the set $\mathcal{Q}_2 \subset \mathbb{R}^8$ to be the image of $\mathbb{S}^{d-1}$ under the mapping $w \rightarrow (w^\top \Sigma w, w^\top G w, w^\top D S^{-1} D w, w^\top S^{-1} w, w^\top S^{-1} D w, w^\top B S^{-1} B w, w^\top D S^{-1} B w, w^\top S^{-1} B w)$. Correspondingly, define the vectors $\bar{\rho} = (\bar{\rho}_1, \bar{\rho}_2)^\top$ and $\bar{q} = (\bar{q}_1, \dots, \bar{q}_6)^\top$ and the matrix

$$M(\rho, q; \bar{\rho}, \bar{q}) \\ = (1 - \bar{\rho}^\top \rho - \bar{q}^\top q) I + \bar{\rho}_1 G + \bar{\rho}_2 \Sigma \\ + \bar{q}_1 D S^{-1} D + \bar{q}_2 S^{-1} + \frac{\bar{q}_3}{2} (S^{-1} D + D S^{-1}) + \bar{q}_4 B S^{-1} B \\ + \frac{\bar{q}_5}{2} (D S^{-1} B + B S^{-1} D) + \frac{\bar{q}_6}{2} (S^{-1} B + B S^{-1}), \quad (85)$$

and the set

$$\bar{\mathcal{Q}}_2 = \{(\bar{\rho}, \bar{q}) \in \mathbb{R}^8 | M(\rho, q; \bar{\rho}, \bar{q}) \succ 0\}. \quad (86)$$

Then, we can define the rate function $I : \mathcal{Q}_2 \rightarrow \mathbb{R}$

$$I_2(\rho, q; f, \kappa) = \sup_{\bar{\rho}, \bar{q} \in \bar{\mathcal{Q}}_2} \left(\frac{1}{2d} \log \det [M(\rho, q; \bar{\rho}, \bar{q})] \right), \quad (87)$$

and the free energy

$$F_2^{SP}(\rho, q; \kappa, f) \\ = \frac{1}{2} \log(\beta e) - \frac{1}{2d} \log \det S - \frac{\beta}{2} q_1 \\ - \frac{\beta}{2} q_2 (\kappa_1 - \kappa_2)^2 + \beta q_3 (\kappa_1 - \kappa_2) \\ - \frac{\beta}{2} \frac{(\alpha_1 f_1 \alpha_2 f_2)(q_5 + (\kappa_1 - \kappa_2) q_6)^2}{\alpha_1 f_1 \rho_1 + \alpha_2 f_2 \rho_2 - (\alpha_1 f_1 \alpha_2 f_2) q_4} \\ - I_2(q, \rho; \kappa, f). \quad (88)$$

Given these definitions, for $F : \mathcal{Q}_2 \rightarrow \mathbb{R}$ bounded by an $\mathcal{O}(1)$ constant, we have the following asymptotic result:

$$\lim_{d \rightarrow \infty} \frac{1}{d} \log \int_{\mathbb{S}^{d-1}} dw p_{\nabla L_2(w)|y}(0) \\ \times e^{dF(w^\top \Sigma w, w^\top G w, w^\top D S^{-1} D w, w^\top S^{-1} w, w^\top S^{-1} D w)} \\ = \sup_{\rho, q \in \mathcal{Q}_2} (F_2^{SP}(\rho, q; \kappa, f) + F(\rho, q)). \quad (89)$$

Proof. We need to express the inverse of the covariance function of $\nabla L_2(w)|_y$ in terms of S^{-1} . This can be performed by using the Woodbury inverse formula and its special form, the Sherman-Morrison formula, to get

$$P_w^\perp \left(\alpha_1 f_1 (P_w^\perp G P_w^\perp - u_1 u_1^\top) + \alpha_2 f_2 (P_w^\perp \Sigma P_w^\perp - u_2 u_2^\top) \right)^{-1} P_w^{\perp\top} \\ = P_w^\perp S^{-1} P_w^{\perp\top} + \gamma P_w^\perp S^{-1} (\sqrt{\rho_2} u_1 - \sqrt{\rho_1} u_2) \\ \times (\sqrt{\rho_2} u_1 - \sqrt{\rho_1} u_2)^\top S^{-1} P_w^{\perp\top}, \quad (90)$$

where

$$\gamma = \frac{\alpha_1 f_1 \alpha_2 f_2}{\alpha_1 f_1 \rho_1 + \alpha_2 f_2 \rho_2 - (\alpha_1 f_1 \alpha_2 f_2) w^\top B S^{-1} B w}. \quad (91)$$

Accordingly, the probability density that $\nabla L_2(w)$ vanishes is given to leading order by

$$\frac{1}{d} \log \phi_{\nabla L_2|y}(0) \\ = \frac{1}{2} \log(m_1 + m_2) - \frac{1}{2} \log(2\pi) \\ - \frac{1}{2d} \log \det(S) - \frac{\beta}{2} w^\top D^\top S^{-1} D w \\ - \frac{\beta}{2} (w^\top S^{-1} w)(\kappa_1 - \kappa_2)^2 + \beta (w^\top S^{-1} D w)(\kappa_1 - \kappa_2) \\ - \frac{\beta}{2} \gamma (w^\top D^\top S^{-1} B w - (\kappa_1 - \kappa_2) w^\top S^{-1} B w)^2. \quad (92)$$

From here, we apply Proposition A.1 as we did in the proof of Proposition IV.3. The rest of the proof has no significant technical difficulty beyond what is presented in the proof of Proposition IV.3. ■

Throughout the calculation until now, the random vectors y_1 and y_2 determine only averages of functions of ζ_1 and ζ_2

through the pair of κ variables, the pair of f variables and the two expectations in (74). Using Sanov's LDP [63], we can get a rate function for these statistics and the loss, ℓ . Specifically, we obtain a Sanov free energy and a modification of the RMT free energy

$$\begin{aligned} F_2^{RM}(\phi, \bar{\phi}; \rho, \kappa, \epsilon; \nu) &= \frac{\beta}{\alpha_1} \log \mathbb{E}_{\zeta_1} \left| 1 + \frac{\alpha_1}{\beta} \sigma_1''(\sqrt{\rho_1} \zeta_1) \phi_1 \right| \\ &+ \frac{\beta}{\alpha_2} \log \mathbb{E}_{\zeta_2} \left| 1 - \frac{\alpha_2}{\beta} \sigma_2''(\sqrt{\rho_2} \zeta_2) \phi_2 \right| \\ &+ \frac{1}{d} \log |\det((\kappa_1 - \kappa_2 + i\epsilon)(I + \bar{\phi}_1 G - \bar{\phi}_2 \Sigma))| \\ &+ \text{Re}((\kappa_1 - \kappa_2 + i\epsilon)(\bar{\phi}_1 \phi_1 - \bar{\phi}_2 \phi_2)), \end{aligned} \quad (93)$$

where $\nu = (\nu_1, \nu_2)$ depicts the dependence on the distributions of ζ_1 and ζ_2 , which appears as described in the following:

$$F_2^S(\ell, \kappa, f, \nu; \rho) = \bar{\ell} \ell + \bar{f}_1 f_1 + \bar{f}_2 f_2 + \bar{\kappa}_1 \kappa_1 + \bar{\kappa}_2 \kappa_2 + \frac{\beta}{\alpha_1} \log \mathcal{Z}_{1, \beta/\alpha_1}(\bar{\ell}, \bar{f}_1, \bar{\kappa}_1; \rho_1) + \frac{\beta}{\alpha_2} \log \mathcal{Z}_{2, \beta/\alpha_2}(-\bar{\ell}, \bar{f}_2, \bar{\kappa}_2; \rho_2). \quad (97)$$

Finally, let $p_S : \mathbb{R}^5 \rightarrow \mathbb{R}$ be the joint probability density of the random variables $\mathbb{E}_{\zeta_j}(\sigma_j'(\sqrt{\rho_j} \zeta_j))^2$ and $\mathbb{E}_{\zeta_j} \sqrt{\rho_j} \zeta_j \sigma_j'(\sqrt{\rho_j} \zeta_j)$ for $j = 1, 2$ and $\mathbb{E}_{\zeta_1} \sigma_1(\sqrt{\rho_1} \zeta_1) - \mathbb{E}_{\zeta_2} \sigma_2(\sqrt{\rho_2} \zeta_2)$. Then, we have

$$\begin{aligned} &\log \mathbb{E} \left[\exp \left(\text{extr}_{\phi, \bar{\phi} \in \mathbb{C}^2} \text{Re}[\tilde{F}_2^{RM}(\phi, \bar{\phi}; \rho, \kappa, 0^+)] \right) \middle| f, \kappa, \ell \right] + \log p_S(f, \kappa, \ell) \\ &= \beta \inf_{\nu \in \mathcal{K}_{\rho, \alpha, \beta}} \left(\text{extr}_{\phi, \bar{\phi} \in \mathbb{C}^2} F_2^{RM}(\phi, \bar{\phi}; \rho, \kappa, 0^+; \nu) + F_2^S(\ell, \kappa, f, \nu; \rho) \right), \end{aligned} \quad (98)$$

where $\mathbb{E}[\cdot | f, \kappa, \ell]$ denotes the expectation conditional on the equations $\kappa_j = \mathbb{E}_{\zeta_j}(\sigma_j'(\sqrt{\rho_j} \zeta_j))^2$ and $f_j = \mathbb{E}_{\zeta_j} \sqrt{\rho_j} \zeta_j \sigma_j'(\sqrt{\rho_j} \zeta_j)$ for $j = 1, 2$ and $\ell = \mathbb{E}_{\zeta_1} \sigma_1(\sqrt{\rho_1} \zeta_1) - \mathbb{E}_{\zeta_2} \sigma_2(\sqrt{\rho_2} \zeta_2)$. The notation 0^+ signifies the limit $\lim_{\epsilon \rightarrow 0^+}$, and F_2^{RM} is as defined in Eq. (93).

Putting together Proposition VI.3, Result VI.2, and Result VI.4, we arrive at Result III.3.

VII. CONCLUSION/DISCUSSION

In conclusion, we have applied the methodology introduced in [6] to the case of the landscape complexity of empirical risk of a number of generalized linear models, all in the case where the high-dimensional input data are correlated random Gaussian vectors. The motivation behind correlations is the fact that real-life data have an underlying structure, including sparsity. As a simpler problem, we started by analyzing in detail the loss function of the form given in (1), focusing on the presence of the covariance matrix. Subsequently, we included a preferred direction in the sphere, corresponding to the planted vector, in a teacher-student setting, and finally, we introduced a loss model corresponding to a discriminating perceptron, where two sets of (differently) correlated data vectors are fed to the perceptron, which tries to discriminate between them. In the analysis of the first two models, an important generalization to previous results is the fact that, due to the presence of the data correlations, the Hessian is now a doubly correlated Wishart matrix. In contrast, the third discriminating perceptron model has a Hessian with a difference of two doubly correlated Wishart matrices.

Result VI.4. Let $\mathcal{K}_{\rho, \alpha, \beta}$ be the (five-dimensional) space of all pairs of measures $\nu = (\nu_1, \nu_2)$ of the form

$$\nu_1(d\zeta_1) = \frac{1}{\mathcal{Z}_{1, \beta/\alpha_1}(\bar{\ell}, \bar{\kappa}_1, \bar{f}_1; \rho_1)} e^{-\frac{1}{2} \zeta_1^2 - \frac{\alpha_1}{\beta} F_{1, \nu_1}(\zeta_1; \bar{\ell}, \bar{\kappa}_1, \bar{f}_1, \rho_1)} d\zeta_1, \quad (94)$$

$$\begin{aligned} \nu_2(d\zeta_2) &= \frac{1}{\mathcal{Z}_{2, \beta/\alpha_2}(-\bar{\ell}, \bar{\kappa}_2, \bar{f}_2; \rho_2)} \\ &\times e^{-\frac{1}{2} \zeta_2^2 - \frac{\alpha_2}{\beta} F_{2, \nu_2}(\zeta_2; -\bar{\ell}, \bar{\kappa}_2, \bar{f}_2, \rho_2)} d\zeta_2, \end{aligned} \quad (95)$$

where the F_{j, ν_j} are given for each j by

$$\begin{aligned} F_{j, \nu_j}(\zeta_j; \bar{\ell}, \bar{\kappa}_j, \bar{f}_j, \rho_j) &= \bar{\ell} \sigma_j(\sqrt{\rho_j} \zeta_j) \\ &+ \bar{\kappa}_j \sqrt{\rho_j} \zeta_j \sigma_j'(\sqrt{\rho_j} \zeta_j) + \bar{f}_j \sigma_j'(\sqrt{\rho_j} \zeta_j)^2, \end{aligned} \quad (96)$$

and $\mathcal{Z}_{j, \beta/\alpha_j}$ is defined for each ν_j by normalization. Furthermore, consider the free energy F_2^S defined through

To check the validity of our methodology results, we have analytically evaluated the complexity for the loss $L(w)$ discussed in Sec. IV with a quadratic activation function, $\sigma(x) = \frac{x^2}{2}$. In this case, the condition for a critical point on the sphere, namely $\nabla L(w) = 0$ is simply the condition that w is an eigenvector of the (semicorrelated) Wishart matrix $R^\top Z Z^\top R$ (with $\Lambda = I$). Indeed, expressing the projection $P_w^\perp$ as $P_w^\perp = I - w w^\top$, we have that $R Z Z^\top R^\top w = w w^\top R Z Z^\top R^\top w = 2\ell w$. Hence, for $\ell \in (\frac{\lambda_{\min}}{2}, \frac{\lambda_{\max}}{2})$, there are $O(m)$ critical points in its neighborhood, where $\lambda_{\min, \max}$ are the edges of the bulk spectrum of the above matrix. As a result, the complexity, which is the normalized logarithm of the number of critical points, vanishes in this region. Outside this region, the probability of finding a macroscopic number of eigenvalues vanishes exponentially, and hence the complexity is negative. This behavior can be readily seen in Fig. 1. It should be noted that this activation function is simpler than the general case, because now $\Lambda(y) = I$ and hence the optimization over the LDP measure $\nu(d\zeta)$ is independent from F^{RM} and thus its optimal is a Gaussian, since we have $\ell = \kappa/2 = f/2$. Nevertheless, its calculation is not entirely trivial, since one still has to calculate the constrained $p_{\nabla L|y}(0)$, as in Sec. IV C.

Additionally, in the case of identity covariance $\Sigma = I$, our result given by Eq. (14) reduces to the result that is presented in [6] after a heuristic derivation starting from their main theorem. Specifically, notice that the jointly constrained domain of ρ and q forces both of those variables to take the value 1 in the case of identity covariance, as they are expected to. As a result F^{SP} is reduced to the first two terms appearing in (11).

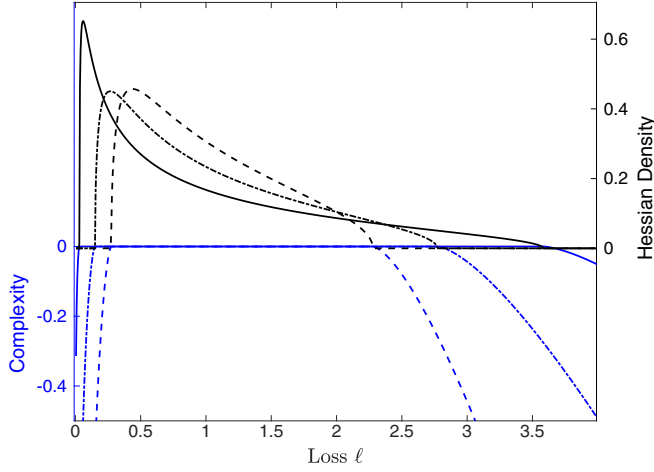


FIG. 1. Complexity function $\Psi(\ell, L)$ (left lower axis and curves in blue) for the case where $\sigma(x) = \frac{x^2}{2}$ as a function of the loss ℓ for three values of $\beta = 1.5, 3, 6$ (solid, dashed, and dash-dot lines, respectively, in blue). These results were obtained after solving (12), (13), and (14) numerically. For comparison, (in black, for the same values of β) we also plot the bulk density of eigenvalues of the loss function in this case, which is equal to $L(w) = \frac{1}{2}w^T R^\top Z^T Z R w$ as a function of its value ℓ , which, when w is an eigenvector, equals half the corresponding eigenvalue. We see that the complexity is exactly zero when the density of eigenvalues is positive, while it is negative when the density vanishes. For concreteness, we used a uniform distribution of $\lambda_\Sigma \in (1, 3)$ [65].

Furthermore, one can use (33) to eliminate $\bar{\phi}$ from F^{RM} in (13).

An important output of our analysis is that the data correlations introduce geometrical constraints on the sphere. Specifically, in the simpler first case we analyzed, corresponding to loss function given by (1), we need to obtain a joint LDP for the quantities $\rho = w^T \Sigma w$ and $q = w^T \Sigma^{-1} w$, which creates nonuniformities in the space of w . This effect is exacerbated in the other two models we analyzed. In the case of $L_1(w)$, the existence of a preferred direction w_* , introduces two additional overlaps that need to be taken into account, namely $p = (w_*^T \Sigma w)^2 / \rho$ and of course $r = w_*^T w$. The more interesting situation arises in the case of the discriminatory perceptron, in which, although without a preferred direction, we need to find joint LDPs, not only for the quantities $\rho_1 = w^T \Sigma w$ and $\rho_2 = w^T G w$, but also for six geometric objects corresponding to ellipsoids, which combine both correlation matrices Σ and G in nontrivial ways. Hence, the optimization procedure over such constrained geometries is very likely nonconvex.

Interestingly, a hypothesis has been formulated in [53], according to which the above analysis can provide a link between the empirical and the generalization loss of each model. For simplicity, we discuss the case of the first model with loss function $L(w)$. When the value of ℓ is such that the critical points of the model have a Hessian with a finite fraction of negative eigenvalues, a local algorithm such as gradient descent and its variants will easily find directions to descend to lower values of ℓ . However, at the value of ℓ , where the bulk of the Hessian spectrum becomes non-negative (notwithstanding

a finite number of negative eigenvalues, with eigendirections that are hard to find), local algorithms are expected to slow down and, at worst, get stuck. Based on the analysis in this paper, that value of ℓ^* will correspond to a specific value of $\rho = w^T \Sigma w$. At the same time, we know that the generalization error $\mathbb{E}_z \sigma(w^T \Sigma^{1/2} z)$ can be expressed uniquely in terms as $\mathbb{E}_x \sigma(x \sqrt{\rho})$, where x is a unit variance Gaussian variable. Hence, the generalization error can be estimated uniquely at this value of $\rho(\ell^*)$. Similar considerations can be made for the generalization errors of the other models discussed here, which depend on the values of other “order parameters” at the value of ℓ , where the system gets stuck. This framework, while tempting, has been called into question by recent advances in the understanding of dynamics in spin glasses, and the significance of the threshold value ℓ^* for descent algorithms remains an open subject of investigation [48,66]. In any case, as long as we are interested in typical saddle points (though not necessarily typical attractors of the dynamics), their loss value is connected to their generalization error.

Building on our results in this paper, we plan to obtain careful numerical results of the optimization problems discussed in the three model cases here. This task is not trivial, given the nonconvexity that appears in the solution spaces. As far as we can tell, the functional under extremization here seems to be more challenging than the functionals that usually appear for the characterizations of free energies and mutual information, for example, in [67,68], mainly because we cannot eliminate most order parameters and the dimension of the phase space remains too large to allow its exploration using brute-force analysis. In our limited numerical experiments, we could follow some phases by iterating the saddle points equations and tuning $\bar{\ell}$, but the full picture remains elusive. It should be stressed that this difficulty appears in exactly the same way in [6]. For example, the parameters f , κ , and ℓ appear there as well, and the difficulty in optimizing over them is similar. In contrast, the new geometric order parameters ρ , q that appear in our work can readily be analyzed in the simpler case of a quadratic activation function, where f , κ , and ℓ are proportional to each other.

In this work, we limited ourselves to the annealed critical point complexity, but the typical amount of saddle points would be more accurately captured by the quenched calculation, which is not sensitive to large deviations. Performing the quenched computation should not be out of reach of the methodology of this paper when supported by the replica method as has been done in [46,49].

Moving on, the complexity of the region of w , which has a Hessian with a finite and zero index, can be obtained by carefully analyzing the large deviation of the minimum eigenvalues of the Hessian function, as was done in [16,53]. This would be the first step in understanding how to count local minima of these models with specific properties, building toward an understanding of their domain of attraction [48].

ACKNOWLEDGMENTS

This work was partially supported by project MIS 5154714 of the National Recovery and Resilience Plan “Greece 2.0” funded by the European Union under the NextGenerationEU Program.

DATA AVAILABILITY

The data that support the findings of this article are openly available [65].

APPENDIX: ASYMPTOTIC VOLUME OF THE INTERSECTION OF A SPHERE AND AN ELLIPSE

Let ω_d be the volume of the $(d-1)$ -dimensional unit sphere $\mathbb{S}_d$. Given a set of $K = \mathcal{O}(1)$ positive definite matrices $\{A_k\}_{k=1}^K$ and positive numbers $\{\lambda_k\}_{k=1}^K$, we are interested in the volume of the shape defined by all $u \in \mathbb{S}_d$ such that all constraints $\{u^\top A_k u = \lambda_k\}_{k=1}^K$ are satisfied. We have the following proposition:

Proposition A.1. Given $K = \mathcal{O}(1)$ positive semidefinite matrices $\{A_k\}_{k=1}^K$ and positive numbers $\{\lambda_k\}_{k=1}^K$, let λ be the K -dimensional vector with elements $\{\lambda_k\}_{k=1}^K$ and let $V(\lambda)$ be the volume of the submanifold of the sphere defined by the constraints $\{u^\top A_k u = \lambda_k\}_{k=1}^K$. Consider also the set

$$\mathcal{L}_\lambda = \left\{ \bar{\lambda} \in \mathbb{R}^K \left| I + \sum_k \bar{\lambda}_k (A_k - \lambda_k I) > 0 \right. \right\}, \quad (\text{A1})$$

and the rate function $I(\cdot, \lambda) : \mathcal{L}_\lambda \rightarrow \mathbb{R}$, defined through

$$I_V(\bar{\lambda}; \lambda) = \lim_{d \rightarrow \infty} \frac{1}{2d} \log \det \left(I + \sum_k \bar{\lambda}_k (A_k - \lambda_k I) \right). \quad (\text{A2})$$

Then, the volume $V(\lambda)$ is asymptotically captured by

$$\lim_{d \rightarrow \infty} \frac{1}{d} \log \frac{1}{\omega_d} V(\lambda) = - \sup_{\bar{\lambda} \in \mathcal{L}_\lambda} I_V(\bar{\lambda}; \lambda). \quad (\text{A3})$$

Proof. Notice that, even if λ_k is not between the smallest and largest eigenvalues of A_k , the supremum of $I_V(\bar{\lambda}; \lambda)$ is taken at infinity and (A3) gives $-\infty$, as it should. Thus, we might assume that λ_k is between the smallest and largest eigenvalues of A_k . We express $V(\lambda)$ using delta functions and proceed with equations to first exponential order in d , denoted by “ $\doteq$ ”:

$$\begin{aligned} \frac{1}{\omega_d} V(\lambda) &= \frac{2}{\omega_d} \int_{\mathbb{S}_d} du \prod_{k=1}^K |A_k u| \delta(u^\top A_k u - \lambda_k) \\ &\doteq \frac{1}{\omega_d} \int_{\mathbb{S}_d} du \prod_{k=1}^K \delta(u^\top A_k u - \lambda_k) \\ &= \frac{1}{\pi^{d/2}} \int_{\mathbb{R}^d} du e^{-\frac{1}{2}|u|^2} \prod_{k=1}^K \delta(u^\top A_k u - \lambda_k u^\top u) \\ &= \frac{1}{\pi^{d/2}} \int_{\mathbb{R}^d} du e^{-\frac{1}{2}|u|^2} \frac{1}{(4\pi)^K} \\ &\quad \times \int_{\mathbb{R}^K} d\bar{\lambda} e^{i\frac{1}{2} \sum_k \bar{\lambda}_k (u^\top A_k u - \lambda_k u^\top u)} \\ &\doteq \int_{\mathbb{R}^K} e^{-\frac{1}{2} \log \det(I + \sum_k \bar{\lambda}_k (A_k - \lambda_k I))}. \end{aligned}$$

Restricted on the bounded convex set $\mathcal{L}_\lambda$, the function $\bar{\lambda} \rightarrow \log \det(I + \sum_k \bar{\lambda}_k (A_k - \lambda_k I))$ is concave and thus has at most a single local maximum. Since the values of this function go to $-\infty$ at the boundary, a local maximum must exist.

Thus, by the method of steepest descent, we get the desired result. ■

-
- [1] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, *Proc. IEEE* **86**, 2278 (1998).
 - [2] H. Xiao, K. Rasul, and R. Vollgraf, [arXiv:1708.07747](#).
 - [3] H. Li, Z. Xu, G. Taylor, C. Studer, and T. Goldstein, in *Advances in Neural Information Processing Systems*, Vol. 31 (Curran Associates, Inc., Red Hook, NY, 2018).
 - [4] A. Choromanska, M. Henaff, M. Mathieu, G. B. Arous, and Y. LeCun, in *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics* (PMLR, 2015), pp. 192–204.
 - [5] N. P. Baskerville, J. P. Keating, F. Mezzadri, and J. Najnudel, *J. Stat. Phys.* **186**, 29 (2022).
 - [6] A. Maillard, G. B. Arous, and G. Biroli, in *Mathematical and Scientific Machine Learning* (PMLR, 2020), pp. 287–327.
 - [7] Y. Arjevani and M. Field, in *Advances in Neural Information Processing Systems*, Vol. 33 (Curran Associates, Inc., Red Hook, NY, 2020), pp. 5441–5452.
 - [8] Y. Arjevani and M. Field, in *Advances in Neural Information Processing Systems*, Vol. 34 (Curran Associates, Inc., Red Hook, NY, 2021), pp. 15162–15174.
 - [9] Y. V. Fyodorov and R. Tublin, *J. Phys. A: Math. Theor.* **55**, 244008 (2022).
 - [10] W. S. McCulloch and W. Pitts, *Bull. Math. Biophys.* **5**, 115 (1943).
 - [11] B. Derrida, E. Gardner, and A. Zippelius, *Europhys. Lett.* **4**, 167 (1987).
 - [12] E. Gardner and B. Derrida, *J. Phys. A: Math. Gen.* **21**, 271 (1988).
 - [13] E. Gardner and B. Derrida, *J. Phys. A: Math. Gen.* **22**, 1983 (1989).
 - [14] S. Franz and G. Parisi, *J. Phys. A: Math. Theor.* **49**, 145001 (2016).
 - [15] J.-M. Azaïs and M. Wschebor, *Level Sets and Extrema of Random Processes and Fields* (John Wiley & Sons, New York, NY, 2009).
 - [16] A. Auffinger, Random matrices, complexity of spin glasses and heavy tailed processes, Ph.D. thesis, New York University, United States – New York, 2011.
 - [17] V. Ros and Y. V. Fyodorov, Spin Glass Theory and Far Beyond: Replica Symmetry Breaking after 40 Years 24 (2022).
 - [18] K. Asgari, A. Montanari, and B. Saeed, [arXiv:2502.01953](#).
 - [19] D. P. Kingma and M. Welling, [arXiv:1312.6114](#).
 - [20] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, in *Advances in Neural Information Processing Systems*, Vol. 27 (Curran Associates, Inc., Red Hook, NY, 2014).
 - [21] M. E. A. Seddik, C. Louart, M. Tamaazousti, and R. Couillet, in *International Conference on Machine Learning* (PMLR, 2020), pp. 8573–8582.

- [22] S. Goldt, B. Loureiro, G. Reeves, F. Krzakala, M. Mézard, and L. Zdeborová, in *Mathematical and Scientific Machine Learning* (PMLR, 2022), pp. 426–471.
- [23] L. Pesce, F. Krzakala, B. Loureiro, and L. Stephan, in *International Conference on Machine Learning* (PMLR, 2023), pp. 27680–27708.
- [24] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli, in *International Conference on Machine Learning* (PMLR, 2015), pp. 2256–2265.
- [25] Y. Song and S. Ermon, in *Advances in Neural Information Processing Systems*, Vol. 32 (Curran Associates, Inc., Red Hook, NY, 2019).
- [26] J. Ho, A. Jain, and P. Abbeel, in *Advances in Neural Information Processing Systems*, Vol. 33 (Curran Associates, Inc., Red Hook, NY, 2020), p. 6840.
- [27] A. Cauchy, in *Comptes rendus de l'Académie des Sciences*, Vol. 25 (French Academy of Sciences, 1847), p. 536.
- [28] H. Robbins and D. Siegmund, in *Optimizing Methods in Statistics* (Elsevier, Amsterdam, 1971), pp. 233–257.
- [29] M. Benaïm, in *Seminaire de Probabilités XXXIII* (Springer, 2006), pp. 1–68.
- [30] I. Panageas and G. Piliouras, [arXiv:1605.00405](https://arxiv.org/abs/1605.00405).
- [31] I. Panageas, G. Piliouras, and X. Wang, in *Advances in Neural Information Processing Systems*, Vol. 32 (Curran Associates, Inc., Red Hook, NY, 2019).
- [32] K. Antonakopoulos, P. Mertikopoulos, G. Piliouras, and X. Wang, in *International Conference on Machine Learning* (PMLR, 2022), pp. 731–771.
- [33] D. Saad and S. A. Solla, *Phys. Rev. Lett.* **74**, 4337 (1995).
- [34] R. Veiga, L. Stephan, B. Loureiro, F. Krzakala, and L. Zdeborová, in *Advances in Neural Information Processing Systems*, Vol. 35 (Curran Associates, Inc., Red Hook, NY, 2022), p. 23244.
- [35] S. S. Mannelli, F. Krzakala, P. Urbani, and L. Zdeborová, in *International Conference on Machine Learning* (PMLR, 2019), pp. 4333–4342.
- [36] F. Mignacco, F. Krzakala, P. Urbani, and L. Zdeborová, *J. Stat. Mech.* (2021) 124008.
- [37] Y. V. Fyodorov, *Phys. Rev. Lett.* **92**, 240601 (2004).
- [38] Y. V. Fyodorov, *Acta Phys. Polonica B* **36**, 2699 (2005).
- [39] Y. V. Fyodorov and I. Williams, *J. Stat. Phys.* **129**, 1081 (2007).
- [40] Y. V. Fyodorov and J. P. Keating, *Philos. Trans. R. Soc. A* **372**, 20120503 (2014).
- [41] G. Ben Arous, P. Bourgade, and B. McKenna, *Probab. Math. Phys.* **3**, 731 (2023).
- [42] A. Auffinger and W.-K. Chen, *J. Stat. Phys.* **157**, 40 (2014).
- [43] Z. Fan, S. Mei, and A. Montanari, *Ann. Probab.* **49**, 1 (2021).
- [44] B. McKenna, in *Annales de l'Institut Henri Poincaré (B) Probabilités et statistiques*, Vol. 60 (1) (Institut Henri Poincaré, 2024), pp. 636–657.
- [45] V. Ros, G. Ben Arous, G. Biroli, and C. Cammarota, *Phys. Rev. X* **9**, 011003 (2019).
- [46] V. Ros, G. Biroli, and C. Cammarota, *Europhys. Lett.* **126**, 20003 (2019).
- [47] J. Kent-Dobias, *Europhys. Lett.* **143**, 61003 (2023).
- [48] J. Kent-Dobias, *Phys. Rev. E* **110**, 064148 (2024).
- [49] J. Kent-Dobias and J. Kurchan, *Phys. Rev. E* **107**, 064111 (2023).
- [50] L. Zdeborová and F. Krzakala, *Adv. Phys.* **65**, 453 (2016).
- [51] J. Barbier, [arXiv:2010.14863](https://arxiv.org/abs/2010.14863).
- [52] D. Gamarnik, C. Moore, and L. Zdeborová, *J. Stat. Mech.* (2022) 114015.
- [53] A. Choromanska, Y. LeCun, and G. B. Arous, in *Proceedings of the 28th Conference on Learning Theory* (PMLR, 2015), pp. 1756–1760.
- [54] G. S. Hartnett, E. Parker, and E. Geist, *Phys. Rev. E* **98**, 022116 (2018).
- [55] A. L. Moustakas, S. H. Simon, and A. M. Sengupta, *IEEE Trans. Inform. Theory* **49**, 2545 (2003).
- [56] R. Couillet, M. Debbah, and J. W. Silverstein, *IEEE Trans. Inform. Theory* **57**, 3493 (2011).
- [57] M. Mézard and A. Montanari, *Information, Physics, and Computation* (Oxford University Press, Oxford, 2009).
- [58] W. Hachem, O. Khorunzhiy, P. Loubaton, J. Najim, and L. Pastur, *IEEE Trans. Inform. Theory* **54**, 3987 (2008).
- [59] A. L. Moustakas and P. Kazakopoulos, *IEEE Trans. Inform. Theory* **59**, 6490 (2013).
- [60] T. Tao, *Topics in Random Matrix Theory*, Vol. 132 (American Mathematical Soc., Providence, 2012).
- [61] D. Paul and J. W. Silverstein, *J. Multivariate Anal.* **100**, 37 (2009).
- [62] B. Bercu, F. Gamboa, and M. Lavielle, *ESAIM Probab. Stat.* **4**, 1 (2000).
- [63] A. Dembo and O. Zeitouni, *Large Deviations Techniques and Applications*, Vol. 38 (Springer Science & Business Media, New York, 2009).
- [64] A. L. Moustakas and S. H. Simon, *IEEE Trans. Inform. Theory* **53**, 3887 (2007).
- [65] Data available at http://scholar.uoa.gr/sites/default/files/arislm/files/tm25_figdata.txt.
- [66] G. Folena and F. Zamponi, *SciPost Phys.* **15**, 109 (2023).
- [67] Y. Kabashima, *J. Phys.: Conf. Ser.* **95**, 012001 (2008).
- [68] A. Maillard, B. Loureiro, F. Krzakala, and L. Zdeborová, in *Advances in Neural Information Processing Systems*, Vol. 33 (Curran Associates, Inc., Red Hook, NY, 2020), pp. 11071–11082.