Baskerville, N. P., Keating, J. P., Mezzadri, F., & Najnudel, J. (2021). A spin-glass model for the loss surfaces of generative adversarial networks. *arXiv*. https://arxiv.org/abs/2101.02524

A SPIN GLASS MODEL FOR THE LOSS SURFACES OF GENERATIVE ADVERSARIAL NETWORKS

Nicholas P. Baskerville[1], Jonathan P. Keating[2], Francesco Mezzadri[1], and Joseph Najnudel[1]

[1]*School of Mathematics, University of Bristol, Fry Building, Bristol, BS8 1UG, UK*
[2]*Mathematical Institute, University of Oxford, Oxford, OX2 6GG, UK*
{n.p.baskerville, F.Mezzadri, joseph.najnudel}@bristol.ac.uk, Jon.Keating@maths.ox.ac.uk

January 8, 2021

ABSTRACT

We present a novel mathematical model that seeks to capture the key design feature of generative adversarial networks (GANs). Our model consists of two interacting spin glasses, and we conduct an extensive theoretical analysis of the complexity of the model's critical points using techniques from Random Matrix Theory. The result is insights into the loss surfaces of large GANs that build upon prior insights for simpler networks, but also reveal new structure unique to this setting.

1 Introduction

By making various modeling assumptions about standard multi-layer perceptron neural networks, [Cho+15] argued heuristically that the training loss surfaces of large networks could be modelled by a spherical multi-spin glass. Using theoretical results of [Auf13], they were able to arrive at quantitative asymptotic characterisations, in particular the existence of a favourable 'banded structure' of local-optima of the loss. There are clear and acknowledged deficiencies with their assumptions [CLA15] and recent observations have shown that the Hessians of real-world deep neural networks do not behave like random matrices from the Gaussian Orthogonal Ensemble (GOE) of Random Matrix Theory at the macroscopic scale [Pap18; Gra+19; Gra20], despite this being implied by the spin-glass model of [Cho+15]. Moreover, there have been questions raised about whether the mean asymptotic properties of loss surfaces for deep neural networks (or energy surfaces of glassy objects) are even relevant practically for gradient-based optimisation in sub-exponential time [Bai+19; Man+19a; FFR19], though interpretation of experiments with deep neural networks remains difficult and the discussion about the true shape of their loss surfaces and the implications thereof is far from settled. Nevertheless, spin-glass models present a tractable example of high-dimensional complex random functions that may well provide insights into aspects of deep learning. Rather than trying to improve or reduce the assumptions of [Cho+15], various authors have recently opted to skip the direct derivation from a neural network to a statistical physics model, instead proposing simple models designed to capture aspects of training dynamics and studying those directly. Examples include: the modified spin glass model of [Ros+19a] with some explicitly added 'signal'; the simple explicitly non-linear model of [MAB19]; the spiked tensor 'signal-in-noise' model of [Man+19b]. In a slightly different direction, [Bas+20] removed one of the main assumptions from the [Cho+15] derivation, and in so doing arrived at a deformed spin-glass model. All of this recent activity sits in the context of earlier work connecting spin-glass objects with simple neural networks [KS87; Gar88; EV01] and, more generally, with image reconstruction and other signal processing problems [Nis01].

One area that has not been much explored in the line of the above-mentioned literature is the study of architectural variants. Modern deep learning contains a very large variety of different design choices in network architecture, such as convolutional networks for image and text data (among others) [Goo+16; Con+17], recurrent networks for sequence data [HS97] and self-attention transformer networks for natural language [Dev+19; Rad+18]. Given the ubiquity of convolutional networks, one might seek to study those, presumably requiring consideration of local correlations in data. One could imagine some study of architectural quirks such as residual connections [He+16], and batch-norm has been considered to some extent by [PW17]. In this work, we propose a novel model for *generative adversarial* *networks* (GANs) [Goo+14] as two interacting spherical spin glasses. GANs have been the focus of intense research and development in recent years, with a large number of variants being proposed [RMC15; Zha+18; LT16; KLA20; MO14; ACB17; Zhu+17] and rapid progress particularly in the field of image generation. From the perspective of optimisation, GANs have much in common with other deep neural networks, being complicated high-dimensional functions optimised using local gradient-based methods such as stochastic gradient descent and variants. On the other hand, the adversarial training objective of GANs, with two deep networks competing, is clearly an important distinguishing feature, and GANs are known to be more challenging to train than single deep networks. Our objective is to capture the essential adversarial aspect of GANs in a tractable model of high-dimensional random complexity which, though being a significant simplification, has established connections to neural networks and high dimensional statistics.

Our model is inspired by [Cho+15; Ros+19b; Man+19b; Aro+19] with spherical multi-spin glasses being used in place of deep neural networks. We thus provide a complicated, random, high-dimensional model with the essential feature of GANs clearly reflected in its construction. By employing standard Kac-Rice complexity calculations [Fyo04; FW07; Auf13] we are able to reduce the loss landscape complexity calculation to a random matrix theoretic calculation. We then employ various Random Matrix Theory techniques as in [Bas+20] to obtain rigorous, explicit leading order asymptotic results. Our calculations rely on the supersymmetric method in Random Matrix Theory, in particular the approach to calculating limiting spectral densities follows [Ver04] and the calculation also follows [GW90; Guh91] in important ways. The greater complexity of the random matrix spectra encountered present some challenges over previous such calculations, which we overcome with a combination of analytical and numerical approaches. Using our complexity results, we are able to draw qualitative implications about GAN loss surfaces analogous to those of [Cho+15] and also investigate the effect of a few key design parameters included in the GAN. We compare the effect of these parameters on our spin glass model and also on the results of experiments training real GANs. Our calculations include some novel details, in particular, we use precise sub-leading terms for a limiting spectral density obtained from supersymmetric methods to prove a required concentration result.

The role that statistical physics models such as spherical multi-spin glasses are to ultimately play in the theory of deep learning is not yet clear, with arguments both for and against their usefulness and applicability. We provide a first attempt to model an important architectural feature of modern deep neural networks within the framework of spin glass models and provide a detailed analysis of properties of the resulting loss (energy) surface. Our analysis reveals potential explanations for observed properties of GANs and demonstrates that it may be possible to inform practical hyperparameter choices using models such as ours. Much of the advancement in practical deep learning has come from innovation in network architecture, so if deep learning theory based on simplified physics models like spin-glasses is to keep pace with practical advances in the field, then it will be necessary to account for architectural details within such models. Our work is a first step in that direction and the mathematical techniques used may prove more widely valuable.

The paper is structured as follows: in Section 2 we introduce the interacting spin glass model; in Section 3 we use a Kac-Rice formula to derive random matrix expressions for the asymptotic complexity of our model; in Section 4 we derive the limiting spectral density of the relevant random matrix ensemble; in Section 5 we use the Coulomb gas approximation to compute the asymptotic complexity, and legitimise its use by proving a concentration result; in Section 6 we derive some implications of our model for GAN training and compare to experimental results from real GANs; in Section 7 we conclude. All code used for numerical calculations of our model, training real GANs, analysing the results and generating plots is made available[1].

2 An interacting spin glass model

We use multi-spin glasses in high dimensions as a toy model for neural network loss surfaces without any further justification, beyond that found in [Cho+15; Bas+20]. GANs are composed of two networks: *generator* (G) and *discriminator* (D). G is a map $\mathbb{R}^m \to \mathbb{R}^d$ and D is a map $\mathbb{R}^d \to \mathbb{R}$. G's purpose is to generate synthetic data samples by transforming random input noise, while D's is to distinguish between real data samples and those generated by G. Given some probability distribution $\mathbb{P}_{data}$ on some $\mathbb{R}^d$, GANs have the following minimax training objective

$$\min_{\Theta_G} \max_{\Theta_D} \left\{ \mathbb{E}_{\boldsymbol{x}\sim\mathbb{P}_{data}} \log D(\boldsymbol{x}) + \mathbb{E}_{\boldsymbol{z}\sim\mathcal{N}(0,\sigma_z^2)} \log(1 - D(G(\boldsymbol{z}))) \right\}, \tag{1}$$

where Θ_D, Θ_G are the parameters of the discriminator and generator respectively. With $\boldsymbol{z} \sim \mathcal{N}(0, \sigma_z^2)$, $G(\boldsymbol{z})$ has some probability distribution $\mathbb{P}_{gen}$. When successfully trained, the initially unstructured $\mathbb{P}_{gen}$ examples are easily distinguished by D, this in turn drives improvements in G, bring $\mathbb{P}_{gen}$ closer to $\mathbb{P}_{data}$. Ultimately, the process successfully terminates when $\mathbb{P}_{gen}$ is very close to $\mathbb{P}_{data}$ and D performs little better than random at the distinguishing

[1] `https://github.com/npbaskerville/loss-surfaces-of-gans`

task. To construct our model, we introduce two spin glasses:

$$\ell^{(D)}(\boldsymbol{w}^{(D)}) = \sum_{i_1,\ldots,i_p=1}^{N_D} X_{i_1,\ldots,i_p} \prod_{k=1}^{p} w_{i_k}^{(D)} \tag{2}$$

$$\ell^{(G)}(\boldsymbol{w}^{(D)}, \boldsymbol{w}^{(G)}) = \sum_{i_1,\ldots,i_p=1}^{N_D} \sum_{j_1,\ldots,j_q=1}^{N_G} Z_{i_1,\ldots,i_p,j_1,\ldots,j_q} \prod_{k=1}^{p} w_{i_k}^{(D)} \prod_{l=1}^{q} w_{i_l}^{(G)} \tag{3}$$

$$\tag{4}$$

where all the $X_{i_1,\ldots,i_p}$ are i.i.d. $\mathcal{N}(0,1)$ and $Z_{j_1,\ldots,j_q}$ are similarly i.i.d. $\mathcal{N}(0,1)$. We then define the generator and discriminator spin glasses:

$$L^{(D)}(\boldsymbol{w}^{(D)}, \boldsymbol{w}^{(G)}) = \ell^{(D)}(\boldsymbol{w}^{(D)}) - \sigma_z \ell^{(G)}(\boldsymbol{w}^{(D)}, \boldsymbol{w}^{(G)}), \tag{5}$$

$$L^{(G)}(\boldsymbol{w}^{(D)}, \boldsymbol{w}^{(G)}) = \sigma_z \ell^{(G)}(\boldsymbol{w}^{(D)}, \boldsymbol{w}^{(G)}). \tag{6}$$

$\ell^{(D)}$ plays the role of the loss of the discriminator network when trying to classify genuine examples as such. $\ell^{(G)}$ plays the role of loss of the discriminator when applied to samples produced by the generator, hence the sign difference between $L^{(D)}$ and $L^{(G)}$. $\boldsymbol{w}^{(D)}$ are the weights of the discriminator, and $\boldsymbol{w}^{(G)}$ the weights of the generator. The $X_{\boldsymbol{i}}$ are surrogates for the training data (i.e. samples from $\mathbb{P}_{data}$) and the $Z_{\boldsymbol{j}}$ are surrogates for the noise distribution of the generator. For convenience, we have chosen to pull the σ_z scale outside of the $Z_{\boldsymbol{j}}$ and include it as a constant multiplier in (5)-(6). In reality, we should like to keep $Z_{\boldsymbol{j}}$ as i.i.d. $\mathcal{N}(0,1)$ but take $X_{\boldsymbol{i}}$ to have some other more interesting distribution, e.g. normally or uniformly distributed on some manifold. Using $[x]$ to denote the integer part of x, we take $N_D = [\kappa N], N_G = [\kappa' N]$ for fixed $\kappa \in (0,1)$, $\kappa' = 1 - \kappa$, and study the regime $N \to \infty$. Note that there is no need to distinguish between $[\kappa N]$ and κN in the $N \to \infty$ limit.

3 Kac-Rice formulae for complexity

Training GANs involves jointly minimising the losses of the discriminator and the generator. Therefore, rather than being interested simply in upper-bounding a single spin-glass and counting its stationary points, the complexity of interest comes from jointly upper bounding both $L^{(D)}$ and $L^{(G)}$ and counting points where both are stationary. Using S^M to denote the M-sphere[2], we define the complexity

$$C_N = \left| \left\{ \boldsymbol{w}^{(D)} \in S^{N_D}, \boldsymbol{w}^{(G)} \in S^{N_G} \ : \ \nabla_D L^{(D)} = 0, \nabla_G L^{(G)} = 0, L^{(D)} \in B_D, L^{(G)} \in B_G \right\} \right| \tag{7}$$

for some Borel sets $B_D, B_G \subset \mathbb{R}$ and where ∇_D, ∇_G denote the conformal derivatives with respect to the discriminator and generator weights respectively. Note:

1. We have chosen to treat the parameters of each network as somewhat separate by placing them on their own hyper-spheres. This reflects the minimax nature of GAN training, where there really are 2 networks being optimised in an adversarial manner rather than one network with some peculiar structure.
2. We could have taken $\nabla = (\nabla_D, \nabla_G)$ and required $\nabla L^{(D)} = \nabla L^{(G)} = 0$ but, as in the previous comment, our choice is more in keeping with the adversarial set-up, with each network seeking to optimize separately its own parameters in spite of the other.
3. We will only be interested in the case $B_D = (-\infty, \sqrt{N} u_D)$ and $B_G = (-\infty, \sqrt{N} u_G)$, for $u_D, u_G \in \mathbb{R}$.

So that the finer structure of local minima and saddle points can be probed, we also define the corresponding complexity with Hessian index prescription

$$C_{N,k_D,k_G} = \Bigg| \Bigg\{ \boldsymbol{w}^{(D)} \in S^{N_D}, \boldsymbol{w}^{(G)} \in S^{N_G} \ : \ \nabla_D L^{(D)} = 0, \nabla_G L^{(G)} = 0, L^{(D)} \in B_D, L^{(G)} \in B_G \\ i(\nabla_D^2 L^{(D)}) = k_D,\ i(\nabla_G^2 L^{(G)}) = k_G \Bigg\} \Bigg|, \tag{8}$$

[2] We use the convention of the M-sphere being the sphere embedded in $\mathbb{R}^M$.

where $i(M)$ is the index of M (i.e. the number of negative eigenvalues of M). To calculate the complexities, we follow the well-trodden route of Kac-Rice formulae as pioneered by [Fyo04; FW07]. For a fully rigorous treatment, we proceed as in [Auf13; Bas+20] by turning to the following result from [AT09].

Theorem 3.1 ([AT09] Theorem 12.1.1)**.** *Let $\mathcal{M}$ be a compact , oriented, N-dimensional C^1 manifold with a C^1 Riemannian metric g. Let $\phi : \mathcal{M} \to \mathbb{R}^N$ and $\psi : \mathcal{M} \to \mathbb{R}^K$ be random fields on $\mathcal{M}$. For an open set $A \subset \mathbb{R}^K$ for which ∂A has dimension $K-1$ and a point $\boldsymbol{u} \in \mathbb{R}^N$ let*

$$N_{\boldsymbol{u}} \equiv |\{x \in \mathcal{M} \mid \phi(x) = \boldsymbol{u},\ \psi(x) \in A\}|. \tag{9}$$

Assume that the following conditions are satisfied for some orthonormal frame field E:

(a) All components of ϕ, $\nabla_E\phi$, and ψ are a.s. continuous and have finite variances (over $\mathcal{M}$).

(b) For all $x \in \mathcal{M}$, the marginal densities p_x of $\phi(x)$ (implicitly assumed to exist) are continuous at $\mathbf{u}$.

(c) The conditional densities $p_x(\cdot|\nabla_E\phi(x), \psi(x))$ of $\phi(x)$ given $\psi(x)$ and $\nabla_E\phi(x)$ (implicitly assumed to exist) are bounded above and continuous at $\mathbf{u}$, uniformly in $\mathcal{M}$.

(d) The conditional densities $p_x(\cdot|\phi(x) = \boldsymbol{z})$ of $\det(\nabla_{E_j}\phi^i(x))$ given are continuous in a neighbourhood of 0 for $\boldsymbol{z}$ in a neighbourhood of $\mathbf{u}$ uniformly in $\mathcal{M}$.

(e) The conditional densities $p_x(\cdot|\phi(x) = \boldsymbol{z})$ are continuous for $\boldsymbol{z}$ in a neighbourhood of $\mathbf{u}$ uniformly in $\mathcal{M}$.

(f) The following moment condition holds

$$\sup_{x\in\mathcal{M}} \max_{1\leq i,j\leq N} \mathbb{E}\left\{\left|\nabla_{E_j}\phi^i(x)\right|^N\right\} < \infty \tag{10}$$

(g) The moduli of continuity with respect to the (canonical) metric induced by g of each component of ψ, each component of ϕ and each $\nabla_{E_j}\phi^i$ all satisfy, for any $\epsilon > 0$

$$\mathbb{P}(\omega(\eta) > \epsilon) = o(\eta^N), \ \ \textit{as } \eta \downarrow 0 \tag{11}$$

where the modulus of continuity *of a real-valued function G on a metric space (T, τ) is defined as (c.f. [AT09] around (1.3.6))*

$$\omega(\eta) := \sup_{s,t:\tau(s,t)\leq\eta} |G(s) - G(t)| \tag{12}$$

Then

$$\mathbb{E}N_{\boldsymbol{u}} = \int_{\mathcal{M}} \mathbb{E}\left\{|\det\nabla_E\phi(x)|\mathbb{1}\{\psi(x)\in A\} \mid \phi(x) = \boldsymbol{u}\right\} p_x(\boldsymbol{u}) Vol_g(x) \tag{13}$$

where p_x is the density of ϕ and Vol_g is the volume element induced by g on $\mathcal{M}$.

In the notation of Theorem 3.1, we make the following choices:

$$\phi = \begin{pmatrix} \nabla_D L^{(D)} \\ \nabla_G L^{(G)} \end{pmatrix}, \quad \psi = \begin{pmatrix} L^{(D)} \\ L^{(G)} \end{pmatrix}$$

and so

$$A = B_D \times B_G, \quad \boldsymbol{u} = 0.$$

and the manifold $\mathcal{M}$ is taken to be $S^{N_D} \times S^{N_G}$ with the product topology.

Lemma 3.2.

$$C_N = \int_{S^{N_D}\times S^{N_G}} d\boldsymbol{w}^{(G)} d\boldsymbol{w}^{(D)} \ \ \varphi_{(\nabla_D L^{(D)}, \nabla_G L^{(G)})}(0)\mathbb{E}\left[|\det\begin{pmatrix} \nabla_D^2 L^{(D)} & \nabla_{GD}L^{(D)} \\ \nabla_{DG}L^{(G)} & \nabla_G^2 L^{(G)}\end{pmatrix}| \middle| \nabla_G L^{(G)} = 0, \nabla_D L^{(D)} = 0\right]$$
$$\mathbb{1}\left\{L^{(D)} \in B_D, L^{(G)} \in B_G\right\} \tag{14}$$

and therefore

$$C_N = \int_{S^{N_D}\times S^{N_G}} d\boldsymbol{w}^{(G)} d\boldsymbol{w}^{(D)} \;\varphi_{(\nabla_D L^{(D)}, \nabla_G L^{(G)})}(0) \int_{B_D} dx_D \int_{B_G} dx_G \;\varphi_{L^{(D)}}(x_D)\varphi_{L^{(G)}}(x_G)$$
$$\mathbb{E}\left[|\det \left(\begin{array}{cc} \nabla_D^2 L^{(D)} & \nabla_{GD} L^{(D)} \\ \nabla_{DG} L^{(G)} & \nabla_G^2 L^{(G)} \end{array} \right) | \;\middle|\; \nabla_G L^{(G)} = 0, \nabla_D L^{(D)} = 0, L^{(D)} = x_D, L^{(G)} = x_G \right]. \tag{15}$$

where $\varphi_{(\nabla_D L^{(D)}, \nabla_G L^{(G)})}$ is the joint density of $(\nabla_D L^{(D)}, \nabla_G L^{(G)})^T$, $\varphi_{L^{(D)}}$ the density of $L^{(D)}$, and $\varphi_{L^{(G)}}$ the density of $L^{(G)}$, all implicitly evaluated at $(\boldsymbol{w}^{(G)}, \boldsymbol{w}^{(D)})$.

Proof. It is sufficient to check the conditions of Theorem 3.1 with the above choices.

Conditions (a)-(f) are satisfied due to Gaussianity and the manifestly smooth definition of $L^{(D)}, L^{(G)}$. The moduli of continuity conditions as in (g) are satisfied separately for $L^{(D)}$ and its derivatives on S^{N_D} and for $L^{(G)}$ and its derivatives on S^{N_G}, as seen in the proof of the analogous result for a single spin glass in [Auf13]. But since $\mathcal{M}$ is just a direct product with product topology, it immediately follows that (g) is satisfied, so Theorem 13 applies and we obtain (14). (15) follows simply, using the rules of conditional expectation. □

Define the Hessian matrix

$$\tilde{H} = \left(\begin{array}{cc} \nabla_D^2 L^{(D)} & \nabla_{GD} L^{(D)} \\ \nabla_{DG} L^{(G)} & \nabla_G^2 L^{(G)} \end{array} \right).$$

To make use of (15), we need the joint distribution of $\left(\ell^{(D)}, \partial_i^{(D)}\ell^{(D)}, \partial_{jk}^{(D)}\ell^{(D)}\right)$ and the independent $\left(\ell^{(G)}, \partial_i^{(G)}\ell^{(G)}, \partial_{jk}^{(G)}\ell^{(G)}, \partial_l^{(D)}\ell^{(G)}, \partial_{mn}^{(D)}\ell^{(G)}\right)$. As in [Auf13], we will simplify the calculation by evaluating in the region of the north poles on each hyper-sphere. $\ell^{(D)}$ behaves just like a single spin glass, and so we have [Auf13]:

$$Var(\ell^{(D)}) = 1, \tag{16}$$
$$Cov(\partial_i^{(D)}\ell^{(D)}, \partial_{jk}^{(D)}\ell^{(D)}) = 0, \tag{17}$$
$$\partial_{ij}^{(D)}\ell^{(D)} \mid \{\ell^{(D)} = x_D\} \sim \sqrt{(N_D-1)p(p-1)}GOE^{N_D-1} - x_D pI. \tag{18}$$

To find the joint and thence conditional distributions for $\ell^{(G)}$, we first note that

$$Cov(\ell^{(G)}(\boldsymbol{w}^{(D)}, \boldsymbol{w}^{(G)}), \ell^{(G)}(\boldsymbol{w}^{(D)'}, \boldsymbol{w}^{(G)'})) = \left(\boldsymbol{w}^{(D)}\cdot\boldsymbol{w}^{(D)'} + \boldsymbol{w}^{(G)}\cdot\boldsymbol{w}^{(G)'}\right)^{p+q} \tag{19}$$

from which, by comparing with [Auf13], one can write down the necessary expressions, at the north poles in a coordinate basis:

$$Var(\ell^{(G)}) = 2^{p+q}, \tag{20}$$
$$Cov(\partial_{ij}^{(G)}\ell^{(G)}, \ell^{(G)}) = -(p+q)2^{p+q}\delta_{ij}, \tag{21}$$
$$Cov(\partial_{ij}^{(D)}\ell^{(G)}, \ell^{(G)}) = -(p+q)2^{p+q}\delta_{ij}, \tag{22}$$
$$Cov(\partial_{ij}^{(G)}\ell^{(G)}, \partial_{kl}^{(G)}\ell^{(G)}) = 2^{p+q}\left[(p+q)(p+q-1)\left(\delta_{ik}\delta_{jl} + \delta_{il}\delta_{jk}\right) + (p+q)^2\delta_{ij}\delta_{kl}\right], \tag{23}$$
$$Cov(\partial_{ij}^{(G)}\ell^{(G)}, \partial_{kl}^{(D)}\ell^{(G)}) = 2^{p+q}(p+q)^2\delta_{ij}\delta_{kl}, \tag{24}$$
$$Cov(\partial_i^{(G)}\partial_j^{(D)}\ell^{(G)}, \partial_k^{(G)}\partial_l^{(D)}\ell^{(G)}) = 2^{p+q}(p+q)(p+q-1)\delta_{ik}\delta_{jl}, \tag{25}$$
$$Cov(\partial_{ij}^{(G)}\ell^{(G)}, \partial_k^{(G)}\partial_l^{(D)}\ell^{(G)}) = 0 \tag{26}$$
$$Cov(\partial_{ij}^{(D)}\ell^{(G)}, \partial_k^{(D)}\partial_l^{(G)}\ell^{(G)}) = 0. \tag{27}$$

Also, all first derivatives of $\ell^{(G)}$ are clearly independent of $\ell^{(G)}$ and its second derivatives by the same reasoning as in [Auf13]. Note that

$$Cov(\partial_i^{(D)}L^{(D)}, \partial_j^{(D)}L^{(D)}) = (p + \sigma_z^2 2^{p+q}(p+q))\delta_{ij} \tag{28}$$
$$Cov(\partial_i^{(G)}L^{(G)}, \partial_j^{(G)}L^{(G)}) = \sigma_z^2 2^{p+q}(p+q)\delta_{ij} \tag{29}$$
$$Cov(\partial_i^{(D)}L^{(D)}, \partial_j^{(G)}L^{(G)}) = 0 \tag{30}$$

and so

$$\varphi_{\left(\nabla_D L^{(D)}, \nabla_G L^{(G)}\right)}(0) = (2\pi)^{-\frac{N-2}{2}} \left(p + \sigma_z^2 2^{p+1}(p+q)\right)^{-\frac{N_D-1}{2}} \left(\sigma_z^2 2^{p+q}(p+q)\right)^{-\frac{N_G-1}{2}}. \tag{31}$$

We need now to calculate the joint distribution of $(\partial_{ij}^{(D)}\ell^{(G)}, \partial_{kl}^{(G)}\ell^{(G)})$ conditional on $\{\ell^{(G)} = x_G\}$. Denote the covariance matrix for $(\partial_{ij}^{(D)}\ell^{(G)}, \partial_{kl}^{(G)}\ell^{(G)}, \ell^{(G)})$ by

$$\Sigma = \left(\begin{array}{cc} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{array}\right) \tag{32}$$

where

$$\begin{aligned}
\Sigma_{11} &= 2^{p+q}\left(\begin{array}{cc} (p+1)(p+q-1)(1+\delta_{ij}) + (p+q)^2\delta_{ij} & (p+q)^2\delta_{ij}\delta_{kl} \\ (p+q)^2\delta_{ij}\delta_{kl} & (p+1)(p+q-1)(1+\delta_{kl}) + (p+q)^2\delta_{kl} \end{array}\right), & (33)\\
\Sigma_{12} &= -2^{p+q}(p+q)\left(\begin{array}{c} \delta_{ij} \\ \delta_{kl} \end{array}\right), & (34)\\
\Sigma_{21} &= -2^{p+q}(p+q)\left(\begin{array}{cc} \delta_{ij} & \delta_{kl} \end{array}\right), & (35)\\
\Sigma_{22} &= 2^{p+q}. & (36)
\end{aligned}$$

The conditional covariance is then

$$\bar{\Sigma} = 2^{p+q}(p+1)(p+q-1)\left(\begin{array}{cc} 1+\delta_{ij} & 0 \\ 0 & 1+\delta_{kl} \end{array}\right). \tag{37}$$

In summary, from (37) and (25-27) we obtain

$$\begin{aligned}
\left(\begin{array}{cc} -\nabla_D^2\ell^{(G)} & -\nabla_G\nabla_D\ell^{(G)} \\ \nabla_D\nabla_G\ell^{(G)} & \nabla^2\ell^{(G)} \end{array}\right) \mid \{\ell^{(G)} = x_G\} &\overset{d}{=} \sqrt{2^{p+q+1}(p+q)(p+q-1)}\left(\begin{array}{cc} \sqrt{N_D-1}M_1^{(D)} & -2^{-1/2}G \\ 2^{-1/2}G^T & \sqrt{N_G-1}M^{(G)} \end{array}\right) \\
&\quad - (p+q)x_G 2^{p+1}\left(\begin{array}{cc} -I_{N_D} & 0 \\ 0 & I_{N_G} \end{array}\right)
\end{aligned} \tag{38}$$

where $M_1^{(D)} \sim GOE^{N_D-1}$ and $M^{(G)} \sim GOE^{N_G-1}$ are independent GOEs and G is an independent $N_D - 1 \times N_G - 1$ Ginibre matrix with entries of unit variance. Therefore, conditional on $\{(\ell^{(D)}, \ell^{(G)}) = (x_D, x_G)\}$,

$$\begin{aligned}
\tilde{H} \overset{d}{=} \sqrt{2p(p-1)}\left(\begin{array}{cc} \sqrt{N_D-1}M_2^{(D)} & 0 \\ 0 & 0 \end{array}\right) &+ \sigma_z\sqrt{2^{p+q+1}(p+q)(p+q-1)}\left(\begin{array}{cc} \sqrt{N_D-1}M_1^{(D)} & -2^{-1/2}G \\ 2^{-1/2}G^T & \sqrt{N_G-1}M^{(G)} \end{array}\right) \\
&- \sigma_z(p+q)x_G 2^{p+q}\left(\begin{array}{cc} -I_{N_D} & 0 \\ 0 & I_{N_G} \end{array}\right) - px_D\left(\begin{array}{cc} I_{N_D} & 0 \\ 0 & 0 \end{array}\right)
\end{aligned} \tag{39}$$

where $M_2^{(D)}$ is another independent GOE^{N_D-1} matrix. We can simplify to obtain:

$$\tilde{H} = \left(\begin{array}{cc} \sigma_D\sqrt{N_D-1}M^{(D)} & -2^{-1/2}\sigma_G G, \\ 2^{-1/2}\sigma_G G^T & \sigma_G\sqrt{N_G-1}M^{(G)} \end{array}\right) - \sigma_z(p+q)x_G 2^{p+q}\left(\begin{array}{cc} -I_{N_D} & 0 \\ 0 & I_{N_G} \end{array}\right) - px_D\left(\begin{array}{cc} I_{N_D} & 0 \\ 0 & 0 \end{array}\right) \tag{40}$$

where

$$\begin{aligned}
\sigma_G &= \sigma_z\sqrt{2^{p+q+1}(p+q)(p+q-1)} & (41)\\
\sigma_D &= \sqrt{\sigma_G^2 + 2p(p-1)} & (42)
\end{aligned}$$

and $M^{(D)} \sim GOE^{N_D-1}$ is a GOE matrix independent of $M^{(G)}$ and G.

Alternatively, because $M_{1,2}^{(D)} \overset{d}{=} -M_{1,2}^{(D)}$, let us write $\tilde{H}$ as

$$\begin{aligned}\tilde{H} =&\sigma_z J \left(\sqrt{2^{p+q+1}(p+q)(p+q-1)(N_D+N_G-2)}M_1 - (p+q)x_G 2^{p+q} I\right)\\ &+ \left(\sqrt{2p(p-1)(N_D-1)}\begin{pmatrix} M_2 & 0 \\ 0 & 0\end{pmatrix} - p x_D \begin{pmatrix} I_{N_D} & 0 \\ 0 & 0\end{pmatrix}\right)\\ \overset{d}{=}&J\Bigg[\sigma_z\sqrt{2^{p+q+1}(p+q)(p+q-1)(N_D+N_G-2)}M_1 - \sigma_z(p+q)x_G 2^{p+q} I\\ &+ \sqrt{2p(p-1)(N_D-1)}\begin{pmatrix} M_2 & 0 \\ 0 & 0\end{pmatrix} + p x_D \begin{pmatrix} I_{N_D} & 0 \\ 0 & 0\end{pmatrix}\Bigg] \end{aligned} \tag{43}$$

where $M_1 \sim GOE^{N_D+N_G-2}$ is a GOE matrix of size N_D+N_G-2, $M_2 \sim GOE^{N_D-1}$ is a GOE matrix of size N_D-1 and

$$J = \begin{pmatrix} -I_{N_D} & 0 \\ 0 & I_{N_G}\end{pmatrix}. \tag{44}$$

If follows that

$$\begin{aligned}|\det\tilde{H}| \overset{d}{=}&\Bigg|\det\Bigg[\sigma_z\sqrt{2^{p+q+1}(p+q)(p+q-1)(N_D+N_G-2)}M_1 - \sigma_z(p+q)x_G 2^{p+q} I\\ &+ \sqrt{2p(p-1)(N_D-1)}\begin{pmatrix} M_2 & 0 \\ 0 & 0\end{pmatrix} + p x_D \begin{pmatrix} I_{N_D} & 0 \\ 0 & 0\end{pmatrix}\Bigg]\Bigg|. \end{aligned} \tag{45}$$

Let $b = \sqrt{2^{p+q}(p+q)(p+q-1)}\sigma_z$, $b_1 = \sqrt{p(p-1)\kappa}$ and

$$x = \frac{\sigma_z(p+q)2^{p+q}}{\sqrt{N-2}}x_G, \quad x_1 = -\frac{p}{\sqrt{N-2}}x_D. \tag{46}$$

Then

$$|\det\tilde{H}| = (2(N-2))^{\frac{N-2}{2}}|\det H(x,x_1)|, \tag{47}$$

$$H(x,x_1) \equiv bM + b_1\begin{pmatrix} M_1 & 0 \\ 0 & 0\end{pmatrix} - x - x_1\begin{pmatrix} I_{N_D} & 0 \\ 0 & 0\end{pmatrix}. \tag{48}$$

The desired complexity term then comes from (15), (31)

$$\mathbb{E}C_N = K_N\int_B \sqrt{\frac{N-2}{2\pi s^2}}e^{-\frac{N-2}{2s^2}x^2}dx\sqrt{\frac{N-2}{2\pi s_1^2}}e^{-\frac{N-2}{2s_1^2}x_1^2}dx_1\mathbb{E}|\det H(x,x_1)| \tag{49}$$

where

$$K_N = \omega_{\kappa N}\omega_{\kappa' N}(2(N-2))^{\frac{N-2}{2}}(2\pi)^{-\frac{N-2}{2}}\left(p+\sigma_z^2 2^{p+1}(p+q)\right)^{-\frac{\kappa N-1}{2}}\left(\sigma_z^2 2^{p+q}(p+q)\right)^{-\frac{\kappa' N-1}{2}} \tag{50}$$

and

$$B = \left\{(x,x_1)\in\mathbb{R}^2 \ : \ x \leq \frac{1}{\sqrt{2}}(p+q)2^{p+q}u_G, \ x_1 \geq -(p+q)^{-1}2^{-(p+q)}px - \frac{p}{\sqrt{2}}u_D\right\} \tag{51}$$

and the variances are (recall (16), (20), (46))

$$s^2 = \frac{1}{2}\sigma_z^2(p+q)^2 2^{3(p+q)}, \quad s_1^2 = \frac{p^2}{2}, \tag{52}$$

and $\omega_N = \frac{2\pi^{N/2}}{\Gamma(N/2)}$ is the surface area of the N sphere. The domain of integration B arises from the contraints $L^{(D)} \in (-\infty, \sqrt{N}u_D)$ and $L^{(G)} \in (-\infty, \sqrt{N}u_G)$.

We will need the asymptotic behaviour of the constant K_N, which we now record in a small lemma.

Lemma 3.3. *As* $N \to \infty$*,*

$$K_N \sim 2^{\frac{N}{2}} \pi^{N/2} \left(\kappa^\kappa \kappa'^{\kappa'}\right)^{-N/2} \sqrt{\kappa\kappa'} \left(p + \sigma_z^2 2^{p+1}(p+q)\right)^{-\frac{\kappa N-1}{2}} \left(\sigma_z^2 2^{p+q}(p+q)\right)^{-\frac{\kappa' N-1}{2}} \tag{53}$$

Proof. By Stirling's formula

$$\begin{aligned} K_N \sim 4\pi^N \left(\frac{4\pi}{\kappa N}\right)^{-1/2} \left(\frac{4\pi}{\kappa' N}\right)^{-1/2} \left(\frac{\kappa N}{2e}\right)^{-\kappa N/2} \left(\frac{\kappa' N}{2e}\right)^{-\kappa' N/2} (2(N-2))^{\frac{N-2}{2}} (2\pi)^{-\frac{N-2}{2}} \\ \left(p + \sigma_z^2 2^{p+1}(p+q)\right)^{-\frac{\kappa N-1}{2}} \left(\sigma_z^2 2^{p+q}(p+q)\right)^{-\frac{\kappa' N-1}{2}} \\ \sim 2^{\frac{N}{2}} \pi^{N/2} \left(\kappa^\kappa \kappa'^{\kappa'}\right)^{-N/2} \sqrt{\kappa\kappa'} \left(p + \sigma_z^2 2^{p+1}(p+q)\right)^{-\frac{\kappa N-1}{2}} \left(\sigma_z^2 2^{p+q}(p+q)\right)^{-\frac{\kappa' N-1}{2}} \end{aligned} \tag{54}$$

where we have used $(N-2)^{\frac{N-2}{2}} = N^{\frac{N-2}{2}} \left(1 - \frac{2}{N}\right)^{\frac{N-2}{2}} \sim N^{\frac{N-2}{2}} e^{-N/2}$. □

4 Limiting spectral density of the Hessian

Our intention now is to compute the the expected complexity $\mathbb{E}C_N$ via the Coulomb gas method. The first step in this calculation is to obtain the limiting spectral density of the random matrix

$$H' = bM + b_1 \begin{pmatrix} M_1 & 0 \\ 0 & 0 \end{pmatrix} - x_1 \begin{pmatrix} I & 0 \\ 0 & 0 \end{pmatrix}, \tag{55}$$

where, note, $H = H' - xI$. Here the upper-left block is of dimension κN, and the overall dimension is N. Let μ_{eq} be the limiting spectral measure of H' and ρ_{eq} its density. The supersymmetric method provides a way of calculating the expected Stieltjes transforms of ρ_{eq} [Ver04]:

$$\langle G(z) \rangle = \frac{1}{N} \frac{\partial}{\partial J}\bigg|_{J=0} Z(J) \tag{56}$$

$$Z(J) := \mathbb{E}_{H'} \frac{\det(z - H' + J)}{\det(z - H')}. \tag{57}$$

Recall that a density and its Stieltjes transform are related by the Stieltjes inversion formula

$$\rho_{eq}(z) = \frac{1}{\pi} \lim_{\epsilon \to 0} \Im \langle G(z + i\epsilon) \rangle. \tag{58}$$

The function $Z(J)$ can be computed using a supersymmetric representation of the ratio of determinants. Firstly, we recall an elementary result from multivariate calculus, where M is a real matrix:

$$\int \prod_{i=1}^N \frac{d\phi_i d\phi_i^*}{2\pi} e^{-i\phi^\dagger M \phi} = \frac{1}{\det M}. \tag{59}$$

By introducing the notion of *Grassmann variables* and *Berezin integration*, we obtain a complimentary expression:

$$\int \prod_{i=1}^N \frac{d\chi_i d\chi_i^*}{-i} e^{-i\chi^\dagger M \chi} = \det M. \tag{60}$$

Here the χ_i, χ_i^* are purely algebraic objects defined by the anti-commutation rule

$$\chi_i \chi_j = -\chi_j \chi_i, \quad \forall i, j \tag{61}$$

and χ_i^* are separate objects, with the complex conjugation unary operator * defined so that $(\chi_i^*)^* = -\chi_i^*$, and Hermitian conjugation is then defined as usual by $\chi^\dagger = (\chi^T)^*$. The set of variables $\{\chi_i, \chi_i^*\}_{i=1}^N$ generate a *graded algebra* over $\mathbb{C}$. Mixed vectors of commuting and anti-commuting variables are called *supervectors*, and they belong to a vector space called *superspace*. The integration symbol $\int d\chi_i d\chi^*$ is defined as a formal algebraic linear operator by the properties

$$\int d\chi_i = 0, \qquad \int d\chi_i \, \chi_j = \delta_{ij}. \tag{62}$$

Functions of the the Grassmann variables are defined by their formal power series, e.g.

$$e^{\chi_i} = 1 + \chi_i + \frac{1}{2}\chi_i^2 + \ldots = 1 + \chi_i \tag{63}$$

where the termination of the series follows from $\chi_i^2 = 0 \;\; \forall i$, which is an immediate consequence of (61). From this it is apparent that (62), along with (61), is sufficient to define Berezin integration over arbitrary functions of arbitrary combinations of Grassmann variables. Using the integral results (59), (60) we can then write

$$\frac{\det(z - H' + J)}{\det(z - H')} = \int d\Psi \exp\left\{-i\phi^\dagger(z - H')\phi - i\chi^\dagger(z + J - H')\chi\right\} \tag{64}$$

where the measure is

$$d\Psi = \prod_{t=1}^{2} \frac{d\phi[t]d\phi^*[t]d\chi[t]d\chi^*[t]}{-2\pi i}, \tag{65}$$

ϕ is a vector of N complex commuting variables, χ and χ^* are vectors of N Grassmann variables, and we use the $[t]$ notation to denote the splitting of each of the vectors into the first κN and last $(1-\kappa)N$ components, as seen in [GW90]:

$$\phi = \begin{pmatrix} \phi[1] \\ \phi[2] \end{pmatrix}. \tag{66}$$

We then split the quadratic form expressions in (64)

$$\begin{aligned} &-\phi^\dagger(z - H')\phi - \chi^\dagger(z + J - H')\chi \\ =&-\phi[1]^\dagger(x_1 - b_1M_1)\phi[1] - \phi^\dagger(z - bM)\phi - \chi[1]^\dagger(x_1 - b_1M_1)\chi[1] - \chi^\dagger(z + J - bM)\chi. \end{aligned} \tag{67}$$

Taking the GOE averages is now simple [Ver04; Noc17]:

$$\mathbb{E}_M \exp\left\{-ib\phi^\dagger M\phi - ib\chi^\dagger M\chi\right\} = \exp\left\{-\frac{b^2}{4N}\text{trg}Q^2\right\}, \tag{68}$$

$$\mathbb{E}_M \exp\left\{-ib_1\phi[1]^\dagger M_1\phi[1] - ib_1\chi[1]^\dagger M_1\chi[1]\right\} = \exp\left\{-\frac{b_1^2}{4\kappa N}\text{trg}Q[1]^2\right\}, \tag{69}$$

where the supersymmetric matrices are given by

$$Q = \begin{pmatrix} \phi^\dagger\phi & \phi^\dagger\chi \\ \chi^\dagger\phi & \chi^\dagger\chi \end{pmatrix}, \quad Q[1] = \begin{pmatrix} \phi[1]^\dagger\phi[1] & \phi[1]^\dagger\chi[1] \\ \chi[1]^\dagger\phi[1] & \chi[1]^\dagger\chi[1] \end{pmatrix}. \tag{70}$$

Introducing the tensor notation

$$\psi = \phi \otimes \begin{pmatrix} 1 \\ 0 \end{pmatrix} + \chi \otimes \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \quad \psi[1] = \phi[1] \otimes \begin{pmatrix} 1 \\ 0 \end{pmatrix} + \chi[1] \otimes \begin{pmatrix} 0 \\ 1 \end{pmatrix} \tag{71}$$

and

$$\zeta = \begin{pmatrix} z & 0 \\ 0 & z + J \end{pmatrix} \tag{72}$$

we can compactly write

$$Z(J) = \int d\Psi \exp\left\{-\frac{b^2}{4N}\text{trg}Q^2 - \frac{b_1^2}{4\kappa N}\text{trg}Q[1]^2 - i\psi[1]^\dagger\psi[1]x_1 - i\psi^\dagger\zeta\psi\right\}. \tag{73}$$

We now perform two Hubbard-Stratonovich transformations [Ver04]

$$Z(J) = \int d\Psi d\sigma d\sigma[1] \exp\left\{-\frac{N}{b^2}\text{trg}\sigma^2 - \frac{\kappa N}{b_1^2}\text{trg}\sigma[1]^2 - i\psi[1]^\dagger(x_1 + \sigma[1])\psi[1] - i\psi^\dagger(\sigma + \zeta)\psi\right\}, \tag{74}$$

where σ and $\sigma[1]$ inherit their form from $Q, Q[1]$

$$\sigma = \begin{pmatrix} \sigma_{BB} & \sigma_{BF} \\ \sigma_{FB} & i\sigma_{FF} \end{pmatrix}, \quad \sigma[1] = \begin{pmatrix} \sigma_{BB}[1] & \sigma_{BF}[1] \\ \sigma_{FB}[1] & i\sigma_{FF}[1] \end{pmatrix} \tag{75}$$

with $\sigma_{BB}, \sigma_{FF}, \sigma_{BB}[1], \sigma_{FF}[1]$ real commuting variables, and $\sigma_{BF}, \sigma_{FB}, \sigma_{BF}[1], \sigma_{FB}[1]$ Grassmanns; the factor i is introduced to ensure convergence. Integrating out over $d\Psi$ is now a straightforward Gaussian integral in superspace, giving

$$\begin{aligned} Z(J) &= \int d\Psi d\sigma d\sigma[1] \exp\left\{-\frac{N}{b^2}\text{trg}\sigma^2 - \frac{\kappa N}{b_1^2}\text{trg}\sigma[1]^2 - i\psi[1]^\dagger(x_1+\zeta+\sigma+\sigma[1])\psi[1] - i\psi[2]^\dagger(\sigma+\zeta)\psi[2]\right\} \\ &= \int d\sigma d\sigma[1] \exp\left\{-\frac{N}{b^2}\text{trg}\sigma^2 - \frac{\kappa N}{b_1^2}\text{trg}\sigma[1]^2 - \kappa N \text{trg}\log(x_1+\zeta+\sigma+\sigma[1]) - \kappa' N\text{trg}\log(\sigma+\zeta)\right\} \\ &= \int d\sigma d\sigma[1] \exp\left\{-\frac{N}{b^2}\text{trg}(\sigma-\zeta)^2 - \frac{\kappa N}{b_1^2}\text{trg}\sigma[1]^2 - \kappa N \text{trg}\log(x_1+\sigma+\sigma[1]) - \kappa' N\text{trg}\log\sigma\right\}. \end{aligned} \tag{76}$$

Recalling the definition of ζ, we have

$$\text{trg}(\sigma-\zeta)^2 = (\sigma_{BB}-z)^2 - (i\sigma_{FF}-z-J)^2 \tag{77}$$

and so one immediately obtains

$$\begin{aligned} \frac{1}{N}\frac{\partial}{\partial J}\bigg|_{J=0} Z(J) &= \frac{2}{b^2}\int d\sigma d\sigma[1](z-i\sigma_{FF})\exp\left\{-\frac{N}{b^2}\text{trg}(\sigma-z)^2 - \frac{\kappa N}{b_1^2}\text{trg}\sigma[1]^2 - \kappa N\text{trg}\log(x_1+\sigma+\sigma[1]) - \kappa' N\text{trg}\log\sigma\right\} \\ &= \frac{2}{b^2}\int d\sigma d\sigma[1](z-i\sigma_{FF})\exp\left\{-\frac{N}{b^2}\text{trg}\sigma^2 - \frac{\kappa N}{b_1^2}\text{trg}\sigma[1]^2 - \kappa N\text{trg}\log(x_1+z+\sigma+\sigma[1]) - \kappa' N\text{trg}\log(z+\sigma)\right\} \end{aligned} \tag{78}$$

To obtain the limiting spectral density (LSD), or rather its Stieltjes transform, one must find the leading order term in the $N\to\infty$ expansion for (78). This can be done by using the saddle point method on the $\sigma, \sigma[1]$ manifolds. We know that the contents of the exponential must vanish at the saddle point, since the LSD is $\mathcal{O}(1)$, so we in fact need only compute σ_{FF} at the saddle point. We can diagonalise σ within the integrand of (78) and absorb the diagonalising graded $U(1/1)$ matrix into $\sigma[1]$. The resulting saddle point equations for the off-diagonal entries of the new (rotated) $\sigma[1]$ dummy variable are trivial and immediately give that $\sigma[1]$ is also diagonal at the saddle point. The saddle point equations are then

$$\frac{2}{b_1^2}\sigma_{BB}[1] + \frac{1}{\sigma_{BB}[1]+\sigma_{BB}+x_1+z} = 0 \tag{79}$$

$$\frac{2}{b^2}\sigma_{BB} + \frac{\kappa}{\sigma_{BB}[1]+\sigma_{BB}+x_1+z} + \frac{\kappa'}{\sigma_{BB}+x} = 0 \tag{80}$$

$$\frac{2}{b_1^2}\sigma_{FF}[1] - \frac{1}{\sigma_{FF}[1]+\sigma_{FF}-ix_1-iz} = 0 \tag{81}$$

$$\frac{2}{b^2}\sigma_{FF} - \frac{\kappa}{\sigma_{FF}[1]+\sigma_{FF}-ix_1-iz} - \frac{\kappa'}{\sigma_{FF}-iz} = 0. \tag{82}$$

(81) and (82) combine to give an explicit expression for $\sigma_{FF}[1]$:

$$\sigma_{FF}[1] = \frac{b_1^2}{2\kappa}\left(\frac{2}{b^2}\sigma_{FF} - \kappa'(\sigma_{FF}-iz)^{-1}\right). \tag{83}$$

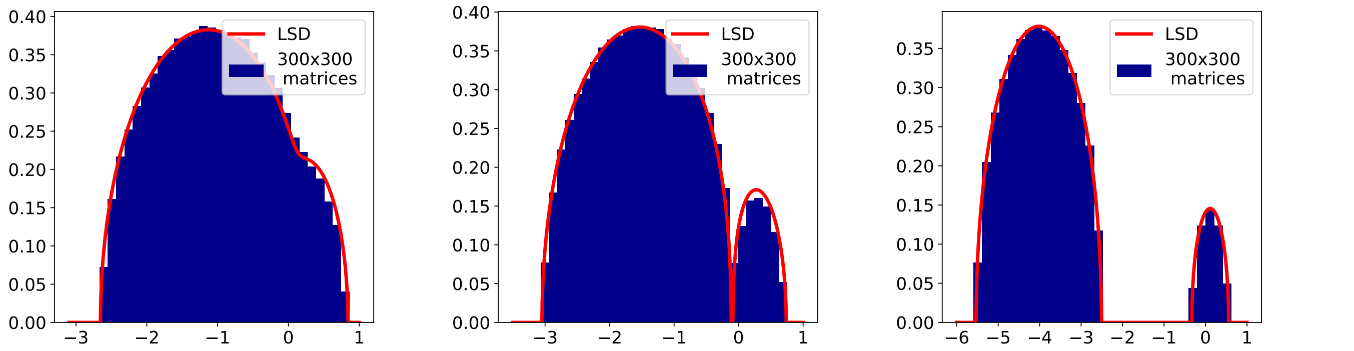
With a view to simplifying the numerical solution of the coming quartic, we define $t = i(\sigma_{FF}-iz)$ and then a line of manipulation with (82) and (83) gives

$$\left(t^2 - zt - \kappa' b^2\right)\left((1+\kappa^{-1}b^{-2}b_1^2)t^2 - (\kappa^{-1}b_1^2 b^{-2}z - x_1)t - \kappa'\kappa^{-1}b_1^2\right) + b^2\kappa t^2 = 0. \tag{84}$$

By solving (84) numerically for fixed values of κ, b, b_1, x_1, we can obtain the four solutions $t_1(z), t_2(z), t_3(z), t_4(z)$. These four solution functions arise from choices of branch for $(z, x_1)\in\mathbb{C}^2$ and determining the correct branch directly is highly non-trivial. However, for any $z\in\mathbb{R}$, at most one of the t_i will lead to a positive LSD, which gives a simple way to compute ρ_{eq} numerically using (58) and (78):

$$\rho_{eq}(z) = \max_i\left\{-\frac{2}{b^2\pi}\Im t_i(z)\right\}. \tag{85}$$

Plots generated using (85) and eigendecompositions of matrices sampled from the distribution of H' are given in Figure 1 and show good agreement between the two.



(a) Merged

(b) Touching

(c) Separate

Figure 1: Example spectra of H' showing empirical spectra from 100 300 × 300 matrices and the corresponding LSDs computed from (84). Here $b = b_1 = 1$, $\kappa = 0.9$, σ_z=1 and x_1 is varied to give the three different behaviours.

5 The asymptotic complexity

In the previous section, we have found the equilibrium measure, μ_{eq}, of the ensemble of random matrices

$$H' = bM + b_1 \begin{pmatrix} M_1 & 0 \\ 0 & 0 \end{pmatrix} - x_1 \begin{pmatrix} I & 0 \\ 0 & 0 \end{pmatrix}, \quad M \sim GOE^N,\ M_1 \sim GOE^{\kappa N}. \tag{86}$$

The Coulomb gas approximation gives us a method of computing $\mathbb{E}|\det(H'-x)|$:

$$\mathbb{E}|\det(H'-x)| \approx \exp\left\{ N \int \log|z-x| d\mu_{eq}(z) \right\}. \tag{87}$$

We have access to the density of μ_{eq} pointwise (in x and x_1) numerically, and so (87) is a matter of one-dimensional quadrature. Recalling (49), we then have

$$\mathbb{E}C_N \approx K'_N \iint_B dx dx_1 \ \exp\left\{ -(N-2)\left(\frac{1}{2s^2}x^2 + \frac{1}{2s_1^2}(x_1)^2 - \int \log|z-x| d\mu_{eq}(z) \right) \right\} \equiv K'_N \iint_B dx dx_1 \ e^{-(N-2)\Phi(x,x_1)} \tag{88}$$

where

$$K'_N = K_N \sqrt{\frac{N-2}{2\pi s_1^2}} \sqrt{\frac{N-2}{2\pi s^2}}. \tag{89}$$

Due to Lemma 3.3, the constant term has asymptotic form

$$\frac{1}{N}\log K'_N \sim \frac{1}{2}\log 2 + \frac{1}{2}\log\pi - \frac{\kappa}{2}\log\left(p + \sigma_z^2 2^{p+q}(p+q)\right) - \frac{\kappa'}{2}\log\left(\sigma_z^2(p+q)2^{p+q}\right) - \frac{\kappa}{2}\log\kappa - \frac{\kappa'}{2}\log\kappa' \equiv K \tag{90}$$

We then define the desired $\Theta(u_D, u_G)$ as

$$\lim \frac{1}{N}\log \mathbb{E}C_N = \Theta(u_D, u_G) \tag{91}$$

and we have

$$\Theta(u_D, u_G) = K - \min_B \Phi. \tag{92}$$

Using these numerical methods, we obtain the plot of Φ in B and a plot of Θ for some example p, q, σ_z, κ values, shown in Figures 2, 3. Numerically obtaining the maximum of Φ on B is not as onerous as it may appear, since $-\Phi$ grows quadratically in $|x|, |x_1|$ at moderate distances from the origin.

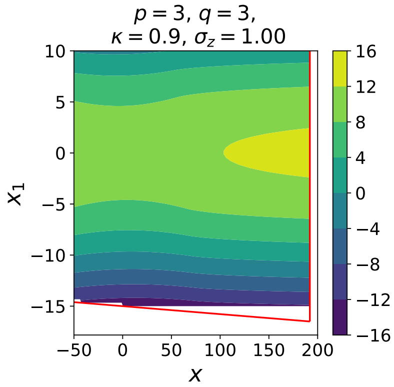


Figure 2: Φ for $p = q = 3, \sigma_z = 1, \kappa = 0.9$. Red lines show the boundary of the integration region B.

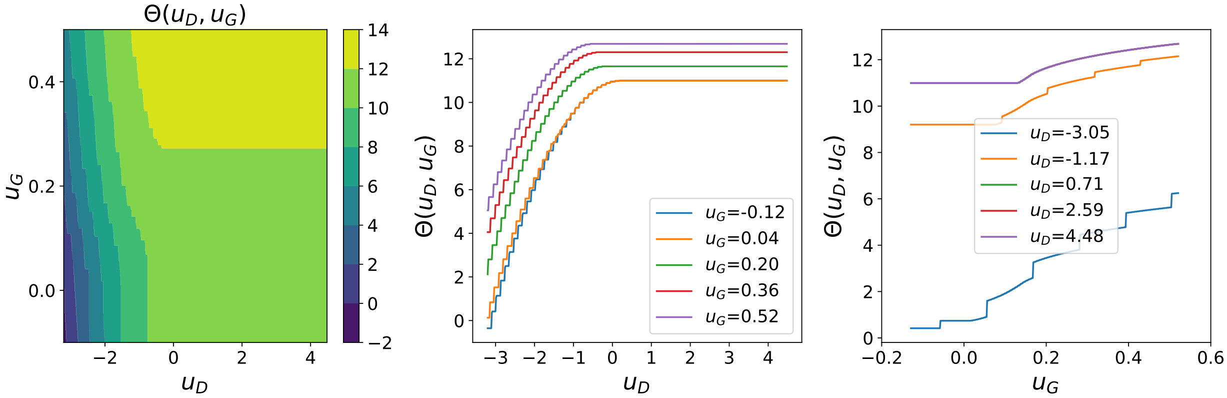


Figure 3: Θ and its cross-sections, fixing separately u_D and u_G. Here $p = q = 3, \sigma_z = 1, \kappa = 0.9$.

We numerically verify the legitimacy of this Coulomb point approximation with Monte Carlo integration

$$\mathbb{E}|\det(H' - x)| \approx \frac{1}{n}\sum_{i=1}^{n}\prod_{j=1}^{N}|\lambda_j^{(i)} - x|, \tag{93}$$

where $\lambda_j^{(i)}$ is the j-th eigenvalues of the i-th i.i.d. sample from the distribution of H'. The results, comparing $N^{-1}\log\mathbb{E}|\det(H' - x)|$ at $N = 50$ for a variety of x, x_1 are show in Figure 4. Note the strong agreement even at such modest N, however to rigorously substantiate the Coulomb gas approximation in (87), we must prove a concentration result.

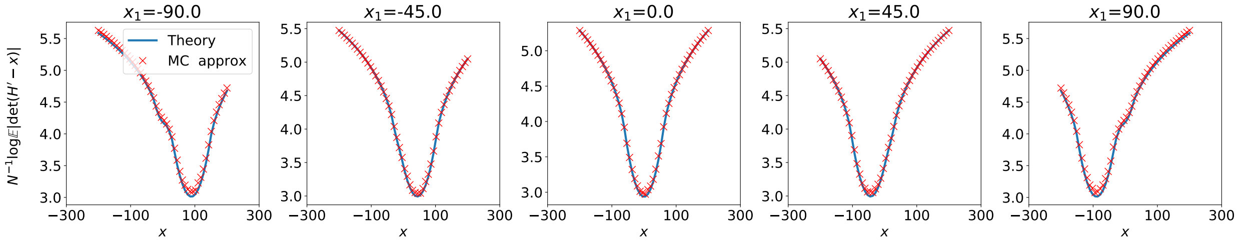


Figure 4: Comparison of (87) and (93), verifying the Coulomb gas approximation numerically. Here $p = q = 3, \sigma_z = 1, \kappa = 0.9$. Sampled matrices for MC approximation are dimension $N = 50$, and $n = 50$ MC samples have been used.

Lemma 5.1. *Let $(H_N)_{N=1}^\infty$ be a sequence of random matrices, where for each N*

$$H_N \overset{d}{=} bM + b_1 \begin{pmatrix} M_1 & 0 \\ 0 & 0 \end{pmatrix} - x_1 \begin{pmatrix} I & 0 \\ 0 & 0 \end{pmatrix} \tag{94}$$

and $M \sim GOE^N$, $M_1 \sim GOE^{\kappa N}$. Let μ_N be the empirical spectral measure of H_N and say $\mu_N \to \mu_{eq}$ weakly almost surely. Then for any $(x, x_1) \in \mathbb{R}^2$

$$\mathbb{E}|\det(H_N - xI)| = \exp\left\{N(1+o(1)) \int \log|z-x| d\mu_{eq}(z)\right\} \tag{95}$$

as $N \to \infty$.

Proof. We begin by establishing an upper bound. For any $\alpha, \beta > 0$

$$\begin{aligned}\mathbb{E}|\det(H_N - xI)| &= \mathbb{E}\left[\exp\left\{N \int \log|z-x| d\mu_N(z)\right\}\right] \\ &\leq \underbrace{\left(\mathbb{E}\left[\exp\left\{2N \int \max\left(-\alpha, \min\left(\log|x-z|, \beta\right)\right) d\mu_N(z)\right\}\right]\right)^{1/2}}_{A_N} \\ &\quad \underbrace{\left(\mathbb{E}\left[\exp\left\{2N \int \log|x-z| \mathbb{1}\{|x-z| \geq e^\beta\} d\mu_N(z)\right\}\right]\right)^{1/2}}_{B_N}.\end{aligned} \tag{96}$$

Considering B_N, we have

$$\log|x-z| \mathbb{1}\{|x-z| \geq e^\beta\} \leq |x-z|^{1/2} \mathbb{1}\{|x-z| \geq e^\beta\} \leq e^{-\beta/2}|x-z| \tag{97}$$

and so

$$\begin{aligned}\mathbb{E}\left[\exp\left\{2N \int \log|x-z| \mathbb{1}\{|x-z| \geq e^\beta\}\right\}\right] &\leq \mathbb{E}\left[\exp\left\{2Ne^{-\beta/2} \frac{\text{Tr}|H_N - xI|}{N}\right\}\right] \\ &= \mathbb{E}\left[\exp\left\{2e^{-\beta/2} \text{Tr}|H_N - xI|\right\}\right].\end{aligned} \tag{98}$$

The entries of H_N are Gaussians with variance $\frac{1}{N}b^2$, $\frac{1}{2N}b^2$, $\frac{1}{N}(b^2 + b_1^2)$ or $\frac{1}{2N}(b^2 + b_1^2)$ and all the diagonal and upper diagonal entries are independent. All of these variances are $\mathcal{O}(N^{-1})$, so

$$|H_N - x|_{ij} \leq |x| + |x_1| + \mathcal{O}(N^{-1/2})|X_{ij}| \tag{99}$$

where the X_{ij} are i.i.d. standard Gaussians for $i \leq j$. It follows that

$$\mathbb{E}\left[\exp\left\{2e^{-\frac{\beta}{2}} \text{Tr}|H_N - xI|\right\}\right] \leq e^{2e^{-\frac{\beta}{2}} N(|x|+|x_1|)} \mathbb{E}_{X \sim \mathcal{N}(0,1)} e^{2e^{-\frac{\beta}{2}} \mathcal{O}(N^{1/2})|X|}. \tag{100}$$

Elementary calculations give

$$\mathbb{E}_{X \sim \mathcal{N}(0,1)} e^{c|X|} \leq \frac{1}{2}\left(e^{-c^2} + e^{c^2}\right) \leq e^{c^2} \tag{101}$$

and so

$$\mathbb{E}\left[\exp\left\{2e^{-\frac{\beta}{2}} \text{Tr}|H_N - xI|\right\}\right] \leq e^{2e^{-\frac{\beta}{2}} N(|x|+|x_1|)} e^{4e^{-\beta}\mathcal{O}(N)} = \exp\left\{2N\left(e^{-\frac{\beta}{2}}(|x| + |x_1|) + e^{-\beta}\mathcal{O}(1)\right)\right\} \tag{102}$$

thus when we take $\beta \to \infty$, we have $B_N \leq e^{o(N)}$.

Considering A_N, it is sufficient now to show

$$\mathbb{E}\left[\exp\left\{2N \int f(z) d\mu_N(z)\right\}\right] = \exp\left\{2N\left(\int f(z) d\mu_{eq}(z) + o(1)\right)\right\} \tag{103}$$

where $f(z) = 2\max\left(\min(\log|x-z|, \beta), -\alpha\right)$, a continuous and bounded function. For any $\epsilon > 0$, we have

$$\mathbb{E}\left[\exp\left\{2N \int f(z) d\mu_N(z)\right\}\right] \leq \exp\left\{2N\left(\int f(z) d\mu_{eq}(z) + \epsilon\right)\right\} + e^{2N||f||_\infty} \mathbb{P}\left(\int f(z) d\mu_N(z) \geq \int f(z) d\mu_{eq}(z) + \epsilon\right). \tag{104}$$

The entries of H_N are Gaussian with $\mathcal{O}(N^{-1})$ variance and so obey a log-Sobolev inequality as required by Theorem 1.5 from [GZ+00]. The constant, c, in the inequality is independent of N, x, x_1, so we need not compute it exactly. The theorem from [GZ+00] then gives

$$\mathbb{P}\left(\int f(z)d\mu_N(z) \geq \int f(z)d\mu_{eq}(z) + \epsilon\right) \leq \exp\left\{-\frac{N^2}{8c}\epsilon^2\right\}. \tag{105}$$

We have shown

$$\mathbb{E}|\det(H_N - xI)| \leq A_N B_N \leq \exp\left\{N(1+o(1))\left(\int f(z)d\mu_{eq}(z)\right)\right\} \leq \exp\left\{N(1+o(1))\left(\int \log|x-z|d\mu_{eq}(z)\right)\right\}. \tag{106}$$

We now need to establish a complimentary lower bound to complete the proof. By Jensen's inequality

$$\begin{aligned}\mathbb{E}|\det(H_N - x)| &\geq \exp\left(N\mathbb{E}\left[\int \log|z-x|d\mu_N(z)\right]\right)\\ &\geq \exp\left(N\mathbb{E}\left[\int \max\left(-\alpha, \log|z-x|\right)d\mu_N(z)\right]\right)\exp\left(N\mathbb{E}\left[\int \log|z-x|\mathbb{1}\{|z-x|\leq e^{-\alpha}\}d\mu_N(z)\right]\right)\\ &\geq \exp\left(N\mathbb{E}\left[\int \min\left(\beta, \max\left(-\alpha, \log|z-x|\right)\right)d\mu_N(z)\right]\right)\exp\left(N\mathbb{E}\left[\int \log|z-x|\mathbb{1}\{|z-x|\leq e^{-\alpha}\}d\mu_N(z)\right]\right)\end{aligned} \tag{107}$$

for any $\alpha, \beta > 0$. Convergence in law of μ_N to μ_{eq} and the dominated convergence theorem give

$$\exp\left(N\mathbb{E}\left[\int \min\left(\beta, \max\left(-\alpha, \log|z-x|\right)\right)d\mu_N(z)\right]\right) \geq \exp\left\{N\left(\int \log|x-z|d\mu_{eq}(z) + o(1)\right)\right\} \tag{108}$$

for large enough β, because μ_{eq} has compact support. It remains to show that the expectation inside the exponent in the second term of (107) converges to zero uniformly in N in the limit $\alpha \to \infty$.

By (58), it is sufficient to consider $\langle G_N(z)\rangle$, which is computed via (78). Let us define the function Ψ so that

$$\langle G_N(z)\rangle = \frac{2}{b^2}\int d\sigma d\sigma[1](z - i\sigma_{FF})e^{-N\Psi(\sigma,\sigma[1])}. \tag{109}$$

Henceforth, $\sigma_{FF}^*, \sigma_{FF}[1]^*, \sigma_{BB}^*, \sigma_{BB}[1]^*$ are the solution to the saddle point equations (79-82) and $\tilde{\sigma}_{FF}, \tilde{\sigma}_{FF}[1], \tilde{\sigma}_{BB}, \tilde{\sigma}_{BB}[1]$ are integration variables. Around the saddle point

$$z - i\sigma_{FF} = z - i\sigma_{FF}^* - iN^{-\frac{1}{r}}\tilde{\sigma}_{FF} \tag{110}$$

for some $r \geq 2$. We use the notation $\boldsymbol{\sigma}$ for $(\sigma_{BB}, \sigma_{BB}[1], \sigma_{FF}, \sigma_{FF}[1])$ and similarly $\boldsymbol{\sigma}_{BB}, \boldsymbol{\sigma}_{FF}$. A superscript asterisk on Ψ or any of its derivatives is short hand for evaluation at the saddle point. While the Hessian of Ψ may not in general vanish at the saddle point,

$$\int d\tilde{\sigma}d\tilde{\sigma}[1]\tilde{\sigma}_{FF}e^{-N\tilde{\boldsymbol{\sigma}}^T\nabla^2\Psi^*\tilde{\boldsymbol{\sigma}}} = 0 \tag{111}$$

and so we must go to at least the cubic term in the expansion of Ψ around the saddle point, i.e.

$$\langle G_N(z)\rangle = G(z) - \frac{2i}{b^2N^{5/3}}\underbrace{\int_{-\infty}^{\infty} d\tilde{\boldsymbol{\sigma}}_{BB}d\tilde{\boldsymbol{\sigma}}_{FF}\tilde{\sigma}_{FF}e^{-\frac{1}{6}\tilde{\sigma}^i\tilde{\sigma}^j\tilde{\sigma}^k\partial_{ijk}\Psi^*}}_{E(z;x_1)} + \text{exponentially smaller terms.} \tag{112}$$

The bosonic (BB) and fermionic (FF) coordinates do not interact, so we can consider derivatives of Φ as block tensors. Simple differentiation gives

$$\begin{aligned}(\nabla\Psi)_B &= \begin{pmatrix}\frac{2}{b^2}\sigma_{BB} - \kappa\left(\sigma_{BB}+\sigma_{BB}[1]+z+x_1\right)^{-1} - \kappa'\left(\sigma_{BB}+z\right)^{-1}\\ \frac{2}{b_1^2}\sigma_{BB}[1] - \left(\sigma_{BB}+\sigma_{BB}[1]+z+x_1\right)^{-1}\end{pmatrix}\\ \implies (\nabla^2\Psi)_B &= \begin{pmatrix}\kappa\left(\sigma_{BB}+\sigma_{BB}[1]+z+x_1\right)^{-2} + \kappa'\left(\sigma_{BB}+z\right)^{-2} & \kappa\left(\sigma_{BB}+\sigma_{BB}[1]+z+x_1\right)^{-2}\\ \left(\sigma_{BB}+\sigma_{BB}[1]+z+x_1\right)^{-2} & \left(\sigma_{BB}+\sigma_{BB}[1]+z+x_1\right)^{-2}\end{pmatrix} \quad (113)\\ \implies (\nabla^3\Psi)_B^* &= \left(\begin{pmatrix}A_B\kappa + B_B\kappa' & A_B\kappa\\ A_B & A_B\end{pmatrix}, A_B\begin{pmatrix}\kappa & \kappa\\ 1 & 1\end{pmatrix}\right), \quad (114)\end{aligned}$$

where

$$A_B = -\frac{2}{\left(\sigma_{BB}^* + \sigma_{BB}^*[1] + z + x_1\right)^3}, \quad B_B = -\frac{2}{\left(\sigma_{BB}^* + z\right)^3}. \tag{115}$$

$(\nabla^3\Psi)_F^*$ follows similarly with

$$A_F = -\frac{2}{\left(\sigma_{FF}^* + \sigma_{FF}^*[1] - iz - ix_1\right)^3}, \quad B_F = -\frac{2}{\left(\sigma_{FF}^* - iz\right)^3}. \tag{116}$$

By the saddle point equations (79)-(82) we have

$$A_B = 2(\sigma_{BB}[1]^*)^3, \;\; B_B = \frac{2}{(\kappa')^3}\left(\frac{2\kappa}{b_1^2}\sigma_{BB}[1]^* - \frac{2}{b^2}\sigma_{BB}^*\right)^3 \tag{117}$$

$$A_F = 2(\sigma_{FF}[1]^*)^3, \;\; B_F = \frac{2}{(\kappa')^3}\left(\frac{2\kappa}{b_1^2}\sigma_{FF}[1]^* - \frac{2}{b^2}\sigma_{FF}^*\right)^3. \tag{118}$$

Let $\xi_1 = \tilde{\sigma}_{BB}, \xi_2 = \tilde{\sigma}_{BB}[1]$. Then

$$\begin{aligned}(\tilde{\sigma}^i\tilde{\sigma}^j\tilde{\sigma}^k\partial_{ijk}\Phi^*)_B &= \left(A_B\kappa + B_B\kappa'\right)\xi_1^3 + A_B(2\kappa+1)\xi_1^2\xi_2[1] + A_B(\kappa+2)\xi_1\xi_2^2 + A_B\xi_2^3 \\ &= A_B\left[\xi_2^3 + (2\kappa+1)\xi_2\xi_1^2 + (2+\kappa)\xi_1\xi_2^2 + C\xi_1^3\right] + \left(B_B\kappa' + A_B\kappa - CA_B\right)\xi_1^3\end{aligned} \tag{119}$$

for any C. Let $\xi_1 = a_1\xi_1'$ and then choose $C = a_1^{-3}$ and $a_1 = (2+\kappa)(2\kappa+1)^{-1}$ to give

$$(\tilde{\sigma}^i\tilde{\sigma}^j\tilde{\sigma}^k\partial_{ijk}\Phi^*)_B = A_B(\xi_1' + \xi_2)^3 + (B_B\kappa' + A_B\kappa - CA_B)a_1^3(\xi_1')^3 \equiv A_B\eta^3 + D_B\xi^3 \tag{120}$$

with $\eta = \xi_1' + \xi_2$, $\xi = \xi_1'$, $D_B = B_B\kappa' + A_B\kappa - a_1^{-3}A_B$. The expressions for $(\tilde{\sigma}^i\tilde{\sigma}^j\tilde{\sigma}^k\partial_{ijk}\Phi^*)_F$ follow identically. We thus have

$$E(z;x_1) \propto \left(\int_0^\infty d\xi\,\xi\int_\xi^\infty d\eta\, e^{A_F\eta^3 + D_F\xi^3}\right)\left(\int_0^\infty d\xi\,\int_\xi^\infty d\eta\, e^{A_B\eta^3 + D_B\xi^3}\right) \tag{121}$$

or perhaps with the the integration ranges reversed depending on the signs of $\Re A_F, \Re A_B, \Re D_F, \Re D_B$. We have

$$\begin{aligned}|E(z;x_1)| &\leq \left|\int_0^\infty d\xi\,\xi\int_\xi^\infty d\eta\, e^{A_F\eta^3 + D_F\xi^3}\right| \cdot \left|\int_0^\infty d\xi\,\int_\xi^\infty d\eta\, e^{A_B\eta^3 + D_B\xi^3}\right| \\ &\leq \int_0^\infty d\xi\,\xi\int_\xi^\infty d\eta\, |e^{A_F\eta^3 + D_F\xi^3}| \cdot \int_0^\infty d\xi\,\int_\xi^\infty d\eta\, |e^{A_B\eta^3 + D_B\xi^3}| \\ &\leq \int_0^\infty d\xi\,\xi\int_0^\infty d\eta\, |e^{A_F\eta^3 + D_F\xi^3}| \cdot \int_0^\infty d\xi\,\int_0^\infty d\eta\, |e^{A_B\eta^3 + D_B\xi^3}| \\ &\leq \left(|\mathfrak{M}D_F|\right)^{-2/3}\left(|\mathfrak{M}A_F|\right)^{-1/3}\left(|\mathfrak{M}D_B|\right)^{-1/3}\left(|\mathfrak{M}A_B|\right)^{-1/3}\left(\int_0^\infty e^{-\xi^3}d\xi\right)^3\left(\int_0^\infty \xi e^{-\xi^3}d\xi\right)\end{aligned} \tag{122}$$

where we have defined

$$\mathfrak{M}y = \begin{cases}\Re y & \text{if } \Re y \neq 0, \\ \Im y & \text{if } \Re y = 0.\end{cases} \tag{123}$$

This last bound follows from a standard Cauchy rotation of integration contour if any of D_F, A_F, D_B, A_B has vanishing real part. (122) is valid for $D_B, A_B, D_F, A_F \neq 0$, but if $D_B = 0$ and $A_B \neq 0$, then the preceding calculations are simplified and we still obtain an upper bound but proportional to $(|\mathfrak{M}A_B|)^{-1/3}$. Similarly with $A_B = 0$ and $D_B \neq 0$ and similarly for A_F, D_F. The only remaining cases are $A_B = D_B = 0$ or $A_F = D_F = 0$. But recall (118) and (81)-(82). We immediately see that $A_F = D_F$ if and only if $\sigma_{FF} = \sigma_{FF}[1] = 0$, which occurs for no finite z, x_1. Therefore, for *fixed* $(x, x_1) \in \mathbb{R}^2$, $\alpha > 0$ and any $z \in (x - e^{-\alpha}, x + e^{-\alpha})$

$$|\mathbb{E}\mu_N(z) - \mu_{eq}(z;x_1)| \lesssim N^{-5/3}C(x_1, |x| + e^{-\alpha}) \tag{124}$$

where $C(|x_1|, |x| + e^{-\alpha})$ is positive and is decreasing in α. Since μ_{eq} is bounded, it follows that $\mathbb{E}\mu_N$ is bounded, and therefore

$$\mathbb{E}\int \log|z - x|\mathbb{1}\{|z - x| \leq e^{-\alpha}\}d\mu_N(z) \to 0 \tag{125}$$

as $\alpha \to \infty$ uniformly in N, and so the lower bound is completed. □

Equipped with this result, we can now prove the legitimacy of the Coulomb gas approximation in our complexity calculation. The proof will require an elementary intermediate result which has undoubtedly appeared in various places before, but we prove it here anyway for the avoidance of doubt.

Lemma 5.2. *Let M_N be a random $N \times N$ symmetric real matrix with independent centred Gaussian upper-diagonal and diagonal entries. Suppose that the variances of the entries are bounded above by cN^{-1} for some constant $c > 0$. Then there exists some constant c_e such that*

$$\mathbb{E}||M_N||_{max}^N \lesssim e^{c_e N}. \tag{126}$$

Proof. Let σ_{ij}^2 denote the variance of M_{ij}. Then

$$\begin{aligned}\mathbb{E}||M||_{max}^N &\leq \sum_{i,j} \mathbb{E}|M_{i,j}|^N \\ &= \sum_{i,j} \mathbb{E}|\mathcal{N}(0, \sigma_{ij}^2)|^N \\ &= \sum_{i,j} \sigma_{ij}^N \mathbb{E}|\mathcal{N}(0,1)|^N \\ &\leq N^2 c^{N/2} N^{-N/2} \mathbb{E}|\mathcal{N}(0,1)|^N. \end{aligned} \tag{127}$$

Simple integration with a change of variables gives

$$\mathbb{E}|\mathcal{N}(0,1)|^N = 2^{\frac{N+1}{2}} \Gamma\left(\frac{N+1}{2}\right) \tag{128}$$

and then, for large enough N, Stirling's formula gives

$$\begin{aligned}\mathbb{E}|\mathcal{N}(0,1)|^N &\sim 2^{\frac{N+1}{2}} \sqrt{\pi(N+1)} \left(\frac{N+1}{2e}\right)^{\frac{N-1}{2}} \\ &\sim 2\sqrt{\pi} e^{-\frac{N-1}{2}} N^{N/2} \left(\frac{N+1}{N}\right)^{N/2} \\ &\sim 2\sqrt{\pi e} N^{N/2}. \end{aligned} \tag{129}$$

So finally

$$\mathbb{E}||M||_{max}^N \lesssim N^2 c^{N/2} = e^{\frac{1}{2}N\log c + 2\log N} \leq e^{\left(\frac{1}{2}\log c + 2\right)N}, \tag{130}$$

so defining $c_e = \frac{1}{2}\log 2 + 2$ gives the result. □

Theorem 5.3. *For any $x_1 \in \mathbb{R}$, let H_N be a random $N \times N$ matrix distributed as in the statement of Lemma 5.1. Then as $N \to \infty$*

$$\begin{aligned}&\iint_B dx dx_1 \ \exp\left\{-N\left(\frac{1}{2s^2}x^2 + \frac{1}{2s_1^2}(x_1)^2\right)\right\} \mathbb{E}|\det(H_N(x_1) - x)| \\ =& \iint_B dx dx_1 \ \exp\left\{-N\left(\frac{1}{2s^2}x^2 + \frac{1}{2s_1^2}(x_1)^2 - \int \log|z - x| d\mu_{eq}(z) + o(1)\right)\right\} + o(1). \end{aligned} \tag{131}$$

Proof. Let $R > 0$ be some constant, independent of N. Introduce the notation $B_{\leq R} = B \cap \{\boldsymbol{z} \in \mathbb{R}^2 \mid |z| \leq R\}$, and then

$$\begin{aligned}&\left| \iint_B dx dx_1 \ \exp\left\{-N\left(\frac{1}{2s^2}x^2 + \frac{1}{2s_1^2}(x_1)^2\right)\right\} \mathbb{E}|\det(H_N(x_1) - x)| \right. \\ &\left. - \iint_{B_{\leq R}} dx dx_1 \ \exp\left\{-N\left(\frac{1}{2s^2}x^2 + \frac{1}{2s_1^2}(x_1)^2\right)\right\} \mathbb{E}|\det(H_N(x_1) - x)| \right| \\ \leq& \iint_{||\boldsymbol{x}|| \geq R} dx dx_1 \ \exp\left\{-N\left(\frac{1}{2s^2}x^2 + \frac{1}{2s_1^2}(x_1)^2\right)\right\} \mathbb{E}|\det(H_N(x_1) - x)|. \end{aligned} \tag{132}$$

We have the upper bound (106) of Lemma 5.1 but this cannot be directly applied to (132) since the bound relies on uniformity in x, x_1 which can only be established for bounded x, x_1. We use a much cruder bound instead. First, let

$$J_N = H_N + x_1 \left(\begin{array}{cc} I & 0 \\ 0 & 0 \end{array} \right) \tag{133}$$

and then

$$|\det\left(H_N - xI\right)| \leq ||J_N||_{\max}^N \max\{|x|, |x_1|\}^N = ||J_N||_{\max}^N \exp\left(N \max\{\log|x|, \log|x_1|\}\right). \tag{134}$$

J_N has centred Gaussian entries with variance $\mathcal{O}(N^{-1})$, so Lemma 5.2 applies, and we find

$$\mathbb{E}|\det\left(H_N - xI\right)| \lesssim \exp\left(N \max\{\log|x|, \log|x_1|\}\right) e^{c_e N} \tag{135}$$

for some constant $c_e > 0$ which is independent of x, x_1 and N, but we need not compute it.

Now we have

$$\begin{aligned}
&\left| \iint_B dx dx_1 \, \exp\left\{-N\left(\frac{1}{2s^2}x^2 + \frac{1}{2s_1^2}(x_1)^2\right)\right\} \mathbb{E}|\det(H_N(x_1) - x)| \right. \\
&\left. - \iint_{B_{\leq R}} dx dx_1 \, \exp\left\{-N\left(\frac{1}{2s^2}x^2 + \frac{1}{2s_1^2}(x_1)^2\right)\right\} \mathbb{E}|\det(H_N(x_1) - x)| \right| \\
&\lesssim \iint_{||\boldsymbol{x}|| \geq R} dx dx_1 \, \exp\left\{-N\left(\frac{1}{2s^2}x^2 + \frac{1}{2s_1^2}(x_1)^2 - \max\{\log|x|, \log|x_1|\} - c_e\right)\right\}.
\end{aligned} \tag{136}$$

But, since μ_{eq} is bounded and has compact support, we can choose R large enough (independent of N) so that

$$\frac{1}{2s^2}x^2 + \frac{1}{2s_1^2}(x_1)^2 - \max\{\log|x|, \log|x_1|\} - c_e > L > 0 \tag{137}$$

for all (x, x_1) with $\sqrt{x^2 + x_1^2} > R$ and for some fixed L independent of N. Whence

$$\begin{aligned}
&\left| \iint_B dx dx_1 \, \exp\left\{-N\left(\frac{1}{2s^2}x^2 + \frac{1}{2s_1^2}(x_1)^2\right)\right\} \mathbb{E}|\det(H_N(x_1) - x)| \right. \\
&\left. - \iint_{B_{\leq R}} dx dx_1 \, \exp\left\{-N\left(\frac{1}{2s^2}x^2 + \frac{1}{2s_1^2}(x_1)^2\right)\right\} \mathbb{E}|\det(H_N(x_1) - x)| \right| \\
&\lesssim N^{-1} e^{-NL} \to 0
\end{aligned} \tag{138}$$

as $N \to \infty$. Finally, for x, x_1 in $B_{\leq R}$, the result of the Lemma 5.1 holds uniformly in x, x_1, so

$$\begin{aligned}
&\iint_{B_{\leq R}} dx dx_1 \, \exp\left\{-N\left(\frac{1}{2s^2}x^2 + \frac{1}{2s_1^2}(x_1)^2\right)\right\} \mathbb{E}|\det(H_N(x_1) - x)| \\
&= \iint_{B_{\leq R}} dx dx_1 \, \exp\left\{-N\left(\frac{1}{2s^2}x^2 + \frac{1}{2s_1^2}(x_1)^2 - \int \log|z - x| d\mu_{eq}(z; x_1) + o(1)\right)\right\}.
\end{aligned} \tag{139}$$

The result follows from (138), (139) and the triangle inequality. □

5.1 Asymptotic complexity with prescribed Hessian index

Recall the complexity defined in (8):

$$\begin{aligned}
C_{N,k_D,k_G} = \Bigg| \Bigg\{ &\boldsymbol{w}^{(D)} \in S^{N_D}, \boldsymbol{w}^{(G)} \in S^{N_G} \; : \; \nabla_D L^{(D)} = 0, \nabla_G L^{(G)} = 0, L^{(D)} \in B_D, L^{(G)} \in B_G \\
&i(\nabla_D^2 L^{(D)}) = k_D, \; i(\nabla_G^2 L^{(G)}) = k_G \Bigg\} \Bigg|.
\end{aligned} \tag{8}$$

The extra Hessian signature conditions in (8) enforce that both generator and discriminator are at low-index saddle points. Our method for computing the complexity C_N in the previous subsection relies on the Coulomb gas approximation applied to the spectrum of H'. However, the Hessian index constraints are formulated in the natural Hessian matrix (40), but our spectral calculations proceed from the rewritten form (45). We find however that we can indeed proceed much as in [Bas+20]. Recall the key Hessian matrix $\tilde{H}$ given in (40) by

$$\tilde{H} = \begin{pmatrix} \sqrt{2(N_D-1)}\sqrt{b^2+b_1^2}M^{(D)} & -bG \\ bG^T & \sqrt{2(N_G-1)}bM^{(G)} \end{pmatrix} - \sqrt{N-2}x\begin{pmatrix} -I_{N_D} & 0 \\ 0 & I_{N_G} \end{pmatrix} + \sqrt{N-2}x_1\begin{pmatrix} I_{N_D} & 0 \\ 0 & 0 \end{pmatrix} \tag{140}$$

where $M^{(D)} \sim GOE^{N_D-1}$, $M^{(G)} \sim GOE^{N_G-1}$, G is $N_D - 1 \times N_G - 1$ Ginibre, and all are independent. Note that we have used (46) to slightly rewrite (40). We must address the problem of computing

$$\mathbb{E}|\det \tilde{H}|\mathbb{1}\left\{i\left(\sqrt{\kappa}(1+\mathcal{O}(N^{-1}))\sqrt{b^2+b_1^2}M_D + \frac{x+x_1}{\sqrt{2}}\right) = k_D,\ i\left(\sqrt{\kappa'}(1+\mathcal{O}(N^{-1}))bM_G - \frac{x}{\sqrt{2}}\right) = k_G\right\}. \tag{141}$$

Indeed, we introduce integration variables $\boldsymbol{y}_1, \boldsymbol{y}_2, \zeta_1, \zeta_1^*, \zeta_2, \zeta_2^*$, being $(N-2)$-vectors of commuting and anti-commuting variables respectively. Use $[t]$ notation to split all vectors into the first $\kappa N - 1$ and last $\kappa' N - 1$ components. Let

$$A[t] = \boldsymbol{y}_1\boldsymbol{y}_1^T + \boldsymbol{y}_2\boldsymbol{y}_2^T + \zeta_1\zeta_1^\dagger + \zeta_2\zeta_2^\dagger. \tag{142}$$

With these definitions, we have [Bas+20]

$$\begin{aligned}|\det \tilde{H}| = (2(N-2))^{\frac{N-2}{2}}\lim_{\epsilon\searrow 0}\int d\Xi \exp\left\{-i\sqrt{\kappa}(1+\mathcal{O}(N^{-1}))\sqrt{b^2+b_1^2}\mathrm{Tr}M^{(D)}A[1] - i\sqrt{\kappa'}(1+\mathcal{O}(N^{-1}))b\mathrm{Tr}M^{(G)}A[2]\right\} \\ \exp\{\mathcal{O}(\epsilon)\}\exp\{\ldots\}\end{aligned} \tag{143}$$

where $d\Xi$ is the normalised measure of the $\boldsymbol{y}_1, \boldsymbol{y}_2, \zeta_1, \zeta_1^*, \zeta_2, \zeta_2^*$ and the ellipsis represents terms with no dependence on $M^{(D)}$ or $M^{(G)}$, which we need not write down. The crux of the matter is that we must compute

$$\mathbb{E}_{M^{(D)}}e^{-i\sqrt{\kappa}\sqrt{b^2+b_1^2}\mathrm{Tr}M^{(D)}A[1]}\mathbb{1}\left\{i\left(M_D + \frac{x+x_1}{\sqrt{\kappa}\sqrt{b^2+b_1^2}}(1+\mathcal{O}(N^{-1}))\right) = k_D\right\}, \tag{144}$$

$$\mathbb{E}_{M^{(G)}}e^{-i\sqrt{\kappa'}b\mathrm{Tr}M^{(G)}A[2]}\mathbb{1}\left\{i\left(M_G - \frac{x}{\sqrt{\kappa'}b}(1+\mathcal{O}(N^{-1}))\right) = k_G\right\}, \tag{145}$$

but [Bas+20] has performed exactly these calculations (see around (5.146) therein) and so there exist constants $K_U^{(D)}, K_L^{(D)}, K_U^{(G)}, K_L^{(G)}$ such that

$$\begin{aligned}&K_L^{(D)}e^{-Nk_D\kappa(1+o(1))I_1(\hat{x}_D;\sqrt{2})}e^{-\frac{1}{2N}(b^2+b_1^2)\mathrm{Tr}A[1]^2} \\ \leq&\Re\mathbb{E}_{M^{(D)}}e^{-i\sqrt{\kappa}\sqrt{b^2+b_1^2}\mathrm{Tr}M^{(D)}A[1]}\mathbb{1}\left\{i\left(M_D + \frac{x+x_1}{\sqrt{\kappa}\sqrt{b^2+b_1^2}}(1+\mathcal{O}(N^{-1}))\right) = k_D\right\} \\ \leq&K_U^{(D)}e^{-Nk_D\kappa(1+o(1))I_1(\hat{x}_D;\sqrt{2})}e^{-\frac{1}{2N}(b^2+b_1^2)\mathrm{Tr}A[1]^2}\end{aligned} \tag{146}$$

and

$$\begin{aligned}&K_L^{(G)}e^{-Nk_G\kappa'(1+o(1))I_1(\hat{x}_G;\sqrt{2})}e^{-\frac{1}{2N}b^2\mathrm{Tr}A[2]^2} \\ \leq&\Re\mathbb{E}_{M^{(G)}}e^{-i\sqrt{\kappa'}b\mathrm{Tr}M^{(G)}A[2]}\mathbb{1}\left\{i\left(M_G - \frac{x}{\sqrt{\kappa'}b}(1+\mathcal{O}(N^{-1}))\right) = k_G\right\} \\ \leq&K_U^{(G)}e^{-Nk_G\kappa'(1+o(1))I_1(\hat{x}_G;\sqrt{2})}e^{-\frac{1}{2N}b^2\mathrm{Tr}A[2]^2}\end{aligned} \tag{147}$$

where

$$\hat{x}_D = -\frac{x+x_1}{\sqrt{\kappa}\sqrt{b^2+b_1^2}},\ \ \hat{x}_G = \frac{x}{\sqrt{\kappa'}b}. \tag{148}$$

Here I_1 is the rate function of the largest eigenvalue of the GOE as obtained in [ADG01] and used in [Auf13; Bas+20]:

$$I_1(u;E) = \begin{cases} \frac{2}{E^2}\int_u^{-E}\sqrt{z^2-E^2}dz & \text{for } u < -E, \\ \frac{2}{E^2}\int_E^u\sqrt{z^2-E^2}dz & \text{for } u > E, \\ \infty & \text{for } |u| < E. \end{cases} \tag{149}$$

Note that for $u < -E$

$$I_1(u;E) = -\frac{u}{E}\sqrt{u^2 - E^2} - \log\left(-u + \sqrt{u^2 - E^2}\right) + \log E \tag{150}$$

and for $u > E$ we simply have $I_1(u;E) = I_1(-u;E)$. Note also that $I_1(ru;E) = I_1(u, E/r)$.

We have successfully dealt with the Hessian index indicators inside the expectation, however we need some way of returning to the form of $\tilde{H}$ in (45) so the complexity calculations using the Coulomb gas approach can proceed as before. We can achieve this with inverse Fourier transforms:

$$e^{-\frac{1}{2N}(b^2+b_1^2)\mathrm{Tr}A[1]^2} = \mathbb{E}_{M_D} e^{-i\sqrt{\kappa}\sqrt{b^2+b_1^2}\mathrm{Tr}M_D A[1]} \tag{151}$$

$$e^{-\frac{1}{2N}b^2\mathrm{Tr}A[2]^2} = \mathbb{E}_{M_G} e^{-i\sqrt{\kappa'}b\mathrm{Tr}M_G A[2]} \tag{152}$$

from which we obtain

$$\begin{aligned}
&K_L e^{-Nk_D\kappa(1+o(1))I_1(\hat{x}_D;\sqrt{2})} e^{-Nk_G\kappa'(1+o(1))I_1(\hat{x}_G;\sqrt{2})}\mathbb{E}|\det\tilde{H}| \\
\leq &\mathbb{E}|\det\tilde{H}|\mathbb{1}\left\{i\left(\sqrt{\kappa}(1+\mathcal{O}(N^{-1}))\sqrt{b^2+b_1^2}M_D + \frac{x+x_1}{\sqrt{2}}\right) = k_D,\ i\left(\sqrt{\kappa'}(1+\mathcal{O}(N^{-1}))bM_G - \frac{x}{\sqrt{2}}\right) = k_G\right\} \quad (153) \\
\leq &K_U e^{-Nk_D\kappa(1+o(1))I_1(\hat{x}_D;\sqrt{2})} e^{-Nk_G\kappa'(1+o(1))I_1(\hat{x}_G;\sqrt{2})}\mathbb{E}|\det\tilde{H}|. \quad (154)
\end{aligned}$$

It follows that

$$\begin{aligned}
&K'_N \iint_B dxdx_1 e^{-(N-2)\left[\Phi(x,x_1)+k_G\kappa' I_1(x;\sqrt{2\kappa'}b)+k_D\kappa I_1\left((-(x+x_1);\sqrt{2\kappa(b^2+b_1^2)}\right)\right](1+o(1))} \\
\lesssim &C_{N,k_D,k_G} \\
\lesssim &K'_N \iint_B dxdx_1 e^{-(N-2)\left[\Phi(x,x_1)+k_G\kappa' I_1(x;\sqrt{2\kappa'}b)+k_D\kappa I_1\left((-(x+x_1);\sqrt{2\kappa(b^2+b_1^2)}\right)\right](1+o(1))}. \quad (155)
\end{aligned}$$

So we see that the relevant exponent in this case is the same as for C_N but with additional GOE eigenvalue large deviation terms, giving the complexity limit

$$\lim \frac{1}{N}\log\mathbb{E}C_{N,k_D,k_G} = \Theta_{k_D,k_G}(u_D,u_G) = K - \min_B\left\{\Phi + k_G\kappa' I_1(x;\sqrt{2\kappa'}b) + k_D\kappa I_1\left(-(x+x_1);\sqrt{2\kappa(b^2+b_1^2)}\right)\right\}. \tag{156}$$

Plots of Θ_{k_D,k_G} for a few values of k_D, k_G are shown in Figure 5.

6 Implications

6.1 Structure of low-index critical points

We examine the fine structure of the low-index critical points for both spin glasses. [Cho+15] used the 'banded structure' of low-index critical points to explain the effectiveness of gradient descent in large multi-layer perceptron neural networks. We undertake to uncover the analogous structure in our dual spin-glass model and thence offer explanations for GAN training dynamics with gradient descent. For a range of (k_D, k_G) values, starting at $(0,0)$, we compute Θ_{k_D,k_G} on an appropriate domain. In the (u_D, u_G) plane, we then find the maximum k_D, and separately k_G, such that $\Theta_{k_D,k_G}(u_D,u_G) > 0$. In the large N limit, this procedure reveals the regions in the (u_D, u_G) plane where critical points of each index of the two spin glasses are found. Figure 6 plots these maximum k_D, k_G values as contours on a shared (u_D, u_G) plane. Figure 7 shows the same results, but on separate filled contour plots; the white regions in the plots clearly show the 'ground state' boundary beyond which no critical points exist. We use some fixed values of the various parameters: $p = q = 3, \sigma_z = 1, \kappa = 0.9$.

These plots reveal, unsurprisingly perhaps, that something resembling the banded structure of [Cho+15] is present, with the higher index critical points being limited to higher loss values for each network. The 2-dimensional analogues of the E_∞ boundary of [Cho+15] are evident in the bunching of the k_D, k_G contours at higher values. There is, however further structure not present in the single spin-glass multi-layer perceptron model. Consider the contour of $k_D = 0$ at the bottom of the full contour plot in Figure 6. Imagine traversing a path near this contour from right to left (decreasing u_D values). At all points along such a path, the only critical points present are exact local minima for both networks, however the losses range over

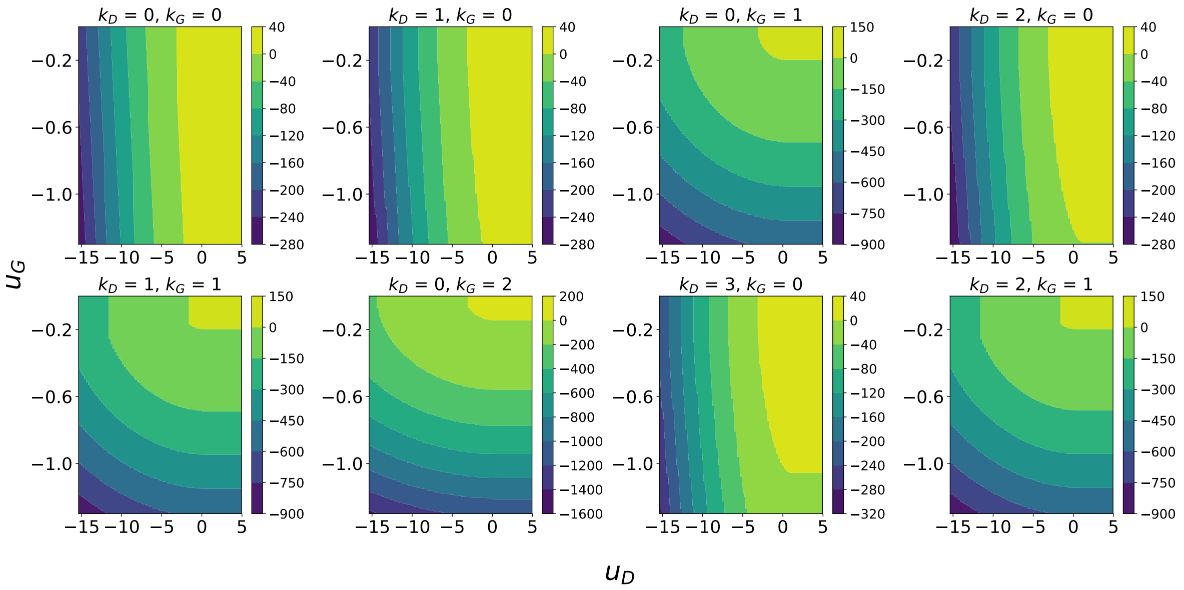


Figure 5: Contour plots of Θ_{k_D,k_G} for a few values of k_D, k_G. Here $p = q = 3, \sigma_z = 1, \kappa = 0.9$.

(i) low generator loss, high discriminator loss;

(ii) some balance between generator and discriminator loss;

(iii) high generator loss, low discriminator loss.

These three states correspond qualitatively to known GAN phenomena:

(i) discriminator collapses to predicting 'real' for all items;

(ii) successfully trained model;

(iii) generator collapses to producing garbage samples which the discriminator trivially identifies.

Overall, the analysis of our model reveals a loss surface that favours convergence to states of low loss for *at least one of the networks*, but not necessarily both. Moreover, our plots of Θ and Θ_{k_D,k_G} in Figures 3, 5 demonstrate clearly the competition between the two networks, with the minimum attainable discriminator loss increasing as the generator loss decreases and vice-versa. We thus have a qualitative similarity between the minimax dynamics of real GANs and our model, including the existence of a Nash equilibrium, but also a new two-dimensional banded critical points structure. This structure offers the following explanation of large GAN training dynamics with gradient descent:

1. As with single feed-forward networks, the loss surface geometry encourages convergence to globally low values of at least one of the network losses.
2. The same favourable geometry encourages convergence to successful states, where both networks achieve reasonably low loss, but also encourages convergence to failure states, where the generator's samples are too easily distinguished by the discriminator, or the discriminator has entirely failed thus providing no useful training signal to the generator.

6.2 Hyperparameter effects

Our proposed model for GANs includes a few fixed hyperparameters that we expect to control features of the model, namely σ_z and κ. Based on the results of [Auf13; Cho+15; Bas+20], and the form of our analytical results above, we do not expect p and q (the number of layers in the discriminator and generator) to have any interesting effect beyond $p, q \geq 3$; this is clearly a limitation of the model. We would expect there to exist an optimal value of σ_z that would result in minimum loss, in some sense. The effect of κ is less clear, though we guess that, in the studied $N \to \infty$ limit, all $\kappa \in (0, 1)$ are effectively equivalent. Intuitively, $\kappa \in \{0, 1\}$ should result in the much larger network beating the smaller in the minimax game, however our results above are valid strictly for $\kappa \in (0, 1)$.

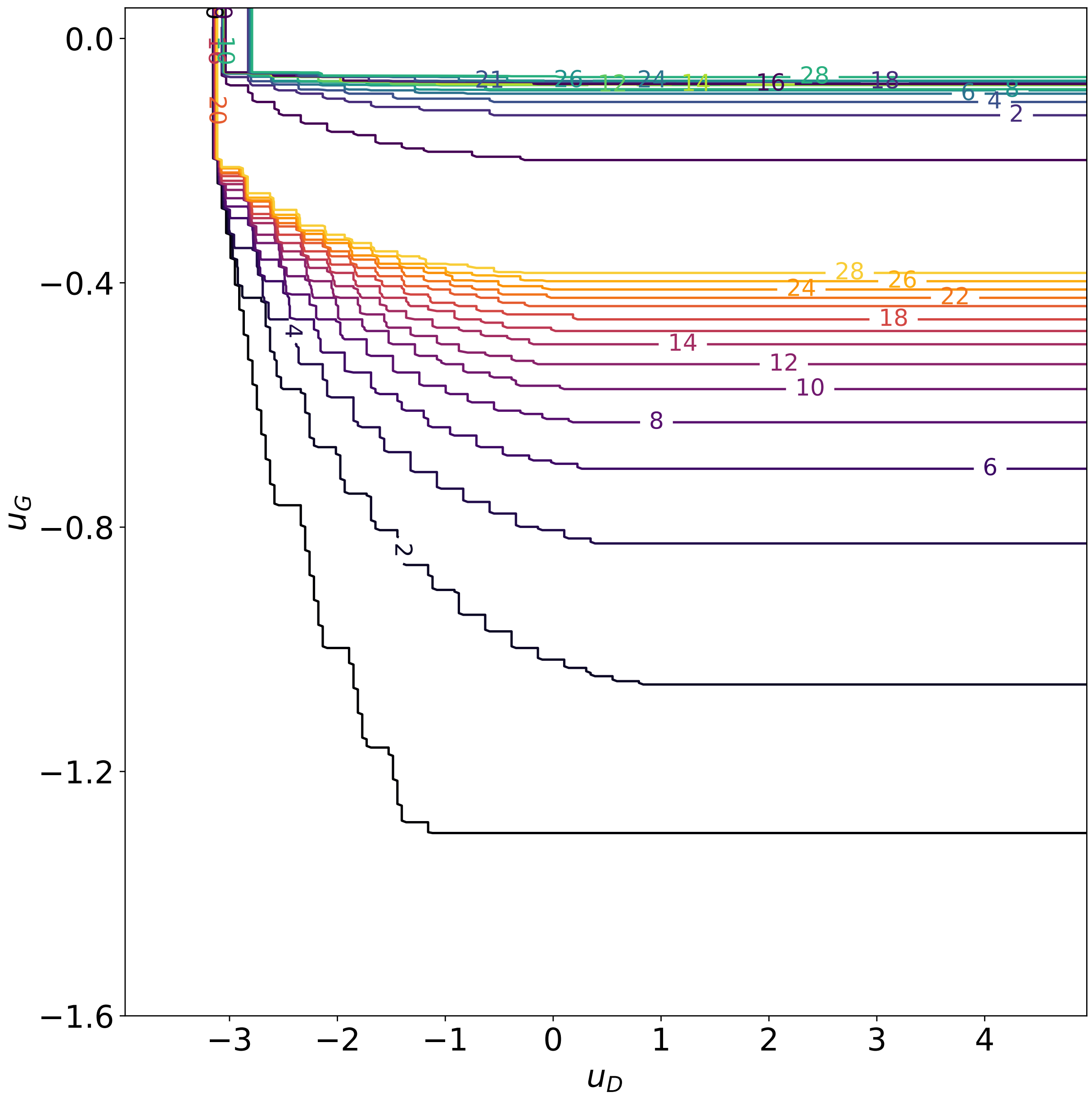


Figure 6: Contours in the (u_D, u_G) plane of the maximum k_D and k_G such that $\Theta_{k_D,k_G}(u_D, u_G) > 0$. k_D results shown with a red colour red scheme, and k_G with green. Only alternate contours are shown, in the interests of clarity. Here $p = q = 3, \sigma_z = 1, \kappa = 0.9$.

6.2.1 Effect of variance ratio

In the definition of complexity, u_D and u_G are upper bounds on the loss of the discriminator and generator, respectively. We are interested in the region of the u_D, u_G plane such that $\Theta(u_D, u_G) > 0$, this being the region where gradient descent algorithms are expected to become trapped. We therefore investigate the minimum loss such that $\Theta > 0$, this being, for a given σ_z, the theoretical minimum loss attainable by the GAN. We consider two natural notions of loss:

1. $\vartheta_D = \min\{u_D \in \mathbb{R} \mid \exists u_G \in \mathbb{R} \ : \ \Theta(u_D, u_G) > 0\}$;
2. $\vartheta_G = \min\{u_G \in \mathbb{R} \mid \exists u_D \in \mathbb{R} \ : \ \Theta(u_D, u_G) > 0\}$.

We vary σ_z over a range of values in $(10^{-5}, 10^2)$ and compute ϑ_D, ϑ_G.

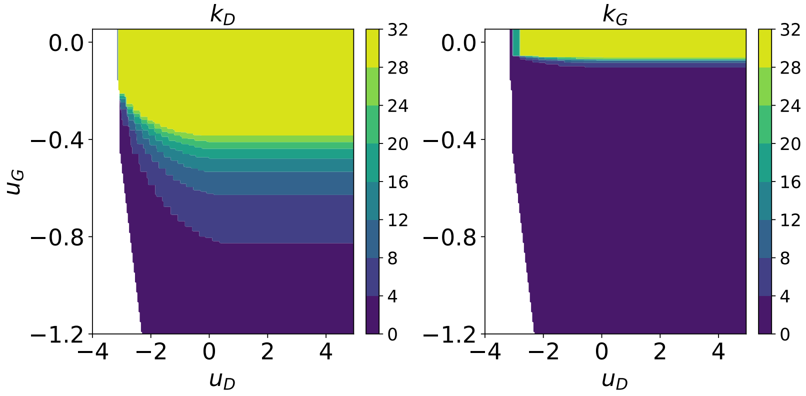
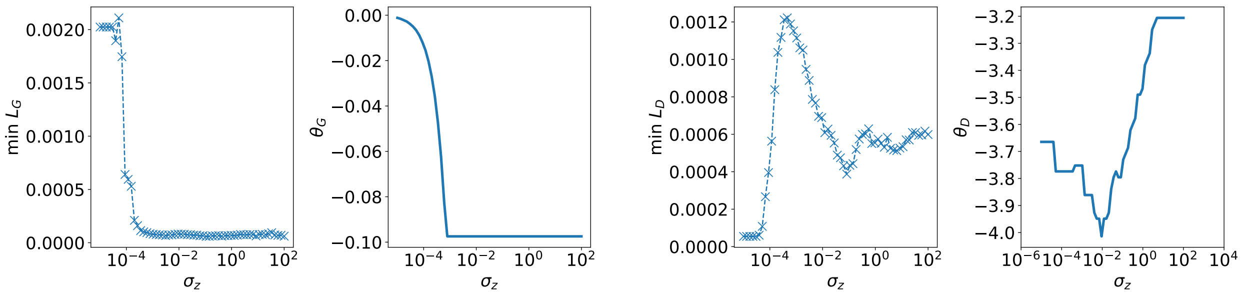


Figure 7: Contours in the (u_D, u_G) plane of the maximum k_D and k_G such that $\Theta_{k_D,k_G}(u_D, u_G) > 0$. k_D results on the left, k_G on the right. Here $p = q = 3, \sigma_z = 1, \kappa = 0.9$.

To compare the theoretical predictions of the effect of σ_z to real GANs, we perform a simple set of experiments. We use a DCGAN architecture [RMC15] with 5 layers in each network, using the reference PyTorch implementation from [Ink18], however we introduce the generator noise scale σ_z. For a given σ_z, we train the GANs for 10 epochs on CIFAR10 [KH+09] and record the generator and discriminator losses. For each σ_z, we repeat the experiment 30 times and average the minimum attained generator and discriminator losses to account for random variations between runs with the same σ_z. We note that the sample variances of the loss were typically very high, despite the PyTorch random seed being fixed across all runs. We plot the sample means, smoothed with rolling averaging over a short window, in the interest of clearly visualising whatever trends are present. The results are shown in Figure 8.

There is a striking similarity between the generator plots, with a sharp decline between $\sigma_z = 10^{-5}$ and around 10^{-3}, after which the minimum loss is approximately constant. The picture for the discriminator is less clear. Focusing on the sections $\sigma_z > 10^{-3}$, both plots show a clear minimum, at around $\sigma_z = 10^{-1}$ in experiments and $\sigma_z = 10^{-2}$ in theory. Note that the scales on the y-axes of these plots should not be considered meaningful. Though there is not precise correspondence between the discriminator curves, we claim that both theory and experiment tell the same qualitative story: increasing σ_z to at least around 10^{-3} gives the lowest theoretical generator loss, and then further increasing to, tentatively, some value in $(10^{-2}, 10^{-1})$ gives the lowest possible discriminator loss at no detriment to the generator.



(a) Generator

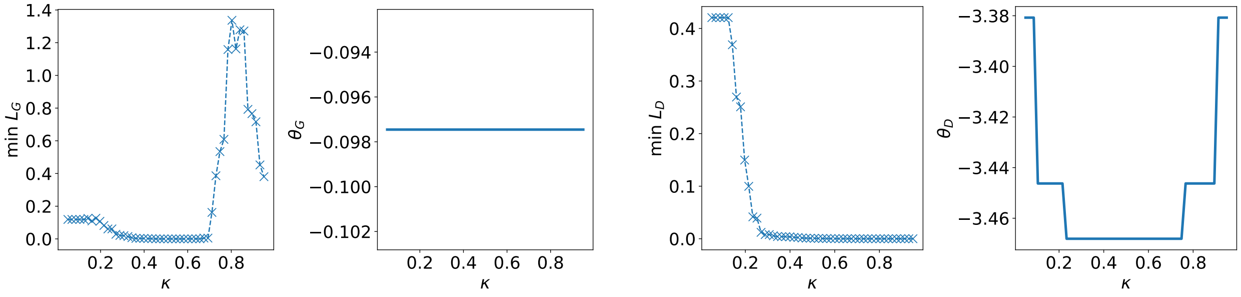
(b) Discriminator

Figure 8: The effect of σ_z. Comparison of theoretical predictions of minimum possible discriminator and generator losses to observed minimum losses when training DCGAN on CIFAR10. The left plots in each case (cross-dashed) show the experimental DCGAN results, and the right plots show the theoretical results θ_G, θ_D. $p = q = 5$ and $\kappa = 0.5$ are used in the theoretical calculations, to best match the DCGAN architecture. σ_z is shown on a log-scale.

6.2.2 Effect of size ratio

Similarly to the previous section, we can investigate the effect of κ using ϑ_D, ϑ_G while varying κ over $(0, 1)$. To achieve this variation in the DCGAN, we vary the number of convolutional filters in each network. The generator and discriminator are essentially mirror images of each other and the number of filters in each intermediate layer are defined as increasing functions[3] of some positive integers n_G, n_D. We fix $n_D + n_G = 128$ and vary n_D to obtain a range of κ values, with $\kappa = \frac{n_d}{n_d+n_g}$. The results are shown in Figure 9.

The theoretical model predicts a a broad range of equivalently optimal κ values centred on $\kappa = 0.5$ from the perspective of the discriminator loss, and no effect of κ on the generator loss. The experimental results similarly show a broad range of equivalently optimal κ centred around $\kappa = 0.5$, however there appear to be deficiencies in our model, particularly for higher κ values. The results of the experiments are intuitively sensible: the generator loss deteriorates for κ closer to 1, i.e. when the discriminator has very many more parameters than the generator, and vice-versa for small κ.



(a) Generator

(b) Discriminator

Figure 9: The effect of κ. Comparison of theoretical predictions of minimum possible discriminator and generator losses to observed minimum losses when training DCGAN on CIFAR10. The left plots in each case (cross-dashed) show the experimental DCGAN results, and the right plots show the theoretical results ϑ_G, ϑ_D. $p = q = 5$ and $\sigma_z = 1$ are used in the theoretical calculations, to best match the DCGAN architecture.

7 Conclusions and outlook

We have contributed a novel model for the study of large neural network gradient descent dynamics with statistical physics techniques, namely an interacting spin-glass model for generative adversarial neural networks. We believe this is the first attempt in the literature to incorporate advanced architectural features of modern neural networks, beyond basic single network multi-layer perceptrons, into such statistical physics style models. We have conducted an asymptotic complexity analysis via Kac-Rice formulae and Random Matrix Theory calculations of the energy surface of this model, acting as a proxy for GAN training loss surfaces of large networks. Our analysis has revealed a banded critical point structure as seen previously for simpler models, explaining the surprising success of gradient descent in such complicated loss surfaces, but with added structural features that offer explanations for the greater difficulty of training GANs compared to single networks. We have used our model to study the effect of some elementary GAN hyper-parameters and compared with experiments training real GANs on a standard computer vision dataset. We believe that the interesting features of our model, and their correspondence with real GANs, are yet further compelling evidence for the role of statistical physics effects in deep learning and the value of studying such models as proxies for real deep learning models, and in particular the value of concocting more sophisticated models that reflect aspects of modern neural network design and practice.

From a mathematical perspective, we have extensively studied the limiting spectral density of a novel random matrix ensemble using supersymmetric methods. In the preparation of this paper, we made considerable efforts to complete the average absolute value determinant calculations directly using a supersymmetric representation, as seen in [Bas+20], however this was found to be analytically intractable (as expected), but also extremely troublesome numerically (essentially due to analytically intractable and highly complicated Riemann sheet structure in $\mathbb{C}^2$). We were able to sidestep these issues by instead using a Coulomb gas approximation, whose validity we have rigorously proved using a novel combination of concentration arguments and supersymmetric asymptotic expansions. We have verified with numerical simulations our derived mean spectral density for the relevant Random Matrix Theory ensemble and also the accuracy of the Coulomb gas approximation.

[3] Number of filters in a layer is either proportional to n_D or n_D^2 depending on the layer (and similarly with n_G).

We hope that future work will be inspired to further study models of neural networks such as we have considered here. Practically, it would be exciting to explore the possibility of using our insights into GAN loss surfaces to devise algorithmic methods of avoiding training failure. Mathematically, the local spectral statistics of our random matrix ensemble may be interesting to study, particularly around the cusp where the two disjoint components of the limiting spectral density merge.

8 Acknowledgements

FM is grateful for support from the University Research Fellowship of the University of Bristol. JPK is pleased to acknowledge support from a Royal Society Wolfson Research Merit Award and ERC Advanced Grant 740900 (LogCorRM). NPB is grateful to Diego Granziol for useful comments and discussions.

References

[ACB17] Martin Arjovsky, Soumith Chintala, and Léon Bottou. "Wasserstein GAN". In: (2017).

[ADG01] G Ben Arous, Amir Dembo, and Alice Guionnet. "Aging of spherical spin glasses". In: *Probability theory and related fields* 120.1 (2001), pp. 1–67.

[Aro+19] Gerard Ben Arous et al. "The landscape of the spiked tensor model". In: *Communications on Pure and Applied Mathematics* 72.11 (2019), pp. 2282–2330.

[AT09] Robert J Adler and Jonathan E Taylor. *Random fields and geometry*. Springer Science & Business Media, 2009.

[Auf13] Auffinger, Antonio and Arous, Gerard Ben and Cerny, Jiri. "Random matrices and complexity of spin glasses". In: *Communications on Pure and Applied Mathematics* 66.2 (2013), pp. 165–201.

[Bai+19] Marco Baity-Jesi et al. "Comparing dynamics: Deep neural networks versus glassy systems". In: *Journal of Statistical Mechanics: Theory and Experiment* 2019.12 (2019), p. 124013.

[Bas+20] Nicholas P Baskerville et al. "The Loss Surfaces of Neural Networks with General Activation Functions". In: *arXiv preprint arXiv:2004.03959* (2020).

[Cho+15] Anna Choromanska et al. "The loss surfaces of multilayer networks". In: *Artificial intelligence and statistics*. 2015, pp. 192–204.

[CLA15] Anna Choromanska, Yann LeCun, and Gérard Ben Arous. "Open problem: The landscape of the loss surfaces of multilayer networks". In: *Conference on Learning Theory*. 2015, pp. 1756–1760.

[Con+17] Alexis Conneau et al. "Very Deep Convolutional Networks for Text Classification". In: *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*. Valencia, Spain: Association for Computational Linguistics, Apr. 2017, pp. 1107–1116. URL: `https://www.aclweb.org/anthology/E17-1104`.

[Dev+19] Jacob Devlin et al. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, June 2019, pp. 4171–4186. DOI: `10.18653/v1/N19-1423`. URL: `https://www.aclweb.org/anthology/N19-1423`.

[EV01] Andreas Engel and Christian Van den Broeck. *Statistical mechanics of learning*. Cambridge University Press, 2001.

[FFR19] Giampaolo Folena, Silvio Franz, and Federico Ricci-Tersenghi. "Rethinking mean-field glassy dynamics and its relation with the energy landscape: the awkward case of the spherical mixed p-spin model". In: *arXiv preprint arXiv:1903.01421* (2019).

[FW07] Yan V Fyodorov and Ian Williams. "Replica symmetry breaking condition exposed by random matrix calculation of landscape complexity". In: *Journal of Statistical Physics* 129.5-6 (2007), pp. 1081–1116.

[Fyo04] Yan V Fyodorov. "Complexity of random energy landscapes, glass transition, and absolute value of the spectral determinant of random matrices". In: *Physical review letters* 92.24 (2004), p. 240601.

[Gar88] Elizabeth Gardner. "The space of interactions in neural network models". In: *Journal of physics A: Mathematical and general* 21.1 (1988), p. 257.

[Goo+14] Ian Goodfellow et al. "Generative Adversarial Nets". In: *Advances in Neural Information Processing Systems 27*. Ed. by Z. Ghahramani et al. Curran Associates, Inc., 2014, pp. 2672–2680. URL: `http://papers.nips.cc/paper/5423-generative-adversarial-nets.pdf`.

[Goo+16] Ian Goodfellow et al. *Deep learning*. Vol. 1. 2. MIT press Cambridge, 2016.

[Gra+19] Diego Granziol et al. “Towards understanding the true loss surface of deep neural networks using random matrix theory and iterative spectral methods”. In: (2019).

[Gra20] Diego Granziol. “Beyond Random Matrix Theory for Deep Networks”. In: *arXiv preprint arXiv:2006.07721* (2020).

[Guh91] Thomas Guhr. “Dyson’s correlation functions and graded symmetry”. In: *Journal of mathematical physics* 32.2 (1991), pp. 336–347.

[GW90] T Guhr and HA Weidenmüller. “Isospin mixing and spectral fluctuation properties”. In: *Annals of Physics* 199.2 (1990), pp. 412–446.

[GZ+00] Alice Guionnet, Ofer Zeitouni, et al. “Concentration of the spectral measure for large matrices”. In: *Electronic Communications in Probability* 5 (2000), pp. 119–136.

[He+16] Kaiming He et al. “Deep residual learning for image recognition”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 770–778.

[HS97] Sepp Hochreiter and Jürgen Schmidhuber. “Long short-term memory”. In: *Neural computation* 9.8 (1997), pp. 1735–1780.

[Ink18] Nathan Inkawhich. *DCGAN Faces Tutorial*. Accessed on 29/10/20. 2018. URL: `https://github.com/pytorch/tutorials/blob/master/beginner_source/dcgan_faces_tutorial.py`.

[KH+09] Alex Krizhevsky, Geoffrey Hinton, et al. “Learning multiple layers of features from tiny images”. In: (2009).

[KLA20] Tero Karras, Samuli Laine, and Timo Aila. “A Style-Based Generator Architecture for Generative Adversarial Networks.” In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2020), pp. 1–1.

[KS87] I Kanter and Haim Sompolinsky. “Associative recall of memory without errors”. In: *Physical Review A* 35.1 (1987), p. 380.

[LT16] Ming-Yu Liu and Oncel Tuzel. “Coupled Generative Adversarial Networks”. In: *Proceedings of the 30th International Conference on Neural Information Processing Systems*. Vol. 29. 2016, pp. 469–477.

[MAB19] Antoine Maillard, Gérard Ben Arous, and Giulio Biroli. “Landscape Complexity for the Empirical Risk of Generalized Linear Models”. In: *arXiv preprint arXiv:1912.02143* (2019).

[Man+19a] Stefano Sarao Mannelli et al. “Passed & spurious: Descent algorithms and local minima in spiked matrix-tensor models”. In: *arXiv preprint arXiv:1902.00139* (2019).

[Man+19b] Stefano Sarao Mannelli et al. “Who is Afraid of Big Bad Minima? Analysis of gradient-flow in spiked matrix-tensor models”. In: *Advances in Neural Information Processing Systems*. 2019, pp. 8676–8686.

[MO14] Mehdi Mirza and Simon Osindero. “Conditional Generative Adversarial Nets”. In: *arXiv preprint arXiv:1411.1784* (2014).

[Nis01] Hidetoshi Nishimori. *Statistical physics of spin glasses and information processing: an introduction*. 111. Clarendon Press, 2001.

[Noc17] Andre Nock. “Characteristic Polynomials of Random Matrices and Quantum Chaotic Scattering”. PhD thesis. Queen Mary University of London, 2017.

[Pap18] Vardan Papyan. “The Full Spectrum of Deepnet Hessians at Scale: Dynamics with SGD Training and Sample Size”. In: *arXiv preprint arXiv:1811.07062* (2018).

[PW17] Jeffrey Pennington and Pratik Worah. “Nonlinear random matrix theory for deep learning”. In: *Advances in Neural Information Processing Systems*. 2017, pp. 2637–2646.

[Rad+18] Alec Radford et al. *Improving language understanding by generative pre-training*. 2018.

[RMC15] Alec Radford, Luke Metz, and Soumith Chintala. “Unsupervised representation learning with deep convolutional generative adversarial networks”. In: *arXiv preprint arXiv:1511.06434* (2015).

[Ros+19a] Valentina Ros et al. “Complex Energy Landscapes in Spiked-Tensor and Simple Glassy Models: Ruggedness, Arrangements of Local Minima, and Phase Transitions”. In: *Phys. Rev. X* 9 (1 Jan. 2019), p. 011003. DOI: `10.1103/PhysRevX.9.011003`. URL: `https://link.aps.org/doi/10.1103/PhysRevX.9.011003`.

[Ros+19b] Valentina Ros et al. “Complex energy landscapes in spiked-tensor and simple glassy models: Ruggedness, arrangements of local minima, and phase transitions”. In: *Physical Review X* 9.1 (2019), p. 011003.

[Ver04] Jacobus Verbaarschot. “The supersymmetric method in random matrix theory and applications to QCD”. In: *AIP Conference Proceedings* (2004). ISSN: 0094-243X. DOI: `10.1063/1.1853204`. URL: `http://dx.doi.org/10.1063/1.1853204`.

[Zha+18] Han Zhang et al. “Self-Attention Generative Adversarial Networks”. In: *International Conference on Machine Learning*. 2018, pp. 7354–7363.

[Zhu+17] Jun-Yan Zhu et al. "Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks". In: *2017 IEEE International Conference on Computer Vision (ICCV)*. 2017, pp. 2242–2251.