Journal of Statistical Physics (2022) 186:29
https://doi.org/10.1007/s10955-022-02875-w

A Spin Glass Model for the Loss Surfaces of Generative Adversarial Networks

Nicholas P. Baskerville[1] · **Jonathan P. Keating**[2] · **Francesco Mezzadri**[1] · **Joseph Najnudel**[1]

Received: 16 June 2021 / Accepted: 3 January 2022 / Published online: 18 January 2022

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 which explains the greater difficulty of training GANs.

Keywords Random matrix theory · Deep learning · Neural networks · Generative adversarial networks · Spin glasses

1 Introduction

By making various modeling assumptions about standard multi-layer perceptron neural networks, [1] argued heuristically that the training loss surfaces of large networks could be modelled by a spherical multi-spin glass. Using theoretical results of [2], 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 [3] and recent observations have shown that the Hessians of

Communicated by Federico Ricci-Tersenghi.

✉ Nicholas P. Baskerville
n.p.baskerville@bristol.ac.uk

Jonathan P. Keating
jon.keating@maths.ox.ac.uk

Francesco Mezzadri
F.Mezzadri@bristol.ac.uk

Joseph Najnudel
joseph.najnudel@bristol.ac.uk

[1] School of Mathematics, University of Bristol, Fry Building, Bristol BS8 1UG, UK

[2] Mathematical Institute, University of Oxford, Oxford OX2 6GG, UK

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 [4–6], despite this being implied by the spin-glass model of [1]. 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 [7–9], 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 [1], 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 [10] with some explicitly added 'signal'; the simple explicitly non-linear model of [11]; the spiked tensor 'signal-in-noise' model of [12]. In a slightly different direction, [13] removed one of the main assumptions from the [1] 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 [14–16] and, more generally, with image reconstruction and other signal processing problems [17].

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) [18,19], recurrent networks for sequence data [20] and self-attention transformer networks for natural language [21,22]. 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 [23], and batch-norm has been considered to some extent by [24]. In this work, we propose a novel model for *generative adversarial networks* (GANs) [25] 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 [26–32] 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 [1,12,33,34] 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 [2,35,36] 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 [13] 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 [37] and the calculation also follows [38,39] 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 [1] 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 to justify the use of the Coulomb gas approximation. We note that our complexity results could be also be obtained in principle using the methods developed in [40], however our work was completed several months before this pre-print appeared. Our approach for computing the limiting spectral density may nevertheless be the simplest and would be used as input to the results of [40].

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 Sect. 2 we introduce the interacting spin glass model; in Sect. 3 we use a Kac-Rice formula to derive random matrix expressions for the asymptotic complexity of our model; in Sect. 4 we derive the limiting spectral density of the relevant random matrix ensemble; in Sect. 5 we use the Coulomb gas approximation to compute the asymptotic complexity, and legitimise its use by proving a concentration result; in Sect. 6 we derive some implications of our model for GAN training and compare to experimental results from real GANs; in Sect. 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 [1,13]. 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

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

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

where Θ_D, Θ_G are the parameters of the discriminator and generator respectively. With $z \sim \mathcal{N}(0, \sigma_z^2)$, $G(z)$ 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 task. To construct our model, we introduce two spin glasses:

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

$$\ell^{(G)}(\boldsymbol{w}^{(D)}, \boldsymbol{w}^{(G)}) = \sum_{i_1,\ldots,i_{p+q}=1}^{N_D+N_G} Z_{i_1,\ldots,i_{p+q}} \prod_{k=1}^{p+q} w_k \tag{3}$$

where $\boldsymbol{w}^T = (\boldsymbol{w}^{(D)^T}, \boldsymbol{w}^{(G)^T})$, all the $X_{i_1,\ldots,i_p}$ are i.i.d. $\mathcal{N}(0, 1)$ and $Z_{j_1,\ldots,j_{p+q}}$ are similarly i.i.d. $\mathcal{N}(0, 1)$. We then define the models for the discriminator and generator losses:

$$L^{(D)}(\boldsymbol{w}^{(D)}, \boldsymbol{w}^{(G)}) = \ell^{(D)}(\boldsymbol{w}^{(D)}) - \sigma_z \ell^{(G)}(\boldsymbol{w}^{(D)}, \boldsymbol{w}^{(G)}), \tag{4}$$

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

$\ell^{(D)}$ 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_i are surrogates for the training data (i.e. samples from $\mathbb{P}_{data}$) and the Z_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_j and include it as a constant multiplier in (4)–(5). In reality, we should like to keep Z_j as i.i.d. $\mathcal{N}(0, 1)$ but take X_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.

Remark 1 Our model is not supposed to have any direct relationship to GANs. Rather, we have used two spin glasses as models for high-dimensional random surfaces. The spin glasses are related by sharing some of their variables, namely the $\boldsymbol{w}^{(D)}$, just as the two training objectives in GANs share the discriminator weights. In prior work modeling neural network loss surfaces as spin glasses, the number of spins corresponds to the number of layers in the network, therefore we have chosen p spins for $\ell^{(D)}$ and $p + q$ for $\ell^{(G)}$, corresponding to p layers in the discriminator and q layers in the generator, but the generator is only ever seen in the losses composed with the discriminator. One could make other choices of $\ell^{(D)}$ and $\ell^{(G)}$ to couple the two glasses and we consider one such example in the appendix Sect. 1.

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

$$C_N = \left| \left\{ \boldsymbol{w}^{(D)} \in S^{N_D}, \boldsymbol{w}^{(G)} \in S^{N_G} \ : \ \nabla_D L^{(D)} = 0, \nabla_G L^{(G)} = 0, L^{(D)} \in B_D, L^{(G)} \in B_G \right\} \right| \tag{6}$$

for some Borel sets $B_D, B_G \subset \mathbb{R}$ and where ∇_D, ∇_G denote the Riemannian covariate derivatives on the hyperspheres 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

$$\begin{aligned} C_{N,k_D,k_G} = \Big| \Big\{ & \boldsymbol{w}^{(D)} \in S^{N_D}, \boldsymbol{w}^{(G)} \in S^{N_G} \ : \ \nabla_D L^{(D)} = 0, \nabla_G L^{(G)} = 0, L^{(D)} \in B_D, L^{(G)} \in B_G \\ & i\left(\nabla_D^2 L^{(D)}\right) = k_D, \ i\left(\nabla_G^2 L^{(G)}\right) = k_G \Big\} \Big|, \end{aligned} \tag{7}$$

where $i(M)$ is the index of M (i.e. the number of negative eigenvalues of M). We have chosen to consider the indices of the Hessians $\nabla_D^2 L^{(D)}$ and $\nabla_G^2 L^{(G)}$ separately, just as we chose to consider separately vanishing derivatives $\nabla_D L^{(D)}$ and $\nabla_G L^{(G)}$. We believe this choice best reflects the standard training loop of GANs, where each iteration updates the discriminator and generator parameters in separate steps.

To calculate the complexities, we follow the well-trodden route of Kac-Rice formulae as pioneered by [35,36]. For a fully rigorous treatment, we proceed as in [2,13].

Lemma 1

$$\begin{aligned} C_N = & \int_{S^{N_D} \times S^{N_G}} d\boldsymbol{w}^{(G)} d\boldsymbol{w}^{(D)} \ \varphi_{(\nabla_D L^{(D)}, \nabla_G L^{(G)})}(0) \\ & \mathbb{E}\left[\left| \det \begin{pmatrix} \nabla_D^2 L^{(D)} & \nabla_{GD} L^{(D)} \\ \nabla_{DG} L^{(G)} & \nabla_G^2 L^{(G)} \end{pmatrix} \right| \ | \ \nabla_G L^{(G)} \right. \\ & \left. = 0, \nabla_D L^{(D)} = 0 \right] \mathbb{1}\left\{ L^{(D)} \in B_D, L^{(G)} \in B_G \right\} \end{aligned} \tag{8}$$

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

and therefore

$$C_N = \int_{S^{N_D}\times S^{N_G}} d\boldsymbol{w}^{(G)} d\boldsymbol{w}^{(D)} \;\; \varphi_{(\nabla_D L^{(D)}, \nabla_G L^{(G)})}(0) \int_{B_D} dx_D \int_{B_G} dx_G \; \varphi_{L^{(D)}}(x_D)\varphi_{L^{(G)}}(x_G)$$
$$\mathbb{E}\Bigg[\left|\det\begin{pmatrix} \nabla_D^2 L^{(D)} & \nabla_{GD} L^{(D)} \\ \nabla_{DG} L^{(G)} & \nabla_G^2 L^{(G)} \end{pmatrix}\right| \;\mid\; \nabla_G L^{(G)}$$
$$= 0, \nabla_D L^{(D)} = 0, L^{(D)} = x_D, L^{(G)} = x_G\Bigg]. \tag{9}$$

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 $(\boldsymbol{w}^{(D)}, \boldsymbol{w}^{(G)})$.

Proof Routine application of a theorem of [41]. See appendix Sect. 1. □

With Lemma 1 in place, we can now establish the following Kac-Rice expression specialised to our model:

Lemma 2 *For $(N-2)\times(N-2)$ GOE matrix M and independent $(N_D-1)\times(N_D-1)$ GOE matrix M_1, define*

$$H(x, x_1) \overset{d}{=} bM + b_1 \begin{pmatrix} M_1 & 0 \\ 0 & 0 \end{pmatrix} - x - x_1 \begin{pmatrix} I_{N_D} & 0 \\ 0 & 0 \end{pmatrix}. \tag{10}$$

For $u_G, u_D \in \mathbb{R}$, define

$$B = \left\{(x, x_1) \in \mathbb{R}^2 \;:\; x \le \frac{1}{\sqrt{2}}(p+q)2^{p+q}u_G, \;\; x_1 \ge -(p+q)^{-1}2^{-(p+q)}px - \frac{p}{\sqrt{2}}u_D\right\}. \tag{11}$$

Define the constant

$$K_N = \omega_{\kappa N}\omega_{\kappa' N}(2(N-2))^{\frac{N-2}{2}}(2\pi)^{-\frac{N-2}{2}}\left(p+\sigma_z^2 2^{p+1}(p+q)\right)^{-\frac{\kappa N-1}{2}}\left(\sigma_z^2 2^{p+q}(p+q)\right)^{-\frac{\kappa' N-1}{2}} \tag{12}$$

where the variances are

$$s^2 = \frac{1}{2}\sigma_z^2(p+q)^2 2^{3(p+q)}, \quad s_1^2 = \frac{p^2}{2}. \tag{13}$$

and $\omega_N = \frac{2\pi^{N/2}}{\Gamma(N/2)}$ is the surface area of the N sphere. The expected complexity C_N is then

$$\mathbb{E}C_N = K_N \int_B \sqrt{\frac{N}{2\pi s^2}} e^{-\frac{N}{2s^2}x^2} dx \sqrt{\frac{N}{2\pi s_1^2}} e^{-\frac{N}{2s_1^2}x_1^2} dx_1 \, \mathbb{E}|\det H(x, x_1)|. \tag{14}$$

Proof Define the matrix

$$\tilde{H} = \begin{pmatrix} \nabla_D^2 L^{(D)} & \nabla_{GD} L^{(D)} \\ \nabla_{DG} L^{(G)} & \nabla_G^2 L^{(G)} \end{pmatrix}$$

appearing in the expression for C_N in Lemma 1. Note that $\tilde{H}$ takes the place of a Hessian (though it is not symmetric). We begin with the distribution of

$$\tilde{H} \;\mid\; \left\{(\ell^{(D)}, \ell^{(G)}) = (x_D, x_G), \;\; \left(\nabla_D \ell^{(D)}, \nabla \ell^{(G)}\right) = (0, 0)\right\}.$$

Note that the integrand in (14) is jointly spherically symmetric in both $\boldsymbol{w}^{(D)}$ and $\boldsymbol{w}^{(G)}$. It is therefore sufficient to consider $\tilde{H}$ in the region of a single point on each sphere. We choose the north poles and coordinate bases on both spheres in the region of their north poles. The remaining calculations are routine Gaussian manipulations which appear in the appendix Sect. 1. One finds

$$\begin{aligned}\tilde{H} \stackrel{d}{=} & \sqrt{2p(p-1)}\begin{pmatrix}\sqrt{N_D-1}M_2^{(D)} & 0\\ 0 & 0\end{pmatrix}\\ & + \sigma_z\sqrt{2^{p+q+1}(p+q)(p+q-1)}\begin{pmatrix}\sqrt{N_D-1}M_1^{(D)} & -2^{-1/2}G\\ 2^{-1/2}G^T & \sqrt{N_G-1}M^{(G)}\end{pmatrix}\\ & - \sigma_z(p+q)x_G 2^{p+q}\begin{pmatrix}-I_{N_D} & 0\\ 0 & I_{N_G}\end{pmatrix} - px_D\begin{pmatrix}I_{N_D} & 0\\ 0 & 0\end{pmatrix}\end{aligned} \tag{15}$$

where $M_1^{(D)}$, $M_2^{(D)}$ are independent GOE^{N_D-1} matrices, $M^{(G)}$ is an independent GOE^{N_G-1} matrix and G is an independent $(N_D-1)\times(N_G-1)$ Ginibre matrix. Note that the dimensions are N_D-1 and N_G-1 rather N_D and N_G. This is simply because the hypersphere S^{N_D} is an N_D-1 dimensional manifold, and similarly S^{N_G}.

We can simplify by summing independent Gaussians to obtain

$$\begin{aligned}\tilde{H} = & \begin{pmatrix}\sigma_D\sqrt{N_D-1}M^{(D)} & -2^{-1/2}\sigma_G G,\\ 2^{-1/2}\sigma_G G^T & \sigma_G\sqrt{N_G-1}M^{(G)}\end{pmatrix}\\ & - \sigma_z(p+q)x_G 2^{p+q}\begin{pmatrix}-I_{N_D} & 0\\ 0 & I_{N_G}\end{pmatrix} - px_D\begin{pmatrix}I_{N_D} & 0\\ 0 & 0\end{pmatrix}\end{aligned} \tag{16}$$

where

$$\sigma_G = \sigma_z\sqrt{2^{p+q+1}(p+q)(p+q-1)} \tag{17}$$

$$\sigma_D = \sqrt{\sigma_G^2 + 2p(p-1)} \tag{18}$$

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

There is an alternative reformulation of $\tilde{H}$ that will also be useful. Indeed, because $M_{1,2}^{(D)} \stackrel{d}{=} -M_{1,2}^{(D)}$, let us write $\tilde{H}$ as

$$\begin{aligned}\tilde{H} = & \sigma_z J\left(\sqrt{2^{p+q+1}(p+q)(p+q-1)(N_D+N_G-2)}M_1 - (p+q)x_G 2^{p+q} I\right)\\ & + \left(\sqrt{2p(p-1)(N_D-1)}\begin{pmatrix}M_2 & 0\\ 0 & 0\end{pmatrix} - px_D\begin{pmatrix}I_{N_D} & 0\\ 0 & 0\end{pmatrix}\right)\\ \stackrel{d}{=} & J\Bigg[\sigma_z\sqrt{2^{p+q+1}(p+q)(p+q-1)(N_D+N_G-2)}M_1 - \sigma_z(p+q)x_G 2^{p+q} I\\ & + \sqrt{2p(p-1)(N_D-1)}\begin{pmatrix}M_2 & 0\\ 0 & 0\end{pmatrix} + px_D\begin{pmatrix}I_{N_D} & 0\\ 0 & 0\end{pmatrix}\Bigg]\end{aligned} \tag{19}$$

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 an independent GOE matrix of size N_D-1 and

$$J = \begin{pmatrix}-I_{N_D} & 0\\ 0 & I_{N_G}\end{pmatrix}. \tag{20}$$

If follows that

$$|\det \tilde{H}| \stackrel{d}{=} \left| \det \left[\sigma_z \sqrt{2^{p+q+1}(p+q)(p+q-1)(N_D+N_G-2)} M_1 - \sigma_z (p+q) x_G 2^{p+q} I \right.\right.$$
$$\left.\left. + \sqrt{2p(p-1)(N_D-1)} \begin{pmatrix} M_2 & 0 \\ 0 & 0 \end{pmatrix} + p x_D \begin{pmatrix} I_{N_D} & 0 \\ 0 & 0 \end{pmatrix} \right] \right|. \tag{21}$$

Now define the constants

$$b = \sqrt{2^{p+q}(p+q)(p+q-1)}\sigma_z, \qquad b_1 = \sqrt{p(p-1)\kappa} \tag{22}$$

$$x = \frac{\sigma_z (p+q) 2^{p+q}}{\sqrt{N}} x_G, \qquad x_1 = -\frac{p}{\sqrt{N}} x_D, \tag{23}$$

and then we arrive at

$$|\det \tilde{H}| \stackrel{d}{=} (2(N-2))^{\frac{N-2}{2}} |\det H(x, x_1)|. \tag{24}$$

The variances of $L^{(D)}$ and $L^{(G)}$ derive from those of $\ell^{(G)}$, $\ell^{(D)}$ computed in appendix Sect. 1 (see (139), (143)):

$$Var(\ell^{(D)}) = 1, \quad Var(\ell^{(G)}) = 2^{p+q}.$$

Similarly the density $\varphi_{(\nabla_D L^{(D)}, \nabla_G L^{(G)})}$ is found in (155):

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

We have now collected all the inputs required for Lemma 1. The domain of integration B arises from the constraints $L^{(D)} \in (-\infty, \sqrt{N} u_D)$ and $L^{(G)} \in (-\infty, \sqrt{N} u_G)$ and the re-scaled variables (23). This completes the proof. □

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

Lemma 3 *As $N \to \infty$,*

$$K_N \sim 2^{\frac{N}{2}} \pi^{N/2} \left(\kappa^{\kappa} \kappa'^{\kappa'}\right)^{-N/2} \sqrt{\kappa\kappa'} \left(p + \sigma_z^2 2^{p+1}(p+q)\right)^{-\frac{\kappa N-1}{2}} \left(\sigma_z^2 2^{p+q}(p+q)\right)^{-\frac{\kappa' N-1}{2}} \tag{25}$$

Proof By Stirling's formula

$$K_N \sim 4\pi^N \left(\frac{4\pi}{\kappa N}\right)^{-1/2} \left(\frac{4\pi}{\kappa' N}\right)^{-1/2} \left(\frac{\kappa N}{2e}\right)^{-\kappa N/2} \left(\frac{\kappa' N}{2e}\right)^{-\kappa' N/2} (2(N-2))^{\frac{N-2}{2}} (2\pi)^{-\frac{N-2}{2}}$$
$$\left(p + \sigma_z^2 2^{p+1}(p+q)\right)^{-\frac{\kappa N-1}{2}} \left(\sigma_z^2 2^{p+q}(p+q)\right)^{-\frac{\kappa' N-1}{2}}$$
$$\sim 2^{\frac{N}{2}} \pi^{N/2} \left(\kappa^{\kappa} \kappa'^{\kappa'}\right)^{-N/2} \sqrt{\kappa\kappa'} \left(p + \sigma_z^2 2^{p+1}(p+q)\right)^{-\frac{\kappa N-1}{2}} \left(\sigma_z^2 2^{p+q}(p+q)\right)^{-\frac{\kappa' N-1}{2}} \tag{26}$$

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

$$H' = bM + b_1 \begin{pmatrix} M_1 & 0 \\ 0 & 0 \end{pmatrix} - x_1 \begin{pmatrix} I & 0 \\ 0 & 0 \end{pmatrix}, \tag{27}$$

where, note, $H' = H + xI$ is just a shifted version of H as defined in Lemma 2. 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} [37]:

$$\langle G(z) \rangle = \frac{1}{N} \frac{\partial}{\partial J} \bigg|_{J=0} Z(J) \tag{28}$$

$$Z(J) := \mathbb{E}_{H'} \frac{\det(z - H' + J)}{\det(z - H')}. \tag{29}$$

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

$$\rho_{eq}(z) = \frac{1}{\pi} \lim_{\epsilon \to 0} \Im \langle G(z + i\epsilon) \rangle. \tag{30}$$

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:

$$\int \prod_{i=1}^{N} \frac{d\phi_i d\phi_i^*}{2\pi} e^{-i\phi^\dagger M \phi} = \frac{1}{\det M}. \tag{31}$$

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

$$\int \frac{1}{-i} \prod_{i=1}^{N} d\chi_i d\chi_i^* e^{-i\chi^\dagger M \chi} = \det M. \tag{32}$$

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

$$\chi_i \chi_j = -\chi_j \chi_i, \quad \forall i, j \tag{33}$$

and χ_i^* are separate objects, with the complex conjugation unary operator * defined so that $\left(\chi_i^*\right)^* = -\chi_i^*$, 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

$$\int d\chi_i = 0, \qquad \int d\chi_i \ \chi_j = \delta_{ij}. \tag{34}$$

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

$$e^{\chi_i} = 1 + \chi_i + \frac{1}{2}\chi_i^2 + \ldots = 1 + \chi_i \tag{35}$$

where the termination of the series follows from $\chi_i^2 = 0 \ \forall i$, which is an immediate consequence of (33). From this it is apparent that (34), along with (33), is sufficient to define Berezin integration over arbitrary functions of arbitrary combinations of Grassmann variables. Finally we establish our notation for supersymmetric (or *graded*) traces of supermatrices. We will encounter supermatrices of the form

$$M = \begin{pmatrix} A & B \\ C & D \end{pmatrix}$$

where A, D are square block matrices of commuting variables and B, C are rectangular block matrices of Grassmann variables. In this case, the graded trace is given by $\mathrm{trg} M = \mathrm{Tr} A - \mathrm{Tr} D$. We refer the reader to [42] for a full introduction to supersymmetric methods.

Using the integral results (31), (32) we can then write

$$\frac{\det(z - H' + J)}{\det(z - H')} = \int d\Psi \exp\left\{-i\phi^\dagger (z - H')\phi - i\chi^\dagger (z + J - H')\chi\right\} \tag{36}$$

where the measure is

$$d\Psi = 1/-i(2\pi)^N \prod_{t=1}^{2} d\phi[t] d\phi^*[t] d\chi[t] d\chi^*[t], \tag{37}$$

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 [38]:

$$\phi = \begin{pmatrix} \phi[1] \\ \phi[2] \end{pmatrix}. \tag{38}$$

We then split the quadratic form expressions in (36)

$$\begin{aligned} &-\phi^\dagger (z - H')\phi - \chi^\dagger (z + J - H')\chi \\ &= -\phi[1]^\dagger \left(x_1 - b_1 M_1\right)\phi[1] - \phi^\dagger (z - bM)\phi - \chi[1]^\dagger \left(x_1 - b_1 M_1\right)\chi[1] - \chi^\dagger (z + J - bM)\chi. \end{aligned} \tag{39}$$

Taking the GOE averages is now simple [37,43]:

$$\mathbb{E}_M \exp\left\{-ib\phi^\dagger M\phi - ib\chi^\dagger M\chi\right\} = \exp\left\{-\frac{b^2}{4N}\mathrm{trg} Q^2\right\}, \tag{40}$$

$$\mathbb{E}_M \exp\left\{-ib_1\phi[1]^\dagger M_1\phi[1] - ib_1\chi[1]^\dagger M_1\chi[1]\right\} = \exp\left\{-\frac{b_1^2}{4\kappa N}\mathrm{trg} Q[1]^2\right\}, \tag{41}$$

where the supersymmetric matrices are given by

$$Q = \begin{pmatrix} \phi^\dagger\phi & \phi^\dagger\chi \\ \chi^\dagger\phi & \chi^\dagger\chi \end{pmatrix}, \quad Q[1] = \begin{pmatrix} \phi[1]^\dagger\phi[1] & \phi[1]^\dagger\chi[1] \\ \chi[1]^\dagger\phi[1] & \chi[1]^\dagger\chi[1] \end{pmatrix}. \tag{42}$$

Introducing the tensor notation

$$\psi = \phi \otimes \begin{pmatrix} 1 \\ 0 \end{pmatrix} + \chi \otimes \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \quad \psi[1] = \phi[1] \otimes \begin{pmatrix} 1 \\ 0 \end{pmatrix} + \chi[1] \otimes \begin{pmatrix} 0 \\ 1 \end{pmatrix} \tag{43}$$

and

$$\zeta = \begin{pmatrix} z & 0 \\ 0 & z + J \end{pmatrix} \tag{44}$$

we can compactly write

$$Z(J) = \int d\Psi \exp\left\{-\frac{b^2}{4N}\mathrm{trg}Q^2 - \frac{b_1^2}{4\kappa N}\mathrm{trg}Q[1]^2 - i\psi[1]^\dagger\psi[1]x_1 - i\psi^\dagger\zeta\psi\right\}. \quad (45)$$

We now perform two Hubbard-Stratonovich transformations [37]

$$Z(J) = \int d\Psi d\sigma d\sigma[1] \exp\left\{-\frac{N}{b^2}\mathrm{trg}\sigma^2 - \frac{\kappa N}{b_1^2}\mathrm{trg}\sigma[1]^2 - i\psi[1]^\dagger(x_1+\sigma[1])\psi[1] - i\psi^\dagger(\sigma+\zeta)\psi\right\}, \quad (46)$$

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

$$\sigma = \begin{pmatrix} \sigma_{BB} & \sigma_{BF} \\ \sigma_{FB} & i\sigma_{FF} \end{pmatrix}, \quad \sigma[1] = \begin{pmatrix} \sigma_{BB}[1] & \sigma_{BF}[1] \\ \sigma_{FB}[1] & i\sigma_{FF}[1] \end{pmatrix} \quad (47)$$

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

$$\begin{aligned} Z(J) &= \int d\Psi d\sigma d\sigma[1] \exp\left\{-\frac{N}{b^2}\mathrm{trg}\sigma^2 - \frac{\kappa N}{b_1^2}\mathrm{trg}\sigma[1]^2 \right. \\ &\quad \left. -i\psi[1]^\dagger(x_1+\zeta+\sigma+\sigma[1])\psi[1] - i\psi[2]^\dagger(\sigma+\zeta)\psi[2]\right\} \\ &= \int d\sigma d\sigma[1] \exp\left\{-\frac{N}{b^2}\mathrm{trg}\sigma^2 - \frac{\kappa N}{b_1^2}\mathrm{trg}\sigma[1]^2 \right. \\ &\quad \left. -\kappa N\mathrm{trg}\log(x_1+\zeta+\sigma+\sigma[1]) - \kappa' N\mathrm{trg}\log(\sigma+\zeta)\right\} \\ &= \int d\sigma d\sigma[1] \exp\left\{-\frac{N}{b^2}\mathrm{trg}(\sigma-\zeta)^2 - \frac{\kappa N}{b_1^2}\mathrm{trg}\sigma[1]^2 \right. \\ &\quad \left. -\kappa N\mathrm{trg}\log(x_1+\sigma+\sigma[1]) - \kappa' N\mathrm{trg}\log\sigma\right\}. \end{aligned} \quad (48)$$

Recalling the definition of ζ, we have

$$\mathrm{trg}(\sigma-\zeta)^2 = (\sigma_{BB}-z)^2 - (i\sigma_{FF}-z-J)^2 \quad (49)$$

and so one immediately obtains

$$\begin{aligned} \frac{1}{N}\frac{\partial}{\partial J}\bigg|_{J=0} Z(J) &= \frac{2}{b^2}\int d\sigma d\sigma[1](z-i\sigma_{FF}) \exp\left\{-\frac{N}{b^2}\mathrm{trg}(\sigma-z)^2 - \frac{\kappa N}{b_1^2}\mathrm{trg}\sigma[1]^2 \right. \\ &\quad \left. -\kappa N\mathrm{trg}\log(x_1+\sigma+\sigma[1]) - \kappa' N\mathrm{trg}\log\sigma\right\} \\ &= \frac{2}{b^2}\int d\sigma d\sigma[1](z-i\sigma_{FF}) \exp\left\{-\frac{N}{b^2}\mathrm{trg}\sigma^2 - \frac{\kappa N}{b_1^2}\mathrm{trg}\sigma[1]^2 \right. \\ &\quad \left. -\kappa N\mathrm{trg}\log(x_1+z+\sigma+\sigma[1]) - \kappa' N\mathrm{trg}\log(z+\sigma)\right\} \end{aligned} \quad (50)$$

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 (50). 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 (50) 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

$$\frac{2}{b_1^2}\sigma_{BB}[1] + \frac{1}{\sigma_{BB}[1] + \sigma_{BB} + x_1 + z} = 0 \tag{51}$$

$$\frac{2}{b^2}\sigma_{BB} + \frac{\kappa}{\sigma_{BB}[1] + \sigma_{BB} + x_1 + z} + \frac{\kappa'}{\sigma_{BB} + x} = 0 \tag{52}$$

$$\frac{2}{b_1^2}\sigma_{FF}[1] - \frac{1}{\sigma_{FF}[1] + \sigma_{FF} - ix_1 - iz} = 0 \tag{53}$$

$$\frac{2}{b^2}\sigma_{FF} - \frac{\kappa}{\sigma_{FF}[1] + \sigma_{FF} - ix_1 - iz} - \frac{\kappa'}{\sigma_{FF} - iz} = 0. \tag{54}$$

(53) and (54) combine to give an explicit expression for $\sigma_{FF}[1]$:

$$\sigma_{FF}[1] = \frac{b_1^2}{2\kappa}\left(\frac{2}{b^2}\sigma_{FF} - \kappa'\left(\sigma_{FF} - iz\right)^{-1}\right). \tag{55}$$

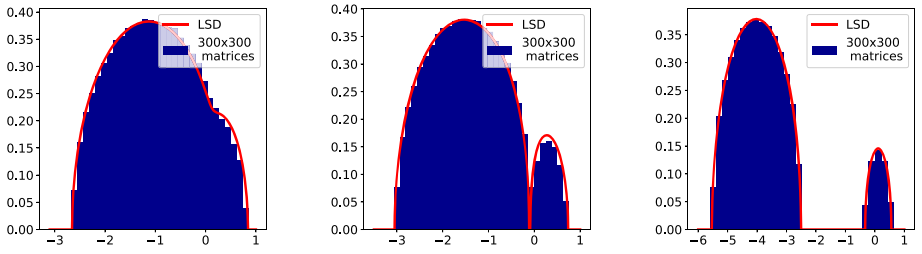
With a view to simplifying the numerical solution of the coming quartic, we define $t = i(\sigma_{FF} - iz)$ and then a line of manipulation with (54) and (55) gives

$$\left(t^2 - zt - \kappa' b^2\right)\left((1 + \kappa^{-1}b^{-2}b_1^2)t^2 - \left(\kappa^{-1}b_1^2 b^{-2}z - x_1\right)t - \kappa'\kappa^{-1}b_1^2\right) + b^2\kappa t^2 = 0. \tag{56}$$

By solving (56) 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 (30) and (50):

$$\rho_{eq}(z) = \max_i \left\{-\frac{2}{b^2\pi}\Im t_i(z)\right\}. \tag{57}$$

Plots generated using (57) and eigendecompositions of matrices sampled from the distribution of H' are given in Fig. 1 and show good agreement between the two. Note the three different forms: single component support, two component support and the transition point between the two, according to the various parameters. In these plots, the larger lobes on the left correspond to the upper left block, which is much larger than the lower-right block (since $\kappa = 0.9$ here). One can see this by considering large x_1, for which there must be a body of eigenvalues in the region of $-x_1$ owing to the upper left block. Since x_1 only features in the upper-left block, not all of the eigenvalues can be located around $-x_1$, and the remainder are found in the other lobe of the density which is around 0 in Fig. 1.



(a) Merged **(b)** Touching **(c)** Separate

Fig. 1 Example spectra of H' showing empirical spectra from 100 300×300 matrices and the corresponding LSDs computed from (56). 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

$$H' = bM + b_1 \begin{pmatrix} M_1 & 0 \\ 0 & 0 \end{pmatrix} - x_1 \begin{pmatrix} I & 0 \\ 0 & 0 \end{pmatrix}, \quad M \sim GOE^N, \; M_1 \sim GOE^{\kappa N}. \tag{58}$$

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

$$\mathbb{E}|\det(H' - x)| \approx \exp\left\{ N \int \log|z - x| d\mu_{eq}(z) \right\}. \tag{59}$$

We have access to the density of μ_{eq} pointwise (in x and x_1) numerically, and so (59) is a matter of one-dimensional quadrature. Recalling (14), we then have

$$\begin{aligned} \mathbb{E}C_N &\approx K'_N \iint_B dx dx_1 \; \exp\left\{ -(N-2)\left(\frac{1}{2s^2}x^2 + \frac{1}{2s_1^2}(x_1)^2 - \int \log|z - x| d\mu_{eq}(z) \right) \right\} \\ &\equiv K'_N \iint_B dx dx_1 \; e^{-(N-2)\Phi(x, x_1)} \end{aligned} \tag{60}$$

where

$$K'_N = K_N \sqrt{\frac{N-2}{2\pi s_1^2}} \sqrt{\frac{N-2}{2\pi s^2}}. \tag{61}$$

Due to Lemma 3, the constant term has asymptotic form

$$\begin{aligned} \frac{1}{N}\log K'_N \sim \frac{1}{2}\log 2 + \frac{1}{2}\log \pi - \frac{\kappa}{2}\log\left(p + \sigma_z^2 2^{p+q}(p+q)\right) \\ - \frac{\kappa'}{2}\log\left(\sigma_z^2 (p+q) 2^{p+q}\right) - \frac{\kappa}{2}\log\kappa - \frac{\kappa'}{2}\log\kappa' \equiv K \end{aligned} \tag{62}$$

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

$$\lim \frac{1}{N}\log \mathbb{E}C_N = \Theta\left(u_D, u_G\right) \tag{63}$$

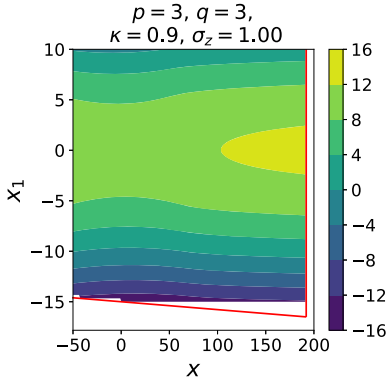


Fig. 2 Φ for $p = q = 3, \sigma_z = 1, \kappa = 0.9$. Red lines show the boundary of the integration region B

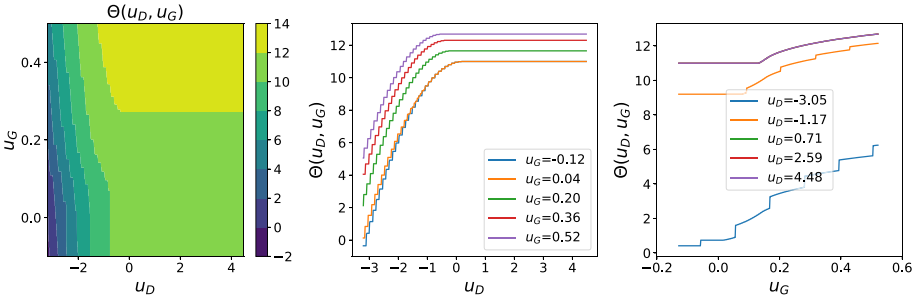


Fig. 3 Θ and its cross-sections, fixing separately u_D and u_G. Here $p = q = 3, \sigma_z = 1, \kappa = 0.9$

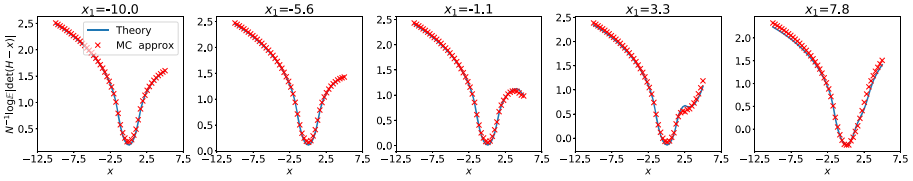


Fig. 4 Comparison of (59) and (65), verifying the Coulomb gas approximation numerically. Here $p = q = 3, \sigma_z = 1, \kappa = 0.9$. Sampled matrices for MC approximation are dimension $N = 50$, and $n = 50$ MC samples have been used

and we have

$$\Theta\left(u_D, u_G\right) = K - \min_B \Phi. \tag{64}$$

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 Figs. 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.

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

$$\mathbb{E}|\det(H' - x)| \approx \frac{1}{n}\sum_{i=1}^{n}\prod_{j=1}^{N}|\lambda_j^{(i)} - x|, \tag{65}$$

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 Fig. 4. Note the strong agreement even at such modest N, however to rigorously substantiate the Coulomb gas approximation in (59), we must prove a concentration result.

Lemma 4 *Let $(H_N)_{N=1}^{\infty}$ be a sequence of random matrices, where for each N*

$$H_N \stackrel{d}{=} bM + b_1\begin{pmatrix} M_1 & 0 \\ 0 & 0 \end{pmatrix} - x_1\begin{pmatrix} I & 0 \\ 0 & 0 \end{pmatrix} \tag{66}$$

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 $(x, x_1) \in \mathbb{R}^2$

$$\mathbb{E}|\det(H_N - xI)| = \exp\left\{N(1+o(1))\int \log|z-x|d\mu_{eq}(z)\right\} \tag{67}$$

as $N \to \infty$.

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

$$\begin{aligned}
&\int \log|z-x|d\mu_N(z) \\
&= \int \log|z-x|\mathbb{1}\{|x-z| \geq e^{\beta}\}d\mu_N(z) + \int \log|z-x|\mathbb{1}\{\log|x-z| < \beta\}d\mu_N(z) \\
&\leq \int \log|z-x|\mathbb{1}\{|x-z| \geq e^{\beta}\}d\mu_N(z) + \int \min(\log|x-z|, \beta)d\mu_N(z).
\end{aligned} \tag{68}$$

Take also any $\alpha > 0$, then trivially

$$\int \min(\log|x-z|, \beta)d\mu_N(z) \leq \int \max(-\alpha, \min(\log|x-z|, \beta))d\mu_N(z). \tag{69}$$

Overall we have, for any $\alpha, \beta > 0$,

$$\begin{aligned}
&\exp\left\{N\int \log|z-x|d\mu_N(z)\right\} \\
&\leq \exp\left\{N\int \log|z-x|\mathbb{1}\{|x-z| \geq e^{\beta}\}d\mu_N(z)\right\} \\
&\exp\left\{N\int \max(-\alpha, \min(\log|x-z|, \beta))d\mu_N(z)\right\}.
\end{aligned} \tag{70}$$

Thence an application of Hölder's inequality gives

$$\begin{aligned}
\mathbb{E}|\det(H_N - xI)| &= \mathbb{E}\left[\exp\left\{N\int \log|z-x|d\mu_N(z)\right\}\right] \\
&\leq \underbrace{\left(\mathbb{E}\left[\exp\left\{2N\int \max(-\alpha, \min(\log|x-z|, \beta))\, d\mu_N(z)\right\}\right]\right)^{1/2}}_{A_N} \\
&\quad \underbrace{\left(\mathbb{E}\left[\exp\left\{2N\int \log|x-z|\mathbb{1}\{|x-z| \geq e^{\beta}\}d\mu_N(z)\right\}\right]\right)^{1/2}}_{B_N}.
\end{aligned} \tag{71}$$

Considering B_N, we have

$$\log|x-z|\mathbb{1}\{|x-z|\geq e^{\beta}\}\leq|x-z|^{1/2}\mathbb{1}\{|x-z|\geq e^{\beta}\}\leq e^{-\beta/2}|x-z| \tag{72}$$

and so

$$\begin{aligned}\mathbb{E}\left[\exp\left\{2N\int\log|x-z|\mathbb{1}\{|x-z|\geq e^{\beta}\}\right\}\right]&\leq\mathbb{E}\left[\exp\left\{2Ne^{-\beta/2}\frac{\mathrm{Tr}|H_N-xI|}{N}\right\}\right]\\&=\mathbb{E}\left[\exp\left\{2e^{-\beta/2}\mathrm{Tr}|H_N-xI|\right\}\right].\end{aligned} \tag{73}$$

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

$$|H_N-x|_{ij}\leq|x|+|x_1|+\mathcal{O}(N^{-1/2})|X_{ij}| \tag{74}$$

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

$$\mathbb{E}\left[\exp\left\{2e^{-\frac{\beta}{2}}\mathrm{Tr}|H_N-xI|\right\}\right]\leq e^{2e^{-\frac{\beta}{2}}N(|x|+|x_1|)}\mathbb{E}_{X\sim\mathcal{N}(0,1)}e^{2e^{-\frac{\beta}{2}}\mathcal{O}(N^{1/2})|X|}. \tag{75}$$

Elementary calculations give

$$\mathbb{E}_{X\sim\mathcal{N}(0,1)}e^{c|X|}\leq\frac{1}{2}\left(e^{-c^2}+e^{c^2}\right)\leq e^{c^2} \tag{76}$$

and so

$$\begin{aligned}\mathbb{E}\left[\exp\left\{2e^{-\frac{\beta}{2}}\mathrm{Tr}|H_N-xI|\right\}\right]&\leq e^{2e^{-\frac{\beta}{2}}N(|x|+|x_1|)}e^{4e^{-\beta}\mathcal{O}(N)}\\&=\exp\left\{2N\left(e^{-\frac{\beta}{2}}(|x|+|x_1|)+e^{-\beta}\mathcal{O}(1)\right)\right\}\end{aligned} \tag{77}$$

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

Considering A_N, it is sufficient now to show

$$\mathbb{E}\left[\exp\left\{2N\int f(z)d\mu_N(z)\right\}\right]=\exp\left\{2N\left(\int f(z)d\mu_{eq}(z)+o(1)\right)\right\} \tag{78}$$

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

$$\begin{aligned}&\mathbb{E}\left[\exp\left\{2N\int f(z)d\mu_N(z)\right\}\right]\\&\leq\exp\left\{2N\left(\int f(z)d\mu_{eq}(z)+\epsilon\right)\right\}+e^{2N\|f\|_\infty}\mathbb{P}\left(\int f(z)d\mu_N(z)\geq\int f(z)d\mu_{eq}(z)+\epsilon\right).\end{aligned} \tag{79}$$

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 [44]. The constant, c, in the inequality is independent of N, x, x_1, so we need not compute it exactly. The theorem from [44] then gives

$$\mathbb{P}\left(\int f(z)d\mu_N(z)\geq\int f(z)d\mu_{eq}(z)+\epsilon\right)\leq\exp\left\{-\frac{N^2}{8c}\epsilon^2\right\}. \tag{80}$$

We have shown

$$\begin{aligned}\mathbb{E}|\det(H_N - xI)| &\leq A_N B_N \leq \exp\left\{N(1+o(1))\left(\int f(z)d\mu_{eq}(z)\right)\right\} \\ &\leq \exp\left\{N(1+o(1))\left(\int \log|x-z|d\mu_{eq}(z)\right)\right\}. \end{aligned} \tag{81}$$

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

$$\begin{aligned}\mathbb{E}|\det(H_N - x)| &\geq \exp\left(N\mathbb{E}\left[\int \log|z-x|d\mu_N(z)\right]\right) \\ &\geq \exp\left(N\mathbb{E}\left[\int \max(-\alpha, \log|z-x|)\, d\mu_N(z)\right]\right) \\ &\quad \exp\left(N\mathbb{E}\left[\int \log|z-x|\mathbb{1}\{|z-x| \leq e^{-\alpha}\}d\mu_N(z)\right]\right) \\ &\geq \exp\left(N\mathbb{E}\left[\int \min(\beta, \max(-\alpha, \log|z-x|))\, d\mu_N(z)\right]\right) \\ &\quad \exp\left(N\mathbb{E}\left[\int \log|z-x|\mathbb{1}\{|z-x| \leq e^{-\alpha}\}d\mu_N(z)\right]\right) \end{aligned} \tag{82}$$

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

$$\exp\left(N\mathbb{E}\left[\int \min(\beta, \max(-\alpha, \log|z-x|))\, d\mu_N(z)\right]\right) \geq \exp\left\{N\left(\int \log|x-z|d\mu_{eq}(z) + o(1)\right)\right\} \tag{83}$$

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 (82) converges to zero uniformly in N in the limit $\alpha \to \infty$.

By (30), it is sufficient to consider $\langle G_N(z)\rangle$, which is computed via (50). Let us define the function Ψ so that

$$\langle G_N(z)\rangle = \frac{2}{b^2}\int d\sigma d\sigma[1]\,(z - i\sigma_{FF})\, e^{-N\Psi(\sigma,\sigma[1])}. \tag{84}$$

Henceforth, $\sigma^*_{FF}, \sigma_{FF}[1]^*, \sigma^*_{BB}, \sigma_{BB}[1]^*$ are the solution to the saddle point equations (51–54) and $\tilde{\sigma}_{FF}, \tilde{\sigma}_{FF}[1], \tilde{\sigma}_{BB}, \tilde{\sigma}_{BB}[1]$ are integration variables. Around the saddle point

$$z - i\sigma_{FF} = z - i\sigma^*_{FF} - iN^{-\frac{1}{r}}\tilde{\sigma}_{FF} \tag{85}$$

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,

$$\int d\tilde{\sigma}d\tilde{\sigma}[1]\tilde{\sigma}_{FF}e^{-N\tilde{\boldsymbol{\sigma}}^T\nabla^2\Psi^*\tilde{\boldsymbol{\sigma}}} = 0 \tag{86}$$

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

$$\langle G_N(z)\rangle = G(z) - \frac{2i}{b^2 N^{5/3}} \underbrace{\int_{-\infty}^{\infty} d\tilde{\sigma}_{BB} d\tilde{\sigma}_{FF} \tilde{\sigma}_{FF} e^{-\frac{1}{6}\tilde{\sigma}^i\tilde{\sigma}^j\tilde{\sigma}^k\partial_{ijk}\Psi^*}}_{E(z;x_1)} + \text{exponentially smaller terms.} \tag{87}$$

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

$$\begin{aligned}(\nabla\Psi)_B &= \begin{pmatrix} \frac{2}{b^2}\sigma_{BB} - \kappa\left(\sigma_{BB}+\sigma_{BB}[1]+z+x_1\right)^{-1} - \kappa'\left(\sigma_{BB}+z\right)^{-1} \\ \frac{2}{b_1^2}\sigma_{BB}[1] - \left(\sigma_{BB}+\sigma_{BB}[1]+z+x_1\right)^{-1} \end{pmatrix} \\ &\implies \left(\nabla^2\Psi\right)_B \\ &= \begin{pmatrix} \kappa\left(\sigma_{BB}+\sigma_{BB}[1]+z+x_1\right)^{-2} + \kappa'\left(\sigma_{BB}+z\right)^{-2} & \kappa\left(\sigma_{BB}+\sigma_{BB}[1]+z+x_1\right)^{-2} \\ \left(\sigma_{BB}+\sigma_{BB}[1]+z+x_1\right)^{-2} & \left(\sigma_{BB}+\sigma_{BB}[1]+z+x_1\right)^{-2} \end{pmatrix} \qquad (88) \\ &\implies \left(\nabla^3\Psi\right)^*_B = \left(\begin{pmatrix} A_B\kappa + B_B\kappa' & A_B\kappa \\ A_B & A_B \end{pmatrix}, A_B \begin{pmatrix} \kappa & \kappa \\ 1 & 1 \end{pmatrix}\right), \qquad (89)\end{aligned}$$

where

$$A_B = -\frac{2}{\left(\sigma^*_{BB}+\sigma^*_{BB}[1]+z+x_1\right)^3}, \quad B_B = -\frac{2}{\left(\sigma^*_{BB}+z\right)^3}. \tag{90}$$

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

$$A_F = -\frac{2}{\left(\sigma^*_{FF}+\sigma^*_{FF}[1]-iz-ix_1\right)^3}, \quad B_F = -\frac{2}{\left(\sigma^*_{FF}-iz\right)^3}. \tag{91}$$

By the saddle point equations (51)–(54) we have

$$A_B = 2\left(\sigma_{BB}[1]^*\right)^3, \quad B_B = \frac{2}{(\kappa')^3}\left(\frac{2\kappa}{b_1^2}\sigma_{BB}[1]^* - \frac{2}{b^2}\sigma^*_{BB}\right)^3 \tag{92}$$

$$A_F = 2\left(\sigma_{FF}[1]^*\right)^3, \quad B_F = \frac{2}{(\kappa')^3}\left(\frac{2\kappa}{b_1^2}\sigma_{FF}[1]^* - \frac{2}{b^2}\sigma^*_{FF}\right)^3. \tag{93}$$

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

$$\begin{aligned}\left(\tilde{\sigma}^i\tilde{\sigma}^j\tilde{\sigma}^k\partial_{ijk}\Phi^*\right)_B &= \left(A_B\kappa + B_B\kappa'\right)\xi_1^3 + A_B(2\kappa+1)\xi_1^2\xi_2[1] + A_B(\kappa+2)\xi_1\xi_2^2 + A_B\xi_2^3 \\ &= A_B\left[\xi_2^3 + (2\kappa+1)\xi_2\xi_1^2 + (2+\kappa)\xi_1\xi_2^2 + C\xi_1^3\right] + \left(B_B\kappa' + A_B\kappa - CA_B\right)\xi_1^3\end{aligned} \tag{94}$$

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

$$\left(\tilde{\sigma}^i\tilde{\sigma}^j\tilde{\sigma}^k\partial_{ijk}\Phi^*\right)_B = A_B\left(\xi_1'+\xi_2\right)^3 + \left(B_B\kappa' + A_B\kappa - CA_B\right)a_1^3(\xi_1')^3 \equiv A_B\eta^3 + D_B\xi^3 \tag{95}$$

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

$$E\,(z;x_1) \propto \left(\int_0^\infty d\xi\ \xi \int_\xi^\infty d\eta\ e^{A_F\eta^3+D_F\xi^3}\right)\left(\int_0^\infty d\xi \int_\xi^\infty d\eta\ e^{A_B\eta^3+D_B\xi^3}\right) \tag{96}$$

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

$$\begin{aligned}
|E\,(z;x_1)\,| &\le \left|\int_0^\infty d\xi\ \xi \int_\xi^\infty d\eta\ e^{A_F\eta^3+D_F\xi^3}\right| \cdot \left|\int_0^\infty d\xi \int_\xi^\infty d\eta\ e^{A_B\eta^3+D_B\xi^3}\right| \\
&\le \int_0^\infty d\xi\ \xi \int_\xi^\infty d\eta\ |e^{A_F\eta^3+D_F\xi^3}| \cdot \int_0^\infty d\xi \int_\xi^\infty d\eta\ |e^{A_B\eta^3+D_B\xi^3}| \\
&\le \int_0^\infty d\xi\ \xi \int_0^\infty d\eta\ |e^{A_F\eta^3+D_F\xi^3}| \cdot \int_0^\infty d\xi \int_0^\infty d\eta\ |e^{A_B\eta^3+D_B\xi^3}| \\
&\le (|\mathfrak{M}D_F|)^{-2/3}\,(|\mathfrak{M}A_F|)^{-1/3}\,(|\mathfrak{M}D_B|)^{-1/3}\,(|\mathfrak{M}A_B|)^{-1/3} \\
&\quad \left(\int_0^\infty e^{-\xi^3}d\xi\right)^3\left(\int_0^\infty \xi e^{-\xi^3}d\xi\right)
\end{aligned} \tag{97}$$

where we have defined

$$\mathfrak{M}y = \begin{cases} \Re y & \text{if } \Re y \neq 0, \\ \Im y & \text{if } \Re y = 0. \end{cases} \tag{98}$$

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. (97) 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 (93) and (53)–(54). 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 $z \in (x - e^{-\alpha}, x + e^{-\alpha})$

$$|\mathbb{E}\mu_N(z) - \mu_{eq}(z;x_1)| \lesssim N^{-5/3}C(x_1, |x| + e^{-\alpha}) \tag{99}$$

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

$$\mathbb{E}\int \log|z - x|\mathbb{1}\{|z - x| \le e^{-\alpha}\}d\mu_N(z) \to 0 \tag{100}$$

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 *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*

$$\mathbb{E}||M_N||_{max}^N \lesssim e^{c_eN}. \tag{101}$$

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

$$\begin{aligned}\mathbb{E}||M||_{max}^N &\leq \sum_{i,j} \mathbb{E}|M_{i,j}|^N \\ &= \sum_{i,j} \mathbb{E}|\mathcal{N}(0,\sigma_{ij}^2)|^N \\ &= \sum_{i,j} \sigma_{ij}^N \mathbb{E}|\mathcal{N}(0,1)|^N \\ &\leq N^2 c^{N/2} N^{-N/2} \mathbb{E}|\mathcal{N}(0,1)|^N. \end{aligned} \tag{102}$$

Simple integration with a change of variables gives

$$\mathbb{E}|\mathcal{N}(0,1)|^N = 2^{\frac{N+1}{2}} \Gamma\left(\frac{N+1}{2}\right) \tag{103}$$

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

$$\begin{aligned}\mathbb{E}|\mathcal{N}(0,1)|^N &\sim 2^{\frac{N+1}{2}} \sqrt{\pi(N+1)} \left(\frac{N+1}{2e}\right)^{\frac{N-1}{2}} \\ &\sim 2\sqrt{\pi} e^{-\frac{N-1}{2}} N^{N/2} \left(\frac{N+1}{N}\right)^{N/2} \\ &\sim 2\sqrt{\pi e} N^{N/2}. \end{aligned} \tag{104}$$

So finally

$$\mathbb{E}||M||_{max}^N \lesssim N^2 c^{N/2} = e^{\frac{1}{2}N\log c + 2\log N} \leq e^{\left(\frac{1}{2}\log c + 2\right)N}, \tag{105}$$

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

Theorem 1 *For any $x_1 \in \mathbb{R}$, let H_N be a random $N \times N$ matrix distributed as in the statement of Lemma 4. Then as $N \to \infty$*

$$\begin{aligned}&\iint_B dx dx_1 \ \exp\left\{-N\left(\frac{1}{2s^2}x^2 + \frac{1}{2s_1^2}(x_1)^2\right)\right\} \mathbb{E}|\det\left(H_N(x_1) - x\right)| \\ &= \iint_B dx dx_1 \ \exp\left\{-N\left(\frac{1}{2s^2}x^2 + \frac{1}{2s_1^2}(x_1)^2 - \int \log|z-x| d\mu_{eq}(z) + o(1)\right)\right\} + o(1). \end{aligned} \tag{106}$$

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

$$\begin{aligned}&\left| \iint_B dx dx_1 \ \exp\left\{-N\left(\frac{1}{2s^2}x^2 + \frac{1}{2s_1^2}(x_1)^2\right)\right\} \mathbb{E}|\det(H_N(x_1) - x)| \right. \\ &\quad \left. - \iint_{B_{\leq R}} dx dx_1 \ \exp\left\{-N\left(\frac{1}{2s^2}x^2 + \frac{1}{2s_1^2}(x_1)^2\right)\right\} \mathbb{E}|\det\left(H_N(x_1) - x\right)| \right| \\ &\leq \iint_{||\boldsymbol{x}|| \geq R} dx dx_1 \ \exp\left\{-N\left(\frac{1}{2s^2}x^2 + \frac{1}{2s_1^2}(x_1)^2\right)\right\} \mathbb{E}|\det\left(H_N(x_1) - x\right)|. \end{aligned} \tag{107}$$

We have the upper bound (81) of Lemma 4 but this cannot be directly applied to (107) 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

$$J_N = H_N + x_1 \begin{pmatrix} I & 0 \\ 0 & 0 \end{pmatrix} \tag{108}$$

and then

$$|\det(H_N - xI)| \leq ||J_N||_{\max}^N \max\{|x|, |x_1|\}^N = ||J_N||_{\max}^N \exp\left(N \max\{\log|x|, \log|x_1|\}\right). \tag{109}$$

J_N has centred Gaussian entries with variance $\mathcal{O}(N^{-1})$, so Lemma 5 applies, and we find

$$\mathbb{E}|\det(H_N - xI)| \lesssim \exp\left(N \max\{\log|x|, \log|x_1|\}\right) e^{c_e N} \tag{110}$$

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

Now we have

$$\begin{aligned} &\left| \iint_B dx dx_1 \ \exp\left\{-N\left(\frac{1}{2s^2}x^2 + \frac{1}{2s_1^2}(x_1)^2\right)\right\} \mathbb{E}|\det(H_N(x_1) - x)| \right. \\ &\quad \left. - \iint_{B_{\leq R}} dx dx_1 \ \exp\left\{-N\left(\frac{1}{2s^2}x^2 + \frac{1}{2s_1^2}(x_1)^2\right)\right\} \mathbb{E}|\det(H_N(x_1) - x)| \right| \\ &\lesssim \iint_{||\boldsymbol{x}||\geq R} dx dx_1 \ \exp\left\{-N\left(\frac{1}{2s^2}x^2 + \frac{1}{2s_1^2}(x_1)^2 - \max\{\log|x|, \log|x_1|\} - c_e\right)\right\}. \end{aligned} \tag{111}$$

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

$$\frac{1}{2s^2}x^2 + \frac{1}{2s_1^2}(x_1)^2 - \max\{\log|x|, \log|x_1|\} - c_e > L > 0 \tag{112}$$

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

$$\begin{aligned} &\left| \iint_B dx dx_1 \ \exp\left\{-N\left(\frac{1}{2s^2}x^2 + \frac{1}{2s_1^2}(x_1)^2\right)\right\} \mathbb{E}|\det(H_N(x_1) - x)| \right. \\ &\quad \left. - \iint_{B_{\leq R}} dx dx_1 \ \exp\left\{-N\left(\frac{1}{2s^2}x^2 + \frac{1}{2s_1^2}(x_1)^2\right)\right\} \mathbb{E}|\det(H_N(x_1) - x)| \right| \\ &\lesssim N^{-1} e^{-NL} \to 0 \end{aligned} \tag{113}$$

as $N \to \infty$. Finally, for x, x_1 in $B_{\leq R}$, the result of the Lemma 4 holds uniformly in x, x_1, so

$$\begin{aligned} &\iint_{B_{\leq R}} dx dx_1 \ \exp\left\{-N\left(\frac{1}{2s^2}x^2 + \frac{1}{2s_1^2}(x_1)^2\right)\right\} \mathbb{E}|\det(H_N(x_1) - x)| \\ &= \iint_{B_{\leq R}} dx dx_1 \ \exp\left\{-N\left(\frac{1}{2s^2}x^2 + \frac{1}{2s_1^2}(x_1)^2 - \int \log|z - x| d\mu_{eq}(z; x_1) + o(1)\right)\right\}. \end{aligned} \tag{114}$$

The result follows from (113), (114) and the triangle inequality. □

5.1 Asymptotic Complexity with Prescribed Hessian Index

Recall the complexity defined in (7):

$$C_{N,k_D,k_G} = \left| \left\{ \boldsymbol{w}^{(D)} \in S^{N_D}, \boldsymbol{w}^{(G)} \in S^{N_G} \ : \ \nabla_D L^{(D)} = 0, \nabla_G L^{(G)} = 0, L^{(D)} \in B_D, L^{(G)} \in B_G \right.\right.$$
$$\left.\left. i\left(\nabla_D^2 L^{(D)}\right) = k_D, \ i\left(\nabla_G^2 L^{(G)}\right) = k_G \right\} \right|. \tag{7}$$

The extra Hessian signature conditions in (7) 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 (16), but our spectral calculations proceed from the rewritten form (21). We find however that we can indeed proceed much as in [13]. Recall the key Hessian matrix $\tilde{H}$ given in (16) by

$$\tilde{H} = \begin{pmatrix} \sqrt{2(N_D-1)}\sqrt{b^2+b_1^2}M^{(D)} & -bG \\ bG^T & \sqrt{2(N_G-1)}bM^{(G)} \end{pmatrix}$$
$$- \sqrt{N-2}x \begin{pmatrix} -I_{N_D} & 0 \\ 0 & I_{N_G} \end{pmatrix} + \sqrt{N-2}x_1 \begin{pmatrix} I_{N_D} & 0 \\ 0 & 0 \end{pmatrix} \tag{115}$$

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 (23) to slightly rewrite (16). We must address the problem of computing

$$\mathbb{E}|\det \tilde{H}|\mathbb{1}\left\{ i\left(\sqrt{\kappa}(1+\mathcal{O}(N^{-1}))\sqrt{b^2+b_1^2}M_D + \frac{x+x_1}{\sqrt{2}}\right)\right.$$
$$\left. = k_D, \ i\left(\sqrt{\kappa'}(1+\mathcal{O}(N^{-1}))bM_G - \frac{x}{\sqrt{2}}\right) = k_G \right\}. \tag{116}$$

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

$$A[t] = \boldsymbol{y}_1 \boldsymbol{y}_1^T + \boldsymbol{y}_2 \boldsymbol{y}_2^T + \zeta_1 \zeta_1^\dagger + \zeta_2 \zeta_2^\dagger. \tag{117}$$

With these definitions, we have [13]

$$|\det \tilde{H}| = (2(N-2))^{\frac{N-2}{2}} \lim_{\epsilon \searrow 0} \int d\Xi$$
$$\exp\left\{ -i\sqrt{\kappa}(1+\mathcal{O}(N^{-1}))\sqrt{b^2+b_1^2}\mathrm{Tr}M^{(D)}A[1] - i\sqrt{\kappa'}(1+\mathcal{O}(N^{-1}))b\mathrm{Tr}M^{(G)}A[2] \right\}$$
$$\exp\{\mathcal{O}(\epsilon)\}\exp\{\ldots\} \tag{118}$$

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

$$\mathbb{E}_{M^{(D)}} e^{-i\sqrt{\kappa}\sqrt{b^2+b_1^2}\mathrm{Tr}M^{(D)}A[1]}\mathbb{1}\left\{i\left(M_D+\frac{x+x_1}{\sqrt{\kappa}\sqrt{b^2+b_1^2}}(1+\mathcal{O}(N^{-1}))\right)=k_D\right\}, \tag{119}$$

$$\mathbb{E}_{M^{(G)}} e^{-i\sqrt{\kappa'}b\mathrm{Tr}M^{(G)}A[2]}\mathbb{1}\left\{i\left(M_G-\frac{x}{\sqrt{\kappa'}b}(1+\mathcal{O}(N^{-1}))\right)=k_G\right\}, \tag{120}$$

but [13] 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

$$\begin{aligned}
&K_L^{(D)} e^{-Nk_D\kappa(1+o(1))I_1(\hat{x}_D;\sqrt{2})} e^{-\frac{1}{2N}(b^2+b_1^2)\mathrm{Tr}A[1]^2}\\
&\leq \Re\mathbb{E}_{M^{(D)}} e^{-i\sqrt{\kappa}\sqrt{b^2+b_1^2}\mathrm{Tr}M^{(D)}A[1]}\mathbb{1}\left\{i\left(M_D+\frac{x+x_1}{\sqrt{\kappa}\sqrt{b^2+b_1^2}}(1+\mathcal{O}(N^{-1}))\right)=k_D\right\}\\
&\leq K_U^{(D)} e^{-Nk_D\kappa(1+o(1))I_1(\hat{x}_D;\sqrt{2})} e^{-\frac{1}{2N}(b^2+b_1^2)\mathrm{Tr}A[1]^2}
\end{aligned} \tag{121}$$

and

$$\begin{aligned}
&K_L^{(G)} e^{-Nk_G\kappa'(1+o(1))I_1(\hat{x}_G;\sqrt{2})} e^{-\frac{1}{2N}b^2\mathrm{Tr}A[2]^2}\\
&\leq \Re\mathbb{E}_{M^{(G)}} e^{-i\sqrt{\kappa'}b\mathrm{Tr}M^{(G)}A[2]}\mathbb{1}\left\{i\left(M_G-\frac{x}{\sqrt{\kappa'}b}(1+\mathcal{O}(N^{-1}))\right)=k_G\right\}\\
&\leq K_U^{(G)} e^{-Nk_G\kappa'(1+o(1))I_1(\hat{x}_G;\sqrt{2})} e^{-\frac{1}{2N}b^2\mathrm{Tr}A[2]^2}
\end{aligned} \tag{122}$$

where

$$\hat{x}_D=-\frac{x+x_1}{\sqrt{\kappa}\sqrt{b^2+b_1^2}},\quad \hat{x}_G=\frac{x}{\sqrt{\kappa'}b}. \tag{123}$$

Here I_1 is the rate function of the largest eigenvalue of the GOE as obtained in [45] and used in [2,13]:

$$I_1(u;E)=\begin{cases}\frac{2}{E^2}\int_u^{-E}\sqrt{z^2-E^2}dz & \text{for } u<-E,\\ \frac{2}{E^2}\int_E^{u}\sqrt{z^2-E^2}dz & \text{for } u>E,\\ \infty & \text{for } |u|<E.\end{cases} \tag{124}$$

Note that for $u<-E$

$$I_1(u;E)=-\frac{u}{E}\sqrt{u^2-E^2}-\log\left(-u+\sqrt{u^2-E^2}\right)+\log E \tag{125}$$

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 (21) so the complexity calculations using the Coulomb gas approach can proceed as before. We can achieve this with inverse Fourier transforms:

$$e^{-\frac{1}{2N}(b^2+b_1^2)\mathrm{Tr}A[1]^2}=\mathbb{E}_{M_D} e^{-i\sqrt{\kappa}\sqrt{b^2+b_1^2}\mathrm{Tr}M_D A[1]} \tag{126}$$

$$e^{-\frac{1}{2N}b^2\mathrm{Tr}A[2]^2}=\mathbb{E}_{M_G} e^{-i\sqrt{\kappa'}b\mathrm{Tr}M_G A[2]} \tag{127}$$

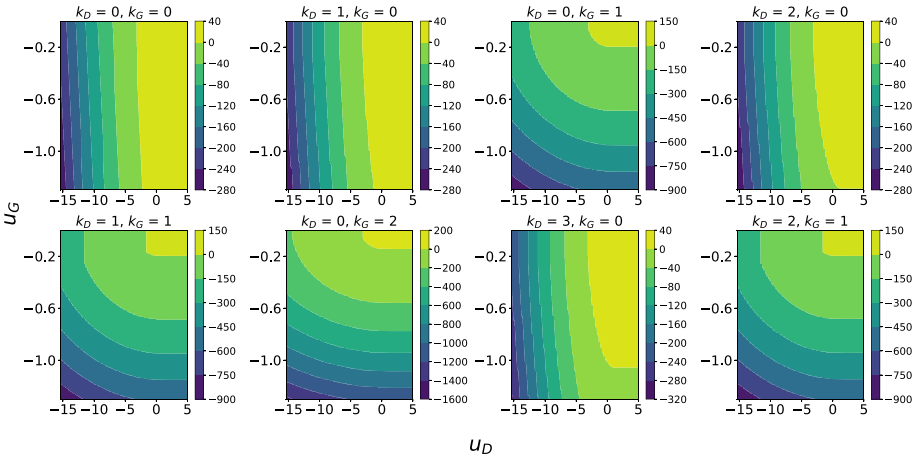


Fig. 5 Contour plots of Θ_{k_D,k_G} for a few values of k_D, k_G. Here $p = q = 3, \sigma_z = 1, \kappa = 0.9$

from which we obtain

$$\begin{aligned} K_L e^{-Nk_D\kappa(1+o(1))I_1(\hat{x}_D;\sqrt{2})} e^{-Nk_G\kappa'(1+o(1))I_1(\hat{x}_G;\sqrt{2})} \mathbb{E}|\det \tilde{H}| \\ \leq \mathbb{E}|\det \tilde{H}| \mathbb{1}\left\{ i\left(\sqrt{\kappa}(1+\mathcal{O}(N^{-1}))\sqrt{b^2+b_1^2} M_D + \frac{x+x_1}{\sqrt{2}}\right)\right. \\ \left. = k_D,\ i\left(\sqrt{\kappa'}(1+\mathcal{O}(N^{-1}))bM_G - \frac{x}{\sqrt{2}}\right) = k_G \right\} \qquad (128) \\ \leq K_U e^{-Nk_D\kappa(1+o(1))I_1(\hat{x}_D;\sqrt{2})} e^{-Nk_G\kappa'(1+o(1))I_1(\hat{x}_G;\sqrt{2})} \mathbb{E}|\det \tilde{H}|. \qquad (129) \end{aligned}$$

It follows that

$$\begin{aligned} & K'_N \iint_B dx dx_1 e^{-(N-2)\left[\Phi(x,x_1)+k_G\kappa' I_1(x;\sqrt{2\kappa'}b)+k_D\kappa I_1\left((-(x+x_1);\sqrt{2\kappa(b^2+b_1^2)}\right)\right](1+o(1))} \\ & \lesssim C_{N,k_D,k_G} \\ & \lesssim K'_N \iint_B dx dx_1 e^{-(N-2)\left[\Phi(x,x_1)+k_G\kappa' I_1(x;\sqrt{2\kappa'}b)+k_D\kappa I_1\left((-(x+x_1);\sqrt{2\kappa(b^2+b_1^2)}\right)\right](1+o(1))}. \end{aligned} \qquad (130)$$

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

$$\begin{aligned} \lim \frac{1}{N} \log \mathbb{E} C_{N,k_D,k_G} &= \Theta_{k_D,k_G}(u_D, u_G) \\ &= K - \min_B \left\{ \Phi + k_G\kappa' I_1(x;\sqrt{2\kappa'}b) + k_D\kappa I_1\left(-(x+x_1);\sqrt{2\kappa(b^2+b_1^2)}\right)\right\}. \end{aligned} \qquad (131)$$

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

Remark 2 Recall that the limiting spectral measure of the Hessian displays a transition as the support splits from one component to two, as shown in Fig. 1. Let us comment on the relevance of this feature to the complexity. The spectral measure appears in one place in

the above complexity calculations: the Coulomb gas integral $\int d\mu_{eq}(z) \log|z - x|$. The effect of integrating against the measure μ_{eq} is to smooth out the transition point. In other words, if μ_{eq} has two components or is at the transition point, one expects to be able to construct another measure ν supported on a single component such that $\int d\nu(z) \log|z-x| = \int d\mu_{eq}(z) \log|z - x|$. We interpret this to mean that the Coulomb gas integral term does not display any features that can be unambiguously attributed to the transition behaviour of the spectral measure.

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. [1] 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. The grey region in the plot clearly shows 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 [1] 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 [1] 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 Fig. 6. Imagine traversing a path near this contour from right to left (decreasing u_D values); an example path is approximately indicated by a black arrow on the figure. At all points along such a path, the only critical points present are exact local minima for both networks, however the losses range over

(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 Figs. 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

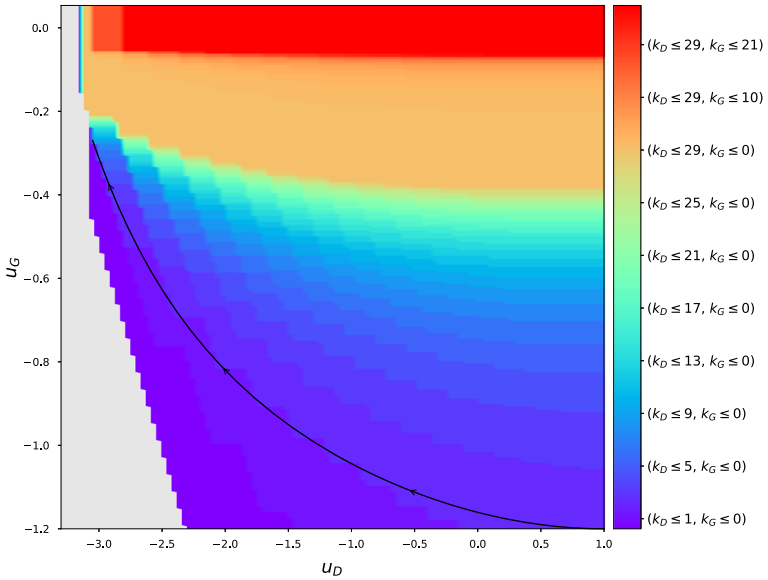


Fig. 6 Contours in the (u_D, u_G) plane of the maximum k_D and k_G such that $\Theta_{k_D,k_G}(u_D, u_G) > 0$. k_D results shown with a red colour red scheme, and k_G with blue/green. The grey region on the left lies outside the domain of definition of Θ_{k_D,k_G}. Here $p = q = 3, \sigma_z = 1, \kappa = 0.9$. The arrow indicates the approximate location of the contour discussed in the main text

GANs and our model, but also a new two-dimensional banded critical points structure. We can further illuminate the structure by plotting, for each (u_D, u_G), the approximate proportion of minima with both $L_D \leq u_D$ and $L_G \leq u_G$ out of all points where at at least one of those conditions holds. The expression is

$$\Theta\left(u_D, u_G\right) - \max\left\{\Theta\left(u_D, \infty\right), \Theta\left(\infty, u_G\right)\right\} \tag{132}$$

which gives the log of the ratio in units of N. We show the plot in Fig. 7. Note that, for large N, any region of the plot away from a value of zero contains exponentially more bad minima – where one of the networks has collapsed – than good minima, with equilibrium between the networks. The model therefore predicts the existence of good local minima (in the bottom left of Fig. 7) that are effectively inaccessible due to their being exponentially outnumbered by bad local minima.

The structure revealed by our analysis 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.

Remark 3 A natural question in the context of our analysis of low-index critical points is: do such points reflect the points typically reached by gradient descent algorithms used to train real GANs? There has been much discussion in the literature of the analogous question for single networks and spin glasses [1,7,9]. It is not clear how to settle this question in our case,

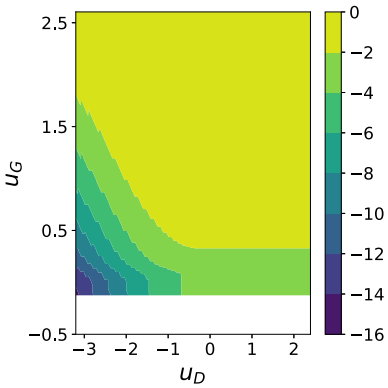


Fig. 7 Contour plot of the log ratio quantity given in (132). This is the approximate proportion of minima with both $L_D \leq u_D$ and $L_G \leq u_G$ out of all points where at at least one of those conditions holds

but we believe our model and its low-index critical points give a description of the baseline properties to be expected of high-dimensional adversarial optimisation problems late in the optimisation procedure. In addition, the unstructured random noise present in spin glasses may be more appropriate in our model for GANs than it is for single spin-glass models of single networks, as GAN generators do genuinely contain unstructured latent noise, rather than just the highly-structured data distributions seen on real data.

Remark 4 The issue of meta-stability is also worth mentioning. In single spin glasses, the boundary E_∞ between fixed index and unbounded index critical points is meta-stable [46,47]. From the random matrix theory perspective, the E_∞ boundary corresponds to the left edge of the Wigner semi-circle [2]. There are $O(N)$ eigenvalues in any finite interval at the left of the Wigner semi-circle, corresponding to $O(N)$ Hessian eigenvalues in any neighbourhood around zero. The 2D analogue of the E_∞ boundary in our double spin-glass model is expected to possess the same meta-stability: the Wigner semi-circle is replaced by the measure studied in Sect. 4, to which the preceding arguments apply. In the context of deep neural networks, there is a related discussion concerning "wide and flat local optima" of the loss surface, i.e. local optima for which many of the Hessian eigenvalues are close to zero. There are strong indications that deep neural networks converge under gradient-based optimisation to such optima [48–53] and that they are perhaps better for generalisation (i.e. test set loss) than other local optima, however some authors have challenged this view [54–58]. It is beyond the scope of the present work to analyse the role of meta-stability further, however we note that the indications from machine learning are that it is most significant when considering generalisation, however our work simplifies to the case of a single loss rather than separately considering training and test loss.

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 [1,2,13], 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, choosing $\kappa = 0, 1$ corresponds to one network having a negligible number of parameters when compared with the other

and we would expect the much larger network to prevail in the minimax game, however our theoretical results above are valid strictly for $\kappa \in (0, 1)$.

In the following two subsections we examine effect of σ_z and κ in our theoretical and in real experiments with a DCGAN [26]. Additional supporting plots are given in the appendix.

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:

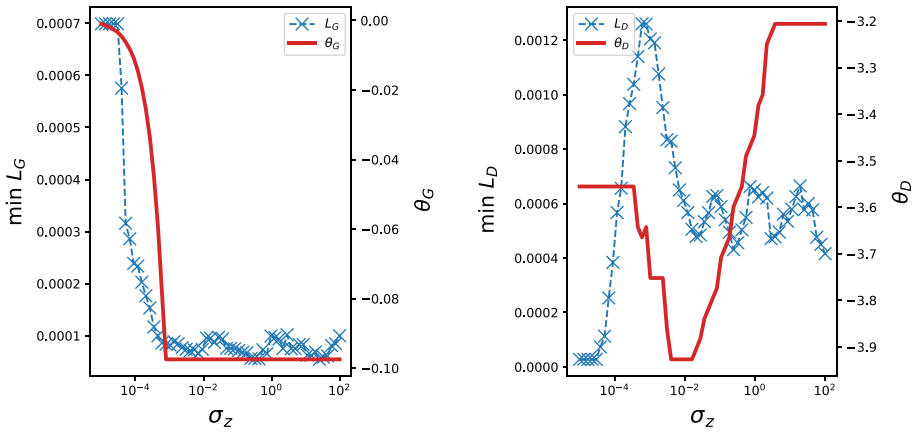
1. $\vartheta_D = \min\{u_D \in \mathbb{R} \mid \exists u_G \in \mathbb{R} \ : \ \Theta(u_D, u_G) > 0\}$;
2. $\vartheta_G = \min\{u_G \in \mathbb{R} \mid \exists u_D \in \mathbb{R} \ : \ \Theta(u_D, u_G) > 0\}$.

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

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 [26] with 5 layers in each network, using the reference PyTorch implementation from [59], however we introduce the generator noise scale σ_z. That is, the latent input noise vector $\boldsymbol{z}$ for the generator is sampled from $\mathcal{N}(0, \sigma_z^2 I)$. For a given σ_z, we train the GANs for 10 epochs on CIFAR10 [60] 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 Fig. 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.

We are not aware of σ_z tuning being widely used in practice for real GANs, rather it is typically taken to be unity. We have chosen this parameter, as it can be directly paralleled in our spin glass model, therefore allowing for the above experimental comparison. Naturally there are other parameters of real GANs that one might wish to study (such as learning rates and batch sizes) however these are much less readily mirrored in the spin glass model and complexity analysis, precluding comparisons between theory and experiment. Nevertheless, the experimental results in Fig. 8 do demonstrate that tuning σ_z in real GANs could be of benefit, as $\sigma_z = 1$ does not appear to be the optimal value.



(a) Generator **(b)** Discriminator

Fig. 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 blue cross-dashed lines show the experimental DCGAN results, and solid red lines 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

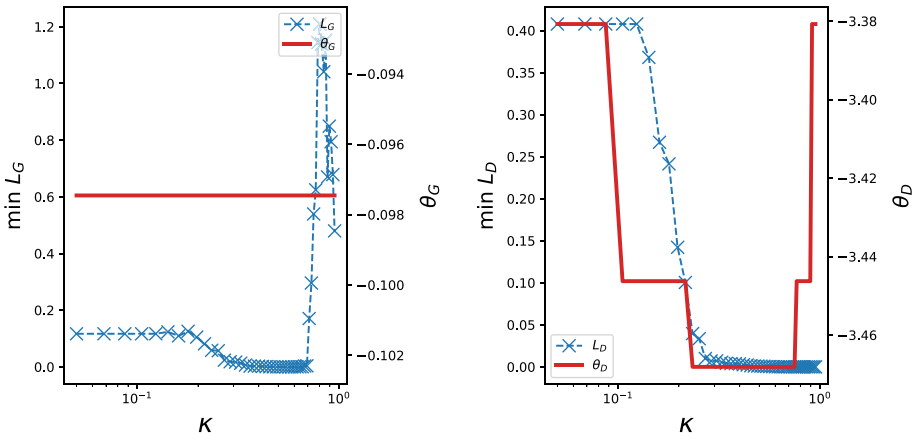
Similarly to the previous section, we can investigate the effect of κ using ϑ_D, ϑ_G while varying κ over $(0, 1)$. To achieve this variation in the DCGAN, we vary the number of convolutional filters in each network. The generator and discriminator are essentially mirror images of each other and the 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 Fig. 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.

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

[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).



(a) Generator **(b)** Discriminator

Fig. 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 blue cross-dashed lines show the experimental DCGAN results, and the solid red 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

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.

Our analysis has focused on the annealed complexity of our spin glass model (i.e. taking the logarithm after the expectation) rather than the quenched complexity (i.e. taking the expectation after the logarithm). Ideally one would compute both, as the quenched complexity is often considered to reflect the typical number of stationary points and is bounded above by the annealed complexity. Computing the quenched complexity is typically more challenging than the annealed and such a calculation for our model could be the subject of a further work requiring considerable technical innovations. Even the elegant and very general methods presented recently in [40] are restricted only to the annealed case. Agreement between annealed and quenched is known only in a few special cases closely related to spherical spin glasses [61–63] and is not expected in general [33]. It is conceivable that quenched and annealed complexity agree in the case of our model, as it closely related to spin glasses and possesses no distinguished directions (i.e. spikes) such as are present in [33]. Establishing agreement by existing methods requires analysis of pairs of correlated GOE-like matrices. Such an approach for our model may well require analysis of at least 4 correlated matrices (2 per diagonal block), and quite possibly more, including correlations between blocks. We leave this considerable challenge for future work.

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 [13], 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.

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.

Acknowledgements FM is grateful for support from the University Research Fellowship of the University of Bristol. JPK is pleased to acknowledge support from European Research Council Advanced Grant 740900 (LogCorRM). NPB is grateful to Diego Granziol for useful discussions (in particular suggesting Fig. 7), to Jonathan Hodgson for help designing the contour plots and to the Advanced Computing Research Centre of the University of Bristol for the GPU resources to perform the experiments. The authors are grateful to several anonymous reviewers whose comments led to considerable improvements in this paper.

Appendix A: Kac-Rice Results

We repeat here the foundational Kac-Rice result on which our complexity calculations are built.

Theorem 2 ([41] 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*

$$N_{\boldsymbol{u}} \equiv |\{x \in \mathcal{M} \mid \phi(x) = \boldsymbol{u},\ \psi(x) \in A\}| \,. \tag{133}$$

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 $\boldsymbol{u}$.

(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 $\boldsymbol{u}$, uniformly in $\mathcal{M}$.

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

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

(f) *The following moment condition holds*

$$\sup_{x\in\mathcal{M}} \max_{1\le i,j\le N} \mathbb{E}\left\{\left|\nabla_{E_j}\phi^i(x)\right|^N\right\} < \infty \tag{134}$$

(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* $\epsilon > 0$

$$\mathbb{P}(\omega(\eta) > \epsilon) = o(\eta^N), \quad \text{as } \eta \downarrow 0 \tag{135}$$

where the modulus of continuity *of a real-valued function* G *on a metric space* (T, τ) *is defined as (c.f. [41] around (1.3.6))*

$$\omega(\eta) := \sup_{s,t:\tau(s,t)\le\eta} |G(s) - G(t)| \tag{136}$$

Then

$$\mathbb{E}N_{\boldsymbol{u}} = \int_{\mathcal{M}} \mathbb{E}\{|\det \nabla_E\phi(x)|\mathbb{1}\{\psi(x) \in A\} \mid \phi(x) = \boldsymbol{u}\}\, p_x(\boldsymbol{u}) Vol_g(x) \tag{137}$$

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

Proof of Lemma 1. In the notation of Theorem 2, 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. It is sufficient to check the conditions of Theorem 2 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 [2]. But since $\mathcal{M}$ is just a direct product with product topology, it immediately follows that (g) is satisfied, so Theorem 2 applies and we obtain (8). (9) follows simply, using the rules of conditional expectation.

Appendix B: Gaussian Hessian Calculations

In this section we give the full details of the Gaussian calculations for the distribution of the Hessian:

$$\begin{pmatrix} \nabla_D^2 L^{(D)} & \nabla_{GD} L^{(D)} \\ \nabla_{DG} L^{(G)} & \nabla_G^2 L^{(G)} \end{pmatrix} \Bigg| \nabla_G L^{(G)} = 0, \nabla_D L^{(D)} = 0, L^{(D)} \in B_D, L^{(G)} \in B_G. \tag{138}$$

These calculations are routine and consist of repeated application of standard results for conditioning multivariate Gaussians, but the details are nevertheless intricate.

Recall the definitions

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

and

$$\ell^{(D)}(\boldsymbol{w}^{(D)}) = \sum_{i_1,\ldots,i_p=1}^{N_D} X_{i_1,\ldots,i_p} \prod_{k=1}^{p} w_{i_k}^{(D)}$$
$$\ell^{(G)}(\boldsymbol{w}^{(D)}, \boldsymbol{w}^{(G)}) = \sum_{i_1,\ldots,i_{p+q}=1}^{N_D+N_G} Z_{i_1,\ldots,i_{p+q}} \prod_{k=1}^{p+q} w_{i_k}$$

for i.i.d. Gaussian X and Z, where $\boldsymbol{w}^T = (\boldsymbol{w}^{(D)^T}, \boldsymbol{w}^{(G)^T})$. As mentioned in the main text, we have spherical symmetry in both $\boldsymbol{w}^{(D)}$ and $\boldsymbol{w}^{(G)}$, so it sufficient to consider the distribution (138) around some fixed specific points on the spheres S^{N_D} and S^{N_G}. Following [2], we choose the north poles. We can select a coordinate basis around both poles, e.g. with

$$\boldsymbol{w}^{(D)} = \left(\sqrt{1-\boldsymbol{u}^2}, \boldsymbol{u}\right), \quad \boldsymbol{w}^{(G)} = \left(\sqrt{1-\boldsymbol{v}^2}, \boldsymbol{v}\right),$$

for $\boldsymbol{u} \in \mathbb{R}^{N_D-1}$, $\boldsymbol{v} \in \mathbb{R}^{N_G-1}$ with $\boldsymbol{u}^2 \leq 1$, $\boldsymbol{v}^2 \leq 1$.

We need the joint distributions

$$\left(\ell^{(D)}, \partial_i^{(D)}\ell^{(D)}, \partial_{jk}^{(D)}\ell^{(D)}\right), \quad \left(\ell^{(G)}, \partial_i^{(G)}\ell^{(G)}, \partial_{jk}^{(G)}\ell^{(G)}, \partial_l^{(D)}\ell^{(G)}, \partial_{mn}^{(D)}\ell^{(G)}\right)$$

where the two groups are independent from of each other. *The derivatives $\partial^{(D)}$, $\partial^{(G)}$ are now Euclidean derivatives with respect to the coordinates $\boldsymbol{u}$, $\boldsymbol{v}$.* $\ell^{(D)}$ behaves just like a single spin glass, and so we have [2]:

$$Var(\ell^{(D)}) = 1, \tag{139}$$
$$Cov(\partial_i^{(D)}\ell^{(D)}, \partial_{jk}^{(D)}\ell^{(D)}) = 0, \tag{140}$$
$$\partial_{ij}^{(D)}\ell^{(D)} \mid \{\ell^{(D)} = x_D\} \sim \sqrt{(N_D-1)p(p-1)}GOE^{N_D-1} - x_D pI. \tag{141}$$

To find the joint and thence conditional distributions for $\ell^{(G)}$, we first note that $\ell^{(G)}$ is simply a spin glass on a partitioned vector $\boldsymbol{w}^T = (\boldsymbol{w}^{(D)^T}, \boldsymbol{w}^{(G)^T})$, so

$$Cov(\ell^{(G)}(\boldsymbol{w}^{(D)}, \boldsymbol{w}^{(G)}), \ell^{(G)}(\boldsymbol{w}^{(D)'}, \boldsymbol{w}^{(G)'})) = \left(\boldsymbol{w}^{(D)} \cdot \boldsymbol{w}^{(D)'} + \boldsymbol{w}^{(G)} \cdot \boldsymbol{w}^{(G)'}\right)^{p+q} \tag{142}$$

from which, by comparing with [2], one can obtain the necessary expressions, at the north poles in a coordinate basis. Practically, one writes $\boldsymbol{w}^{(D)^T} = (\sqrt{1-\sum_j u_j^2}, u_1, \ldots, u_{N_D-1})$, and similarly for $\boldsymbol{w}^{(G)}$. Then one takes derivatives of (142) with respect to these new variables around the north poles. Finally, one sets $\boldsymbol{w}^{(D)} = \boldsymbol{w}^{(D)'}$ and takes $u_j = 0\ \forall j$, and similarly for $\boldsymbol{w}^{(G)}$. The resulting expressions are largely familiar from the standard spin glass in [2],

except there are extra cross terms between $\boldsymbol{w}^{(D)}$ and $\boldsymbol{w}^{(G)}$:

$$Var(\ell^{(G)}) = 2^{p+q}, \tag{143}$$

$$Cov\left(\partial_{ij}^{(G)}\ell^{(G)}, \ell^{(G)}\right) = -(p+q)2^{p+q}\delta_{ij}, \tag{144}$$

$$Cov\left(\partial_{ij}^{(D)}\ell^{(G)}, \ell^{(G)}\right) = -(p+q)2^{p+q}\delta_{ij}, \tag{145}$$

$$\begin{aligned} Cov\left(\partial_{ij}^{(G)}\ell^{(G)}, \partial_{kl}^{(G)}\ell^{(G)}\right) &= 2^{p+q}\left[(p+q)(p+q-1)\left(\delta_{ik}\delta_{jl}+\delta_{il}\delta_{jk}\right)\right. \\ &\quad \left.+(p+q)^2\delta_{ij}\delta_{kl}\right], \end{aligned} \tag{146}$$

$$Cov\left(\partial_{ij}^{(G)}\ell^{(G)}, \partial_{kl}^{(D)}\ell^{(G)}\right) = 2^{p+q}(p+q)^2\delta_{ij}\delta_{kl}, \tag{147}$$

$$Cov\left(\partial_{i}^{(G)}\partial_{j}^{(D)}\ell^{(G)}, \partial_{k}^{(G)}\partial_{l}^{(D)}\ell^{(G)}\right) = 2^{p+q}(p+q)(p+q-1)\delta_{ik}\delta_{jl}, \tag{148}$$

$$Cov\left(\partial_{ij}^{(G)}\ell^{(G)}, \partial_{k}^{(G)}\partial_{l}^{(D)}\ell^{(G)}\right) = 0 \tag{149}$$

$$Cov\left(\partial_{ij}^{(D)}\ell^{(G)}, \partial_{k}^{(D)}\partial_{l}^{(G)}\ell^{(G)}\right) = 0, \tag{150}$$

$$Cov\left(\partial_{i}^{(D)}\partial_{j}^{(G)}\ell^{(G)}, \ell^{(G)}\right) = 0. \tag{151}$$

Also, all first derivatives of $\ell^{(G)}$ are clearly independent of $\ell^{(G)}$ and its second derivatives by the same reasoning as in [2]. Note that

$$Cov\left(\partial_{i}^{(D)}L^{(D)}, \partial_{j}^{(D)}L^{(D)}\right) = \left(p+\sigma_z^2 2^{p+q}(p+q)\right)\delta_{ij} \tag{152}$$

$$Cov\left(\partial_{i}^{(G)}L^{(G)}, \partial_{j}^{(G)}L^{(G)}\right) = \sigma_z^2 2^{p+q}(p+q)\delta_{ij} \tag{153}$$

$$Cov\left(\partial_{i}^{(D)}L^{(D)}, \partial_{j}^{(G)}L^{(G)}\right) = 0 \tag{154}$$

and so

$$\varphi_{(\nabla_D L^{(D)}, \nabla_G L^{(G)})}(0) = (2\pi)^{-\frac{N-2}{2}}\left(p+\sigma_z^2 2^{p+1}(p+q)\right)^{-\frac{N_D-1}{2}}\left(\sigma_z^2 2^{p+q}(p+q)\right)^{-\frac{N_G-1}{2}}. \tag{155}$$

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

$$\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} \tag{156}$$

where

$$\Sigma_{11} = 2^{p+q}\begin{pmatrix} (p+1)(p+q-1)(1+\delta_{ij})+(p+q)^2\delta_{ij} & (p+q)^2\delta_{ij}\delta_{kl} \\ (p+q)^2\delta_{ij}\delta_{kl} & (p+1)(p+q-1)(1+\delta_{kl})+(p+q)^2\delta_{kl} \end{pmatrix}, \tag{157}$$

$$\Sigma_{12} = -2^{p+q}(p+q)\begin{pmatrix} \delta_{ij} \\ \delta_{kl} \end{pmatrix}, \tag{158}$$

$$\Sigma_{21} = -2^{p+q}(p+q)\begin{pmatrix} \delta_{ij} & \delta_{kl} \end{pmatrix}, \tag{159}$$

$$\Sigma_{22} = 2^{p+q}. \tag{160}$$

The conditional covariance is then

$$\bar{\Sigma} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} = 2^{p+q}(p+1)(p+q-1)\begin{pmatrix} 1+\delta_{ij} & 0 \\ 0 & 1+\delta_{kl} \end{pmatrix}. \tag{161}$$

Identical reasoning applied to $(\partial_{ij}^{(G)}\ell^{(G)}, \partial_{kl}^{(G)}\ell^{(G)}, \ell^{(G)})$ and $(\partial_{ij}^{(D)}\ell^{(G)}, \partial_{kl}^{(D)}\ell^{(G)}, \ell^{(G)})$ shows that, conditional on $\{\ell^{(G)} = x_G\}$, $\nabla_G^2\ell^{(G)}$ and $\nabla_D^2\ell^{(G)}$ have independent entries up-to symmetry, so 161 demonstrates they are independent GOEs and we have:

$$\begin{aligned}
&\begin{pmatrix} -\nabla_D^2\ell^{(G)} & -\nabla_G\nabla_D\ell^{(G)} \\ \nabla_D\nabla_G\ell^{(G)} & \nabla^2\ell^{(G)} \end{pmatrix} \mid \{\ell^{(G)} = x_G\} \\
&\overset{d}{=} \sqrt{2^{p+q+1}(p+q)(p+q-1)} \begin{pmatrix} \sqrt{N_D-1}M_1^{(D)} & -2^{-1/2}G \\ 2^{-1/2}G^T & \sqrt{N_G-1}M^{(G)} \end{pmatrix} \\
&\quad - (p+q)x_G 2^{p+1} \begin{pmatrix} -I_{N_D} & 0 \\ 0 & I_{N_G} \end{pmatrix}
\end{aligned} \tag{162}$$

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.

Appendix C: Bipartite Spin-Glass Formulation

Recalling the expression for $\ell^{(G)}$, one could argue that a more natural formulation would be

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

for i.i.d. Gaussian Z. In this case, each term in the sum contains exactly p weights from the discriminator network and q weights from the generator. This object is known as a bipartite spin glass. We will now present the Gaussian calculations. We need the joint distributions

$$\left(\ell^{(D)}, \partial_i^{(D)}\ell^{(D)}, \partial_{jk}^{(D)}\ell^{(D)}\right), \quad \left(\ell^{(G)}, \partial_i^{(G)}\ell^{(G)}, \partial_{jk}^{(G)}\ell^{(G)}, \partial_l^{(D)}\ell^{(G)}, \partial_{mn}^{(D)}\ell^{(G)}\right)$$

where the two groups are independent from of each other. As in [2], 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 [2]:

$$Var(\ell^{(D)}) = 1, \tag{163}$$

$$Cov(\partial_i^{(D)}\ell^{(D)}, \partial_{jk}^{(D)}\ell^{(D)}) = 0, \tag{164}$$

$$\partial_{ij}^{(D)}\ell^{(D)} \mid \{\ell^{(D)} = x_D\} \sim \sqrt{(N_D-1)p(p-1)}GOE^{N_D-1} - x_D pI, \tag{165}$$

$$Cov(\partial_i^{(D)}\ell^{(D)}, \partial_j^{(D)}\ell^{(D)}) = p\delta_{ij}. \tag{166}$$

To find the joint and thence conditional distributions for $\ell^{(G)}$, we first compute the covariance function, which follows from the independence of the Z:

$$Cov(\ell^{(G)}(\boldsymbol{w}^{(D)}, \boldsymbol{w}^{(G)}), \ell^{(G)}(\boldsymbol{w}^{(D)'}, \boldsymbol{w}^{(G)'})) \tag{167}$$

$$= \sum_{\substack{i_1,\ldots,i_p=1 \\ i'_1,\ldots,i'_p=1}}^{N_D} \sum_{\substack{j_1,\ldots,j_q=1 \\ j'_1,\ldots,j'_q=1}}^{N_G} \mathbb{E} Z_{\boldsymbol{i}\boldsymbol{i}} Z_{\boldsymbol{i}'\boldsymbol{j}'} \prod_{k=1}^{p} w_{i_k}^{(D)} w_{i'_k}^{(D)'} \prod_{l=1}^{q} w_{j_l}^{(G)} w_{j'_l}^{(G)'} \tag{168}$$

$$= \sum_{i_1,\ldots,i_p=1}^{N_D} \sum_{j_1,\ldots,j_q=1}^{N_G} \prod_{k=1}^{p} w_{i_k}^{(D)} w_{i_k}^{(D)'} \prod_{l=1}^{q} w_{j_l}^{(G)} w_{j_l}^{(G)'} \tag{169}$$

$$= \left(\boldsymbol{w}^{(D)} \cdot \boldsymbol{w}^{(D)'}\right)^p (\boldsymbol{w}^{(G)} \cdot \boldsymbol{w}^{(G)'})^q \tag{170}$$

The product structure of the covariance function implies that we can write down the following covariances directly from the simple spin-glass case, as the $\partial^{(D)}$ and $\partial^{(G)}$ derivatives act independently on their respective terms:

$$Var(\ell^{(G)}) = 1, \tag{171}$$

$$Cov\left(\partial_{ij}^{(G)} \ell^{(G)}, \ell^{(G)}\right) = -q\delta_{ij}, \tag{172}$$

$$Cov\left(\partial_{ij}^{(D)} \ell^{(G)}, \ell^{(G)}\right) = -p\delta_{ij}, \tag{173}$$

$$Cov\left(\partial_{ij}^{(G)} \ell^{(G)}, \partial_{kl}^{(G)} \ell^{(G)}\right) = q(q-1)\left(\delta_{ik}\delta_{jl} + \delta_{il}\delta_{jk}\right) + q^2\delta_{ij}\delta_{kl}, \tag{174}$$

$$Cov\left(\partial_{ij}^{(D)} \ell^{(G)}, \partial_{kl}^{(D)} \ell^{(G)}\right) = p(p-1)\left(\delta_{ik}\delta_{jl} + \delta_{il}\delta_{jk}\right) + p^2\delta_{ij}\delta_{kl}, \tag{175}$$

$$Cov\left(\partial_{ij}^{(G)} \ell^{(G)}, \partial_{kl}^{(D)} \ell^{(G)}\right) = pq\delta_{ij}\delta_{kl}, \tag{176}$$

$$Cov\left(\partial_{i}^{(G)}\partial_{j}^{(D)} \ell^{(G)}, \partial_{k}^{(G)}\partial_{l}^{(D)} \ell^{(G)}\right) = pq\delta_{ik}\delta_{jl}, \tag{177}$$

$$Cov\left(\partial_{ij}^{(G)} \ell^{(G)}, \partial_{k}^{(G)}\partial_{l}^{(D)} \ell^{(G)}\right) = 0 \tag{178}$$

$$Cov\left(\partial_{ij}^{(D)} \ell^{(G)}, \partial_{k}^{(D)}\partial_{l}^{(G)} \ell^{(G)}\right) = 0, \tag{179}$$

$$Cov\left(\partial_{i}^{(D)}\partial_{j}^{(G)} \ell^{(G)}, \ell^{(G)}\right) = 0. \tag{180}$$

Also, all first derivatives of $\ell^{(G)}$ are clearly independent of $\ell^{(G)}$ and its second derivatives by the same reasoning and

$$Cov\left(\partial_{i}^{(G)} \ell^{(G)}, \partial_{j}^{(G)} \ell^{(G)}\right) = q\delta_{ij}, \tag{181}$$

$$Cov\left(\partial_{i}^{(D)} \ell^{(G)}, \partial_{j}^{(D)} \ell^{(G)}\right) = p\delta_{ij}, \tag{182}$$

$$Cov\left(\partial_{i}^{(D)} \ell^{(G)}, \partial_{j}^{(G)} \ell^{(G)}\right) = 0. \tag{183}$$

We caw deduce the full gradient covariances, recalling that $\ell^{(D)}$ and $\ell^{(G)}$ are independent:

$$Cov\left(\partial_i^{(D)} L^{(D)}, \partial_j^{(D)} L^{(D)}\right) = p\left(1+\sigma_z^2\right)\delta_{ij} \tag{184}$$

$$Cov\left(\partial_i^{(G)} L^{(G)}, \partial_j^{(G)} L^{(G)}\right) = \sigma_z^2 q \delta_{ij} \tag{185}$$

$$Cov\left(\partial_i^{(D)} L^{(D)}, \partial_j^{(G)} L^{(G)}\right) = 0 \tag{186}$$

and so

$$\varphi_{(\nabla_D L^{(D)}, \nabla_G L^{(G)})}(0) = (2\pi)^{-\frac{N-2}{2}} \left(p+\sigma_z^2 p\right)^{-\frac{N_D-1}{2}} \left(\sigma_z^2 q\right)^{-\frac{N_G-1}{2}}. \tag{187}$$

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

$$\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} \tag{188}$$

where

$$\Sigma_{11} = \begin{pmatrix} p(p-1)(1+\delta_{ij}) + p^2\delta_{ij} & pq\delta_{ij}\delta_{kl} \\ pq\delta_{ij}\delta_{kl} & q(q-1)(1+\delta_{kl}) + q^2\delta_{kl} \end{pmatrix}, \tag{189}$$

$$\Sigma_{12} = -\begin{pmatrix} p\delta_{ij} \\ q\delta_{kl} \end{pmatrix}, \tag{190}$$

$$\Sigma_{21} = -\begin{pmatrix} p\delta_{ij} & q\delta_{kl} \end{pmatrix}, \tag{191}$$

$$\Sigma_{22} = 1. \tag{192}$$

The conditional covariance is then

$$\bar{\Sigma} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} \tag{193}$$

$$= \begin{pmatrix} p(p-1)(1+\delta_{ij}) & 0 \\ 0 & q(q-1)(1+\delta_{kl}) \end{pmatrix}. \tag{194}$$

Repeating this calculation for $(\partial_{ij}^{(G)}\ell^{(G)}, \partial_{kl}^{(G)}\ell^{(G)}, \ell^{(G)})$ demonstrates that $\nabla_G^2\ell^{(G)} \mid \{\ell^{(G)} = x_G\}$ has independent entries, up-to symmetry. The result (194) demonstrates that, conditional on $\{\ell^{(G)} = x_G\}$, $\nabla_G^2\ell^{(G)}$ and $\nabla_D^2\ell^{(G)}$ are independent GOEs. In summary, from (194) and (177–179) we obtain

$$\begin{pmatrix} -\nabla_D^2\ell^{(G)} & -\nabla_G\nabla_D\ell^{(G)} \\ \nabla_D\nabla_G\ell^{(G)} & \nabla^2\ell^{(G)} \end{pmatrix} \mid \{\ell^{(G)} = x_G\}$$

$$\stackrel{d}{=} \sqrt{2}\begin{pmatrix} \sqrt{N_D-1}\sqrt{p(p-1)}M^{(D)} & -2^{-1/2}\sqrt{pq}G \\ 2^{-1/2}\sqrt{pq}G^T & \sqrt{N_G-1}\sqrt{q(q-1)}M^{(G)} \end{pmatrix}$$

$$- x_G\begin{pmatrix} -pI_{N_D} & 0 \\ 0 & qI_{N_G} \end{pmatrix} \tag{195}$$

where $M^{(D)} \sim GOE^{N_D-1}$ 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.

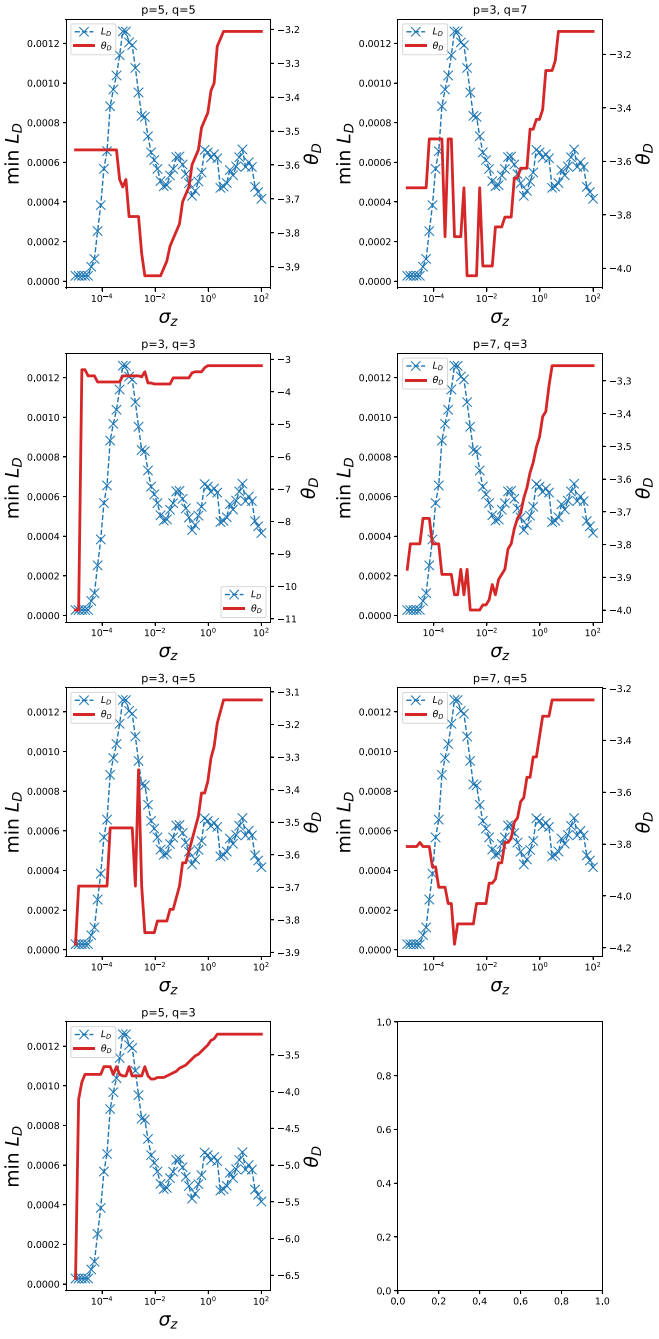


Fig. 10 The effect of σ_z on minimum L_D. Comparison of theoretical predictions of minimum possible discriminator and generator losses to observed minimum losses when training DCGAN on CIFAR10. The blue cross-dashed lines show the experimental DCGAN results, and the solid red show the theoretical results ϑ_G, ϑ_D. $\kappa = 0.5$ is used and p, q are varied

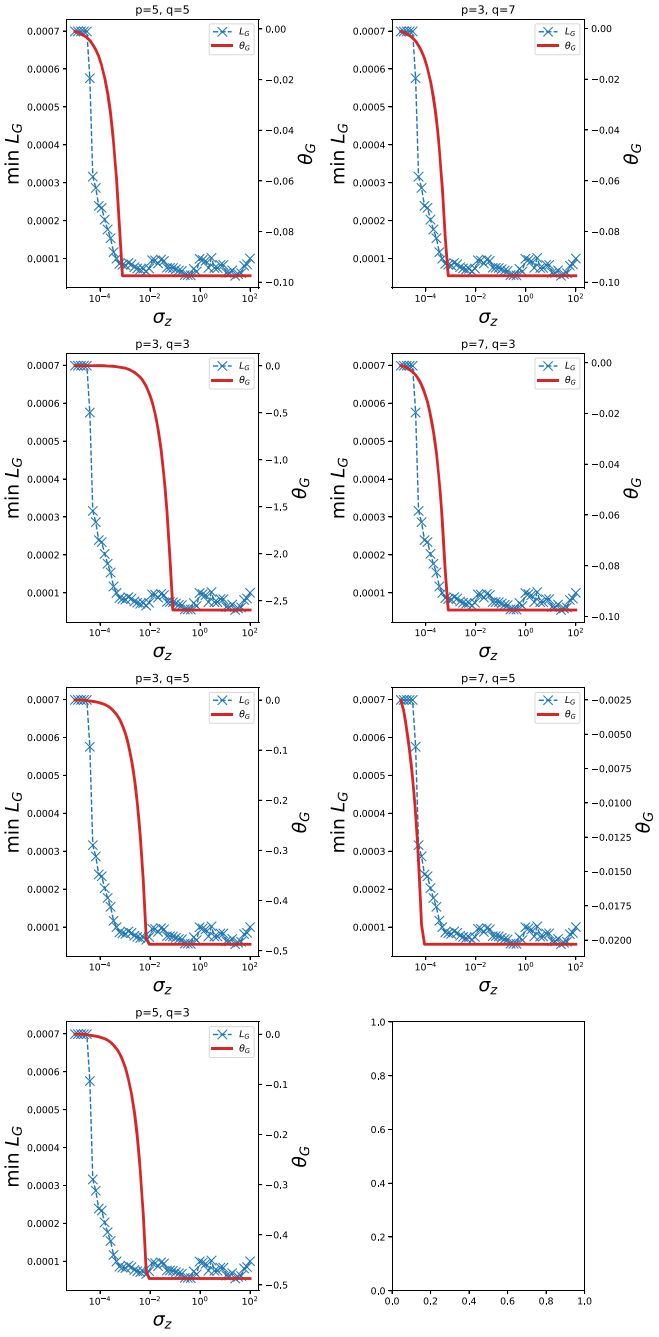


Fig. 11 The effect of σ_z on minimum L_G. Comparison of theoretical predictions of minimum possible discriminator and generator losses to observed minimum losses when training DCGAN on CIFAR10. The blue cross-dashed lines show the experimental DCGAN results, and the solid red show the theoretical results ϑ_G, ϑ_D. $\kappa = 0.5$ is used and p, q are varied

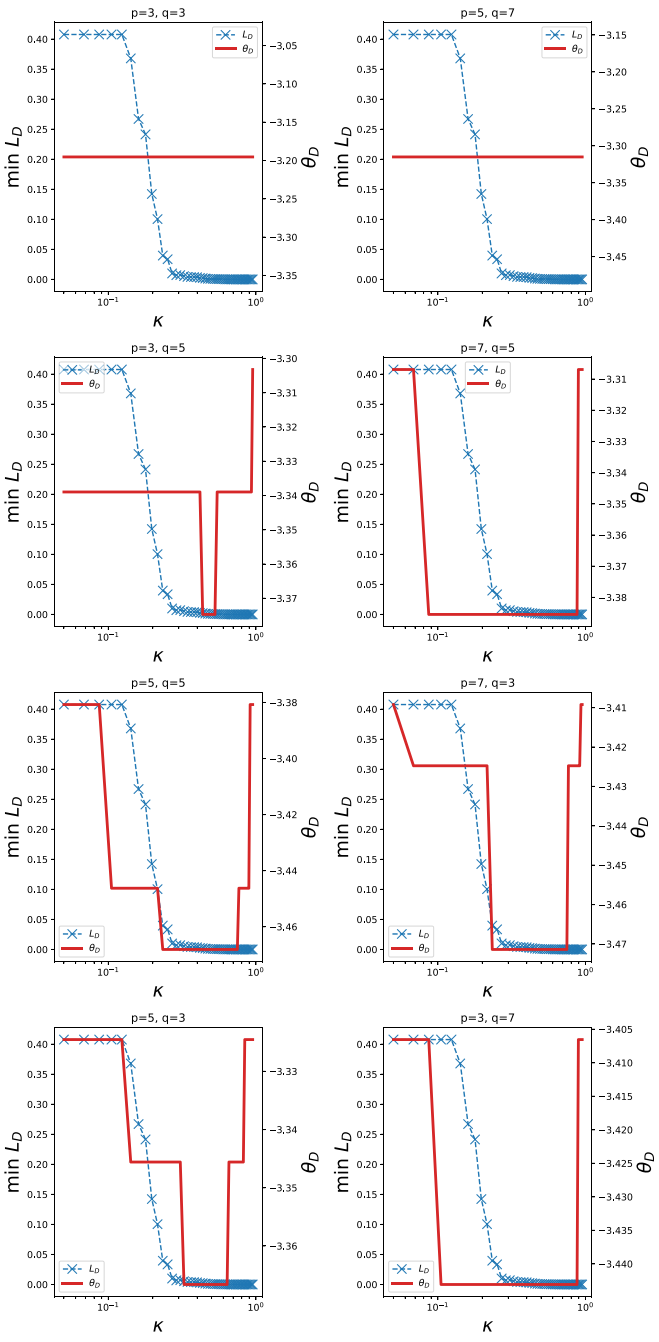


Fig. 12 The effect of κ on minimum L_D. Comparison of theoretical predictions of minimum possible discriminator and generator losses to observed minimum losses when training DCGAN on CIFAR10. The blue cross-dashed lines show the experimental DCGAN results, and the solid red show the theoretical results ϑ_G, ϑ_D. $\sigma_z = 1$ is used and p, q are varied

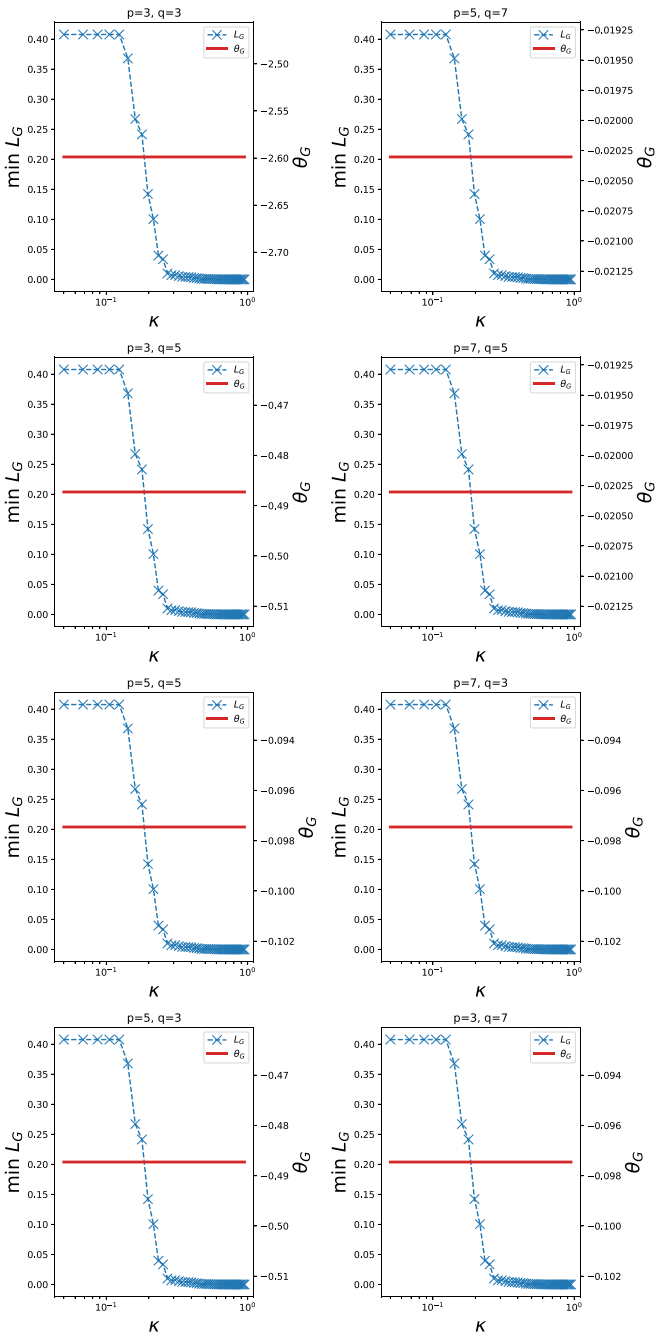


Fig. 13 The effect of κ on minimum L_G. Comparison of theoretical predictions of minimum possible discriminator and generator losses to observed minimum losses when training DCGAN on CIFAR10. The blue cross-dashed lines show the experimental DCGAN results, and the solid red show the theoretical results ϑ_G, ϑ_D. $\sigma_z = 1$ is used and p, q are varied

At this point a problem becomes apparent. Suppose that $q \leq p$, then the variance of the lower-right block is strictly less than that of the off diagonal blocks. If we proceed with the strategy in the main text, there is no way of decomposing the lower-right block as a sum of two independent smaller variance GOEs with one matching the variance of the off diagonal blocks. Similarly, if $q > p$, then the final Hessian involving $L^{(D)}$, $L^{(G)}$ will have lower-variance in the upper-left block than the off-diagonals unless very specific undesirable conditions hold on p, q and σ_z. In either of these cases, we cannot decompose the final Hessian as a sum of a large $N - 2 \times N - 2$ GOE and some smaller GOEs in the upper-left or lower-right blocks. We would therefore have to truly compute the Ginibre averages in the supersymmetric method, which we believe is intractable.

We could complete the complexity calculation via the methods of this paper supposing that the appropriate conditions hold on p, q and σ_z. It would look much the same as the calculation in the main text, though the resulting polynomial for the spectral density would be different. Since this work was completed, the complexity results for bipartite spin glasses were obtained in [64] using an entirely new method developed in the companion paper [40]. Applying this method arguably presents more technical hurdles than the supersymmetric approach to complexity calculations, however it is much more general and can be applied to the above model for any p, q and σ_z.

Appendix D: Extra Plots

This section contains some extra plots to back up the comparisons between our model's predictions and the experimental DCGAN results in Sect. 6.2 (Figs. 10, 11, 12, 13). In particular, we produce versions of the plots in Figs. 8 and 9 but for various values of p and q other than $p = q = 5$. Since $p = q = 5$ is the structurally correct choice for the DCGAN, it is natural to ask if any agreement between theory and experiment is most closely obtained with $p = q = 5$. Figure 13 shows that the model has the same deficiency in κ for all p, q values tested. Figure 12 shows best agreement for $p = q = 5$, $p = 3, q = 7$ and $p = 7, q = 3$, and similarly in Fig. 11. There is perhaps weak evidence that the role of p and q as representing the number of layers in the networks has some merit experimentally.

References

1. Choromanska, A., Henaff, M., Mathieu, M., Arous, G.B., LeCun, Y.: The loss surfaces of multilayer networks In: Artificial Intelligence and Statistics, pp. 192–204 (2015)
2. Auffinger, A., Arous, G.B., Cerny, J.: Random matrices and complexity of spin glasses. Commun. Pure Appl. Math. **66**(2), 165 (2013)
3. Choromanska, A., LeCun, Y., Arous, G.B.: Open problem: The landscape of the loss surfaces of multilayer networks. In: Conference on Learning Theory, pp. 1756–1760 (2015)
4. Papyan, V.: The Full Spectrum of Deepnet Hessians at Scale: Dynamics with SGD Training and Sample Size, arXiv preprint arXiv:1811.07062 (2018)
5. Granziol, D., Garipov, T., Vetrov, D., Zohren, S., Roberts, S., Wilson, A.G.: Towards understanding the true loss surface of deep neural networks using random matrix theory and iterative spectral methods. https://openreview.net/forum?id=H1gza2NtwH. Accessed: 2021-06-15 (2019)
6. Granziol, D.: Beyond Random Matrix Theory for Deep Networks, arXiv preprint arXiv:2006.07721 (2020)
7. Baity-Jesi, M., Sagun, L., Geiger, M., Spigler, S., Arous, G.B., Cammarota, C., LeCun, Y., Wyart, M., Biroli, G.: Comparing dynamics: Deep neural networks versus glassy systems. J. Stat. Mech. Theory Exp. **2019**(12), 124013 (2019)

8. Mannelli, S.S., Krzakala, F., Urbani, P., Zdeborova, L.: Passed & spurious: Descent algorithms and local minima in spiked matrix-tensor models, arXiv preprint arXiv:1902.00139 (2019)
9. Folena, G., Franz, S., Ricci-Tersenghi, F.: Rethinking mean-field glassy dynamics and its relation with the energy landscape: the awkward case of the spherical mixed p-spin model, arXiv preprint arXiv:1903.01421 (2019)
10. Ros, V., Ben Arous, G., Biroli, G., Cammarota, C.: Complex energy landscapes in spiked-tensor and simple glassy models: ruggedness, arrangements of local minima, and phase transitions. Phys. Rev. X (2019). https://doi.org/10.1103/PhysRevX.9.011003
11. Maillard, A., Arous, G.B., Biroli, G.: Landscape Complexity for the Empirical Risk of Generalized Linear Models, arXiv preprint arXiv:1912.02143 (2019)
12. Mannelli, S.S., Biroli, G., Cammarota, C., Krzakala, F.: L. Zdeborová, Who is Afraid of Big Bad Minima? Analysis of gradient-flow in spiked matrix-tensor models. In: Advances in Neural Information Processing Systems, pp. 8676–8686 (2019)
13. Baskerville, N.P., Keating, J.P., Mezzadri, F., Najnudel, J.: The loss surfaces of neural networks with general activation functions. J. Stat. Mech: Theory Exp. **2021**(6), 064001 (2021)
14. Kanter, I., Sompolinsky, H.: Associative recall of memory without errors. Phys. Rev. A **35**(1), 380 (1987)
15. Gardner, E.: The space of interactions in neural network models. J. Phys. A **21**(1), 257 (1988)
16. Engel, A., Van den Broeck, C.: Statistical Mechanics of Learning. Cambridge University Press, Cambridge (2001)
17. Nishimori, H.: Statistical Physics of Spin Glasses and Information Processing: An Introduction, vol. 111. Clarendon Press, Oxford (2001)
18. Goodfellow, I., Bengio, Y., Courville, A., Bengio, Y.: Deep Learning, vol. 1. MIT Press, Cambridge (2016)
19. Conneau, A., Schwenk, H., Barrault, L., Lecun, Y.: 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 (Association for Computational Linguistics, Valencia, Spain), pp. 1107–1116 (2017). https://www.aclweb.org/anthology/E17-1104
20. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. **9**(8), 1735 (1997)
21. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: 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) (Association for Computational Linguistics, Minneapolis, Minnesota), pp. 4171–4186 (2019). https://doi.org/10.18653/v1/N19-1423. https://www.aclweb.org/anthology/N19-1423
22. Radford, A., Narasimhan, K., Salimans, T., Sutskever, I.: Improving language understanding by generative pre-training (2018)
23. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition, In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
24. Pennington, J., Worah, P.: Nonlinear random matrix theory for deep learning, In: Advances in Neural Information Processing Systems, pp. 2637–2646 (2017)
25. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative Adversarial Nets, In: Advances in Neural Information Processing Systems 27, ed. by Z. Ghahramani, M. Welling, C. Cortes, N.D. Lawrence, K.Q. Weinberger Curran Associates, Inc., pp. 2672–2680 (2014). http://papers.nips.cc/paper/5423-generative-adversarial-nets.pdf
26. Radford, A., Metz, L., Chintala, S.: Unsupervised representation learning with deep convolutional generative adversarial networks, arXiv preprint arXiv:1511.06434 (2015)
27. Zhang, H., Goodfellow, I.J., Metaxas, D.N., Odena, A.: Self-attention generative adversarial networks. In: International Conference on Machine Learning, pp. 7354–7363 (2018)
28. Liu, M.Y., Tuzel, O.: Coupled Generative Adversarial Networks, In: Proceedings of the 30th International Conference on Neural Information Processing Systems **29**, pp. 469–477 (2016)
29. Karras, T., Laine, S., Aila, T.: A Style-based generator architecture for generative adversarial networks. In: IEEE Transactions on Pattern Analysis and Machine Intelligence, pp. 1–1 (2020)
30. Mirza, M., Osindero, S.: Conditional Generative Adversarial Nets, arXiv preprint arXiv:1411.1784 (2014)
31. Arjovsky, M., Chintala, S., Bottou, L.: Wasserstein generative adversarial networks, In: International Conference on Machine Learning (PMLR), pp. 214–223 (2017)
32. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: 2017 IEEE International Conference on Computer Vision (ICCV), pp. 2242–2251 (2017)
33. Ros, V., Arous, G.B., Biroli, G., Cammarota, C.: Complex energy landscapes in spiked-tensor and simple glassy models: Ruggedness, arrangements of local minima, and phase transitions. Phys. Rev. X **9**(1), 011003 (2019)

34. Arous, G.B., Mei, S., Montanari, A., Nica, M.: The landscape of the spiked tensor model. Commun. Pure Appl. Math. **72**(11), 2282 (2019)
35. Fyodorov, Y.V.: Complexity of random energy landscapes, glass transition, and absolute value of the spectral determinant of random matrices. Phys. Rev. Lett. **92**(24), 240601 (2004)
36. Fyodorov, Y.V., Williams, I.: Replica symmetry breaking condition exposed by random matrix calculation of landscape complexity. J. Stat. Phys. **129**(5–6), 1081 (2007)
37. Verbaarschot, J.: The supersymmetric method in random matrix theory and applications to QCD. AIP Conf. Proc. (2004). https://doi.org/10.1063/1.1853204
38. Guhr, T., Weidenmüller, H.: Isospin mixing and spectral fluctuation properties. Ann. Phys. **199**(2), 412 (1990)
39. Guhr, T.: Dyson's correlation functions and graded symmetry. J. Math. Phys. **32**(2), 336 (1991)
40. Arous, G.B., Bourgade, P., McKenna, B.: Exponential growth of random determinants beyond invariance, arXiv preprint arXiv:2105.05000 (2021)
41. Adler, R.J., Taylor, J.E.: Random Fields and Geometry. Springer, New York (2009)
42. Efetov, K.: Supermathematics. Cambridge University Press, Cambridge, pp. 8-28 (1996). https://doi.org/10.1017/CBO9780511573057.003
43. Nock, A.: Characteristic polynomials of random matrices and quantum chaotic scattering. Ph.D. thesis, Queen Mary University of London (2017)
44. Guionnet, A., Zeitouni, O., et al.: Concentration of the spectral measure for large matrices. Electron. Commun. Probab. **5**, 119 (2000)
45. Arous, G.B., Dembo, A., Guionnet, A.: Aging of spherical spin glasses. Probab. Theory Relat. Fields **120**(1), 1 (2001)
46. Crisanti, A., Sommers, H.J.: Thouless-Anderson-Palmer approach to the spherical p-spin spin glass model. J. Phys. I **5**(7), 805 (1995)
47. Kurchan, J., Parisi, G., Virasoro, M.A.: Barriers and metastable states as saddle points in the replica approach. J. Phys. I **3**(8), 1819 (1993)
48. Hochreiter, S., Schmidhuber, J.: Flat minima. Neural Comput. **9**(1), 1 (1997). https://doi.org/10.1162/neco.1997.9.1.1
49. Chaudhari, P., Choromanska, A., Soatto, S., LeCun, Y., Baldassi, C., Borgs, C., Chayes, J., Sagun, L., Zecchina, R.: Entropy-SGD: biasing gradient descent into wide valleys. J. Stat. Mech. Theory Exp. (2019). https://doi.org/10.1088/1742-5468/ab39d9
50. Keskar, N.S., Mudigere, D., Nocedal, J., Smelyanskiy, M., Tang, P.T.P.: On large-batch training for deep learning: generalization gap and sharp minima. In: 5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings (OpenReview.net) (2017). https://openreview.net/forum?id=H1oyRlYgg
51. Kleinberg, B., Li, Y., Yuan, Y.: An alternative view: when does SGD escape local minima?, In: International Conference on Machine Learning (PMLR), pp. 2698–2707 (2018)
52. Baldassi, C., Lauditi, C., Malatesta, E.M., Perugini, G., Zecchina, R.: Unveiling the structure of wide flat minima in neural networks, arXiv preprint arXiv:2107.01163 (2021)
53. Baldassi, C., Pittorino, F., Zecchina, R.: Shaping the learning landscape in neural networks around wide flat minima. Proc. Natl. Acad. Sci. U.S.A. **117**(1), 161 (2020)
54. Dinh, L., Pascanu, R., Bengio, S., Bengio, Y.: Sharp minima can generalize for deep nets. In: Proceedings of the 34th International Conference on Machine Learning, Proceedings of Machine Learning Research, vol. 70, ed. by D. Precup, Y.W. Teh (PMLR), Proceedings of Machine Learning Research, vol. 70, pp. 1019–1028 (2017). https://proceedings.mlr.press/v70/dinh17b.html
55. Hoffer, E., Hubara, I., Soudry, D.: Train longer, generalize better: closing the generalization gap in large batch training of neural networks, In: Advances in Neural Information Processing Systems, vol. 30, ed. by I. Guyon, U.V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett (Curran Associates, Inc.), vol. 30 (2017). https://proceedings.neurips.cc/paper/2017/file/a5e0ff62be0b08456fc7f1e88812af3d-Paper.pdf
56. Kawaguchi, K., Kaelbling, L.P., Bengio, Y.: Generalization in deep learning (2020)
57. He, H., Huang, G., Yuan, Y.: Asymmetric valleys: Beyond sharp and flat local minima, arXiv preprint arXiv:1902.00744 (2019)
58. Granziol, D.: Flatness is a False Friend, arXiv preprint arXiv:2006.09091 (2020)
59. Dcgan faces tutorial. https://github.com/pytorch/tutorials/blob/master/beginner_source/dcgan_faces_tutorial.py (2018). Accessed 30 Sept 2020
60. Krizhevsky, A., Hinton, G.: Learning multiple layers of features from tiny images. Tech. Rep. 0, University of Toronto, Toronto, Ontario (2009)
61. Subag, E.: The complexity of spherical p-spin models—a second moment approach. Ann. Probab. **45**(5), 3385 (2017)

62. Auffinger, A., Gold, J.: The number of saddles of the spherical p-spin model, arXiv preprint arXiv:2007.09269v1.q (2020)
63. Arous, G.B., Subag, E., Zeitouni, O.: Geometry and temperature chaos in mixed spherical spin glasses at low temperature: the perturbative regime. Comm. Pure Appl. Math. **73**(8), 1732 (2020)
64. McKenna, B.: Complexity of bipartite spherical spin glasses, arXiv preprint arXiv:2105.05043 (2021)