Solving overparametrized systems of random equations: I. Model and algorithms for approximate solutions

Andrea Montanari * Eliran Subag†

Abstract

Consider the problem of solving a system of equations $\boldsymbol{F}(\boldsymbol{x}) = \boldsymbol{0}$, subject to $\|\boldsymbol{x}\|_2 = 1$, whereby $\boldsymbol{F} : \mathbb{R}^d \to \mathbb{R}^n$ is a random nonlinear map. More precisely, $\boldsymbol{F}(\boldsymbol{x}) = (F_1(\boldsymbol{x}), \dots, F_n(\boldsymbol{x}))$ where the $F_i(\,\cdot\,)$'s are i.i.d. rotationally invariant Gaussian processes. We study this problem under the proportional asymptotics $n, d \to \infty$, $n/d \to \alpha \in [0, 1)$ and establish results about the existence of solutions and polynomial-time algorithms to find them.

First, we establish upper and lower bounds $\alpha_{\rm UB}$, $\alpha_{\rm LB}$ on the threshold for existence of solutions. Namely, if the number of equations per variable satisfies $\alpha < \alpha_{\rm LB}$, then the system admits exact solutions with high probability, while for $\alpha > \alpha_{\rm UB}$, no solutions exist, even in an approximate sense.

We then analyze several algorithms to find solutions: gradient descent, Hessian descent, and a two-phase algorithm. In particular, for Hessian descent and the two-phase algorithm, we characterize their thresholds $\alpha_{\rm HD}$, $\alpha_{\rm TP}$. Namely, for $\alpha < \alpha_{\rm HD}$ (or $\alpha < \alpha_{\rm TP}$) the algorithm finds an approximate solution with high probability, while for $\alpha > \alpha_{\rm HD}$ (respectively $\alpha > \alpha_{\rm TP}$), it does not.

Finally, we compare the theoretical predictions within this model to empirical results obtained with structured systems of nonlinear equations.

1 Introduction

Given data $(\boldsymbol{z}_i, y_i) \in \mathbb{R}^D \times \mathbb{R}$, $i \le n$, and a function class $\mathcal{F} \subseteq \{f : \mathbb{R}^D \to \mathbb{R}\}$, an *interpolator* of the data in $\mathcal{F}$ is any function $f \in \mathcal{F}$ such that $f(\boldsymbol{z}_i) = y_i$ for all $i \le n$. Existence and construction of interpolators are fundamental problems in pure and applied mathematics. For instance, Kirszbraun (Lipschitz) extension theorem completely characterizes the existence of interpolators when $\mathcal{F}$ is the class of L-Lipschitz functions.

Recently, existence and algorithmic construction of interpolators has attracted new attention because of applications to machine learning [BHMM19, BLLT20, ZBH+21, BMR21]. In this setting, the data is normally assumed to be random $\{(\boldsymbol{z}_i, y_i)\}_{i \le n} \sim_{i.i.d.} \mathbb{P}$, and the function class is parametric. Namely, there exists a parametric form $f : \mathbb{R}^D \times \mathbb{R}^d \to \mathbb{R}$ such that $\mathcal{F} = \{f(\,\cdot\,; \boldsymbol{x}) : \ \boldsymbol{x} \in \mathbb{S}^{d-1}\}$. Here $\mathbb{S}^{d-1}$ is the unit sphere in d-dimensions and $\boldsymbol{x}$ is a vector of parameters that parametrizes the function class. The condition $\boldsymbol{x} \in \mathbb{S}^{d-1}$ is a stylized version of typical constraints that are imposed (either implicitly or explicitly) on the parameters' vectors in these applications. The interpolation problem then reduces to

$$\text{Find } \boldsymbol{x} \in \mathbb{S}^{d-1} \text{ such that } f(\boldsymbol{z}_i; \boldsymbol{x}) = y_i \text{ for all } i \le n. \tag{1}$$

*Department of Electrical Engineering and Department of Statistics, Stanford University
†Department of Mathematics, Weizmann Institute of Science

Here a parametric form f is fixed, and the pairs $(\boldsymbol{z}_i, y_i)$ are i.i.d. with common law $\mathbb{P}$.

We can abstract away the parametric form and data, thus leading to the formulation adopted in the present paper.

$$\text{Find } \boldsymbol{x} \in \mathbb{S}^{d-1} \text{ such that } F_i(\boldsymbol{x}) = 0 \text{ for all } i \le n. \tag{2}$$

The original formulation is recovered by choosing the special case $F_i(\,\cdot\,) = y_i - f(\boldsymbol{z}_i; \,\cdot\,)$, i.e. by selecting a special distribution of the functions F_i. Since above we assumed the data $(\boldsymbol{z}_i, y_i)$ to be i.i.d., the F_i's will be i.i.d. random functions $F_i : \mathbb{S}^{d-1} \to \mathbb{R}$. It is convenient to define the mapping $\boldsymbol{F} : \mathbb{R}^d \to \mathbb{R}^n$ via $\boldsymbol{F}(\boldsymbol{x}) := (F_1(\boldsymbol{x}), \dots, F_n(\boldsymbol{x}))$.

In this paper, we study this problem under a simple distribution for the random functions F_i. Namely, we assume the F_i's to be i.i.d. centered Gaussian processes with

$$\mathbb{E}\big[F_i(\boldsymbol{x}^1)F_j(\boldsymbol{x}^2)\big] = \delta_{ij}\xi\big(\langle \boldsymbol{x}^1, \boldsymbol{x}^2\rangle\big)\,. \tag{3}$$

Assuming this specific covariance structure is equivalent to the following two assumptions: (i) F_i and F_j are independent and identically distributed for $i \neq j$; (ii) F_i is invariant (in distribution) under rotations.

The fact that the covariance is positive semidefinite implies $\xi(t) = \sum_{k\ge 0}\xi_k t^k$, with $\xi_k \ge 0$ for all k [Sch42]. We will impose the random function $\boldsymbol{F} : \boldsymbol{x} \mapsto \boldsymbol{F}(\boldsymbol{x})$ to be smooth, by imposing $\xi(1+\varepsilon) < \infty$ for some $\varepsilon > 0$ ($\sum_{k\ge 0}\xi_k k^{2\ell+\varepsilon} < \infty$ is sufficient for $\boldsymbol{F} \in C^\ell(\mathbb{S}^{d-1})$ by the Kolmogorov-Centsov theorem).

This model is closely related to the mixed p-spin model from spin glass theory, and indeed, each coordinate F_i is an independent copy of the so-called 'mixed p-spin Hamiltonian' [CS92, CS95, Sub17b]. The case of multiple equations $n \asymp d$, with energy function $\|\boldsymbol{F}(\boldsymbol{x})\|_2^2/2$ was recently introduced by Urbani [Urb23, KU23] as a model for confluent tissues. In particular, [Urb23] derives the phase diagram of this model using the replica method from spin glass physics. Our focus (and approach) will be different from the one of these works. In [Sub23], one of the authors used the second moment method to prove that, when $\xi(0) = 0$, the Hausdorff volume (or number) of solutions concentrates provided $n \le d - 1$.

We will refer to the model defined above as the 'Nonlinear random Equations Model' (NEM). The random function $\boldsymbol{F}$ is naturally extended to the unit ball $\mathsf{B}^d(1)$, by positing the same covariance of Eq. (3). In particular, for a fixed $\boldsymbol{x} \in \mathbb{R}^d$, $\mathbb{E}[\|\boldsymbol{F}(\boldsymbol{x})\|^2] = n\xi(\|\boldsymbol{x}\|)$. We define the set of ε-approximate solutions by

$$\mathsf{Sol}_{n,d}(\varepsilon) := \Big\{\boldsymbol{x} \in \mathbb{S}^{d-1} : \ \|\boldsymbol{F}(\boldsymbol{x})\|_2^2 \le n\xi(1)\cdot\varepsilon\Big\}\,. \tag{4}$$

A number of natural questions arise within this model:

- **Q1** Do exact solutions exist with high probability, i.e. is $\mathsf{Sol}_{n,d}(0)$ non-empty?
- **Q2** Do approximate solution exist, i.e. is $\mathsf{Sol}_{n,d}(\varepsilon)$ non-empty for some $\varepsilon > 0$?
- **Q3** Can we find these solutions (either exact or approximate ones) in polynomial time?

Questions **Q1** and **Q2** can be addressed (albeit non-rigorously) using the replica method [Urb23].

Here we will focus on the problem of efficiently finding approximate solutions in the high-dimensional limit $n, d \to \infty$, with $n/d \to \alpha \in [0, 1)$. Hence α is the asymptotic number of equations per unknown. Since $\alpha < 1$ the system of non-linear equations $\boldsymbol{F}(\boldsymbol{x})$ is overparametrized.

We present the following results, illustrated in Figure 1 which refers to the special case $\xi(q) = \xi_0 + q^p$, $p \in \{3, 7, 11, 15\}$.

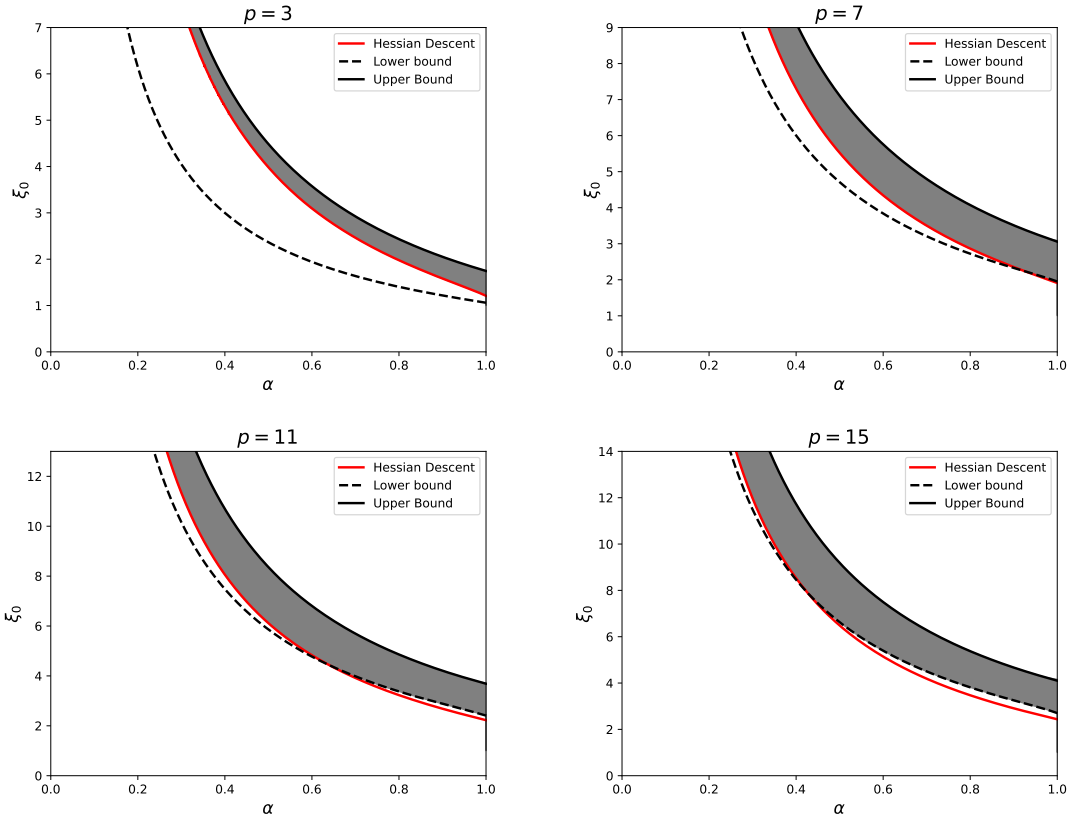

Figure 1: Phase diagram for $\xi(q) = \xi_0 + q^p$, $p \in \{3, 7, 11, 15\}$. Upper (solid) and lower (dashed) black lines are the upper and lower bounds on the threshold for existence of solutions $\alpha_{\rm UB}$ and $\alpha_{\rm LB}$ obtained respectively by Gaussian comparison and second moment method. Red line: The threshold $\alpha_{\rm HD}$ below which the Hessian descent algorithm finds solutions with high probability.

Lower bound. We determine a lower bound threshold $\alpha_{\rm LB} = \alpha_{\rm LB}(\xi)$ such that, if $\alpha < \alpha_{\rm LB}$ then with high probability the system admits exact solutions. Namely, for $\alpha < \alpha_{\rm LB}$, $\mathsf{Sol}_{n,d}(0) \neq \emptyset$ with probability converging to one as $n, d \to \infty$, with $n/d \to \alpha$. The lower bound is based on a second-moment calculation.

In the examples of Figure 1, this lower bound is traced as the dashed black line.

Upper bound. We determine an upper bound threshold $\alpha_{\rm UB} = \alpha_{\rm UB}(\xi)$ such that, if $\alpha > \alpha_{\rm UB}$ then with high probability the system does not admit approximate solutions. Namely, for any $\alpha > \alpha_{\rm UB}$ there exists $\varepsilon > 0$ such that $\mathsf{Sol}_{n,d}(\varepsilon) = \emptyset$ with probability converging to one as $n, d \to \infty$, with $n/d \to \alpha$. Our upper bound is proved using Gaussian comparison inequalities.

In Figure 1, this corresponds to the upper continuous black line.

Gradient descent. We formulate the problem of finding a solution as an optimization problem, with cost function $H(\boldsymbol{x}) := \|\boldsymbol{F}(\boldsymbol{x})\|_2^2/2$. We analyze gradient descent with respect to this cost, and establish a lower bound $\alpha_{\rm GD} = \alpha_{\rm GD}(\xi)$ such that, for $\alpha < \alpha_{\rm GD}$, gradient flow converges to an exact solution.

Our analysis of gradient descent uses techniques recently developed to analyze overparametrized neural networks [DZPS18, COB19, BMR21]. Because of its ubiquity, this approach provides a useful benchmark for the more precise analysis outlined below (which yields better guarantees in terms of overparametrization ratio α.)

Hessian descent. We introduce an Hessian descent algorithm, that instead of optimizing the first order approximation to the cost function $H(\boldsymbol{x})$, follows the best direction according to its Hessian. For this case, we characterize precisely its threshold $\alpha_{\rm HD}(\xi)$. For $\alpha < \alpha_{\rm HD}$ the algorithm converges to an ε-approximate solution, with ε arbitrarily small, with high probability (with respect to the random realization of $\boldsymbol{F}$). For $\alpha > \alpha_{\rm HD}$ it only converges to an ε_0-approximate solution for some $\varepsilon_0(\alpha)$ bounded away from 0. The threshold $\alpha_{\rm HD}$ is reported as a red line in Figure 1.

For a broad set of random Gaussian functions $\boldsymbol{F}$, the guarantee we obtain for Hessian descent are significantly superior to the more standard ones for gradient descent.

Two-phase algorithm. In general the Hessian descent algorithm is not optimal because it disregards almost entirely gradient information. In particular, it is not optimal when $\xi'(0) \neq 0$ (and hence $\nabla H(\boldsymbol{0})$ is bounded away from $\boldsymbol{0}$). This is related to the fact that the set of solution is not centered around $\boldsymbol{0}$ in this case. We describe a two-phase algorithm that is well suited to these cases. In a first phase we use approximate message passing (AMP) to construct a point $\boldsymbol{m}_* \in \mathsf{B}^d(1)$, with $\nabla H(\boldsymbol{m}_*) \approx c\boldsymbol{m}_*$. In the second phase we run Hessian descent in the hyperplane orthogonal to $\boldsymbol{m}_*$.

We characterize the threshold $\alpha_{\rm TP}$ below which the two-phase algorithm finds an approximate solution.

By comparing the second moment lower bound on the existence of solutions, and the analysis of various efficient algorithms, we identify regimes in which solutions exist with high probability, but we do not now of any polynomial-time algorithm to find them. Further, we expect the two-phase algorithm to be essentially optimal among polynomial-time algorithms. We will provide evidence towards this expectation in a forthcoming publication [MS23a].

The five set of results outlined above are presented, respectively, in Sections 2.1, 2.2, 3, 4, 5. Finally, Section 6 discusses, on the basis of numerical simulations, possible generalizations of the present work to other probabilistic models for the equations F_i.

Definitions and notations

The Gaussian random function $\boldsymbol{F} : \mathbb{R}^d \to \mathbb{R}^n$ defined in the previous section can be constructed explicitly by setting

$$\begin{aligned} F_i(\boldsymbol{x}) &:= \sum_{k\ge 0} \sqrt{\xi_k} \sum_{j_1,\dots,j_k=1}^{d} G^{(k)}_{i,j_1\dots j_k} x_{j_1}\cdots x_{j_k} \\ &= \sqrt{\xi_0}\, G^{(0)}_i + \sqrt{\xi_1} \sum_{j=1}^{d} G^{(1)}_{i,j} x_j + \dots , \end{aligned} \tag{5}$$

where the random coefficients $(\boldsymbol{G}^{(k)})_{k\ge 0} := (G^{(k)}_{i,j_1\dots j_k})_{k\ge 0, i\le n, j_1,\dots,j_k\le d} \sim_{i.i.d.} \mathsf{N}(0,1)$ are i.i.d. Gaussian random variables. The relation with the previous definition is given by the coefficients ξ_k.

Given a symmetric matrix $\boldsymbol{A} \in \mathbb{R}^{n\times n}$, we denote by $\lambda_1(\boldsymbol{A}) \ge \lambda_2(\boldsymbol{A}) \ge \cdots \ge \lambda_n(\boldsymbol{A})$ its eigenvalues in decreasing order. For a general matrix $\boldsymbol{M} \in \mathbb{R}^{m\times n}$, $m \le n$, $\sigma_1(\boldsymbol{M}) \ge \sigma_2(\boldsymbol{M}) \ge \cdots \ge \sigma_m(\boldsymbol{M}) \ge 0$ denote its singular values and $\|\boldsymbol{M}\|_{\rm op}$ its operator norm.

We let $\mathsf{B}^d(r) := \{\boldsymbol{x} \in \mathbb{R}^d : \|\boldsymbol{x}\| \le r\}$ denote the ball of radius r in $\mathbb{R}^d$, with $\mathsf{B}^d := \mathsf{B}^r(1)$.

Throughout, we will write $\boldsymbol{W} \sim \mathsf{GOE}(N)$ if $\boldsymbol{W} = \boldsymbol{W}^{\mathsf{T}}$ and $(W_{ij})_{i\le j\le N}$ are independent with $W_{ii} \sim \mathsf{N}(0,2)$, $W_{ij} \sim \mathsf{N}(0,1)$ for $i < j$. We write $\boldsymbol{Z} \sim \mathsf{GOE}(M,N)$ if $(Z_{ij})_{i\le M, j\le N}$ are independent with $Z_{ij} \sim \mathsf{N}(0,1)$.

2 Existence of solutions

In this section we derive upper and lower bounds for the maximum number of equations per variable α, below which solutions do exist. The lower bound $\alpha_{\rm LB}$ (Section 2.1) is based on the second moment method and guarantees the existence of exact solutions for $\alpha < \alpha_{\rm LB}$. The upper bound $\alpha_{\rm UB}$ (Section 2.2) is based on Gaussian comparison inequalities and implies instead that even approximate solutions do not exist for $\alpha > \alpha_{\rm UB}$.

It will be useful to keep in mind the following elementary fact that often allows us to focus on $\mathbb{E}\min_{\boldsymbol{x}\in\mathbb{S}^{d-1}} \|\boldsymbol{F}(\boldsymbol{x})\|_2$.

Remark 2.1. For each $i \le n$, and each $\boldsymbol{x} \in \mathbb{S}^{d-1}$ the function $(G^{(k)}_{i,j_1,\dots,j_k}) \mapsto F_i(\boldsymbol{x})$ is Lipschitz continuous (in ℓ_2), with Lipschitz norm bounded by $\sqrt{\xi(1)}$ (recall that $\xi(1) = \sum_k \xi_k$). As a consequence $(G^{(k)}_{i,j_1,\dots,j_k}) \mapsto \min_{\boldsymbol{x}\in\mathbb{S}^{d-1}} \|\boldsymbol{F}(\boldsymbol{x})\|_2/\sqrt{n}$ is also Lipschitz hence it concentrates:

$$\mathbb{P}\Big(\big| \min_{\boldsymbol{x}\in\mathbb{S}^{d-1}} \|\boldsymbol{F}(\boldsymbol{x})\|_2 - \mathbb{E}\min_{\boldsymbol{x}\in\mathbb{S}^{d-1}} \|\boldsymbol{F}(\boldsymbol{x})\|_2 \big| \ge t \Big) \le 2\, e^{-t^2/(cn)} . \tag{6}$$

2.1 Lower bound

Theorem 1. *Assume $\xi'(0) = \xi''(0) = 0$ and define $\Psi(\,\cdot\,;\alpha,\xi) : [0,1] \to \mathbb{R}$ via*

$$\Psi(r;\alpha,\xi) := \frac{1}{2}\log(1-r^2) - \frac{\alpha}{2}\log\left(1 - \Big(\frac{\xi(r)-\xi(0)}{\xi(1)-\xi(0)}\Big)^2\right) - \frac{\alpha\xi(0)}{\xi(1)+\xi(r)-2\xi(0)} + \frac{\alpha\xi(0)}{\xi(1)-\xi(0)} . \tag{7}$$

Further define

$$\alpha_{\rm LB}(\xi) := \inf\Big\{ \alpha \ge 0 : \sup_{r\in[0,1]} \Psi(r;\alpha,\xi) > 0 \Big\} . \tag{8}$$

If $\alpha < \alpha_{\rm LB}(\xi)$ *then* $\mathsf{Sol}_{n,d}(0) \neq \emptyset$ *with probability converging to one as* $n, d \to \infty$ *with* $n/d \to \alpha \in (0,1)$.

Proof technique. The proof of this result is based on the second moment method. Namely, let $\boldsymbol{F}_{>0}(\boldsymbol{x}) := \boldsymbol{F}(\boldsymbol{x}) - \boldsymbol{F}(\boldsymbol{0})$, and note that $\boldsymbol{F}_{>0}(\,\cdot\,)$ is independent of $\boldsymbol{F}(\boldsymbol{0}) \sim \mathsf{N}(\boldsymbol{0}, \xi(0)\boldsymbol{I}_n)$ and distributed as the original process, with $\xi(q)$ replaced by $\xi_{>0}(q) = \xi(q) - \xi(0)$.

Consider the modified set of solutions

$$\mathsf{Sol}_{n,d}(\boldsymbol{u}; \varepsilon) := \left\{ \boldsymbol{x} \in \mathbb{S}^{d-1} : \ \|\boldsymbol{F}_{>0}(\boldsymbol{x}) - \boldsymbol{u}\|_2^2 \le n\xi(1) \cdot \varepsilon \right\}. \tag{9}$$

We prove that, for $\alpha < \alpha_{\rm LB}$ and $\delta_d \ge 0$ any deterministic sequence such that $\delta_d \to 0$, we have

$$\lim_{n,d\to\infty} \sup_{\|\boldsymbol{u}\|_2/\sqrt{n} \in [\xi_0^{1/2} - \delta_d, \xi_0^{1/2} + \delta_d]} \mathbb{P}(\mathsf{Sol}_{n,d}(\boldsymbol{u}; 0) = \emptyset) = 0\,. \tag{10}$$

This of course implies the claim of the theorem since $\|\boldsymbol{F}(\boldsymbol{0})\|_2/\sqrt{n}$ concentrates around $\xi_0^{1/2}$.

In order to upper bound the probability of $\mathsf{Sol}_{n,d}(\boldsymbol{u}; 0) = \emptyset$, as per Eq. (10), we introduce the random variable

$$\mathrm{V}(\boldsymbol{u}) := \mathrm{Vol}_{d-n-1}\big(\mathsf{Sol}_{n,d}(\boldsymbol{u}; 0)\big)\,, \tag{11}$$

where Vol_i is the Hausdorff measure of dimension i, or the counting measure when $i = 0$.

We use Kac-Rice formula [Kac43, Ric45] to compute the first two moments of $\mathrm{V}(\boldsymbol{u})$ and obtain, by a second moment argument, a sufficient condition for it to concentrate around the mean. For $\boldsymbol{u} = \boldsymbol{0}$ this calculation was carried out in [Sub23]. Here we generalize the argument to the case $\boldsymbol{u} \neq 0$. We present this calculation in Appendix B. □

In the case $\xi(t) = \xi_0 + t^p$ (writing, with an abuse of notation, $\alpha_{\rm LB}(\xi_0, p)$ instead of $\alpha_{\rm LB}(\xi)$), we have the large-p asymptotics (cf. Appendix C)

$$\alpha_{\rm LB}(\xi_0, p) = \frac{\log p}{\xi_0} \cdot \big(1 + o_p(1)\big)\,. \tag{12}$$

(This should be interpreted as $\alpha_{\rm LB}(\xi_0 = \gamma_0 \log p, p) = \gamma_0^{-1}(1 + o_p(1))$ for any $\gamma_0 > 1$.)

2.2 Upper bound

We will prove two upper bounds $\alpha_{\rm UB}^{(1)}$ and $\alpha_{\rm UB}^{(2)}$ on the satisfiability threshold. The first one holds for general ξ and has the advantage of being quite simple, while we prove the second only for the 'pure' model $\xi(q) = \xi_0 + q^p$.

Define

$$E_\star(\xi) := \lim_{d\to\infty} \frac{1}{\sqrt{d}} \mathbb{E} \max_{\boldsymbol{x} \in \mathbb{S}^{d-1}} F_1(\boldsymbol{x})\,. \tag{13}$$

The limit is known to exist and is given by the Parisi formula [AC17, JT17], which we recall below.

Given $\gamma : [0,1) \to \mathbb{R}_{\ge 0}$ and $L \ge \int_0^1 \gamma(s)\mathrm{d}s$, we define

$$\mathsf{P}(\gamma, L) = \frac{1}{2} \int_0^1 \left(\xi''(t)\Gamma(t) + \frac{1}{\Gamma(t)} \right) \mathrm{d}t\,, \tag{14}$$

$$\Gamma(t) := L - \int_0^t \gamma(s)\mathrm{d}s\,. \tag{15}$$

Equivalently, we can view P as a function of $\Gamma : [0,1] \to \mathbb{R}_{\ge 0}$ which is continuous and non-increasing. We then have

$$E_\star(\xi) := \inf\left\{\mathsf{P}(\gamma, L) : \ \gamma \text{ non-decreasing} \quad L \ge \int_0^1 \gamma(t)\,\mathrm{d}t \ \right\}. \tag{16}$$

Proposition 2.1. *Assume that $\xi(0) > 0$ and define $\xi_{>0}(t) := \xi(t) - \xi(0)$ and*

$$\alpha^{(1)}_{\rm UB}(\xi) := \frac{E_\star(\xi_{>0})^2}{\xi(0)}\,. \tag{17}$$

If $\alpha > \alpha^{(1)}_{\rm UB}$ then there exists $\varepsilon > 0$ such that, for $n/d \to \alpha$:

$$\lim_{n,d\to\infty} \mathbb{P}\big(\mathsf{Sol}_{n,d}(\varepsilon) = \emptyset\big) = 1\,.$$

In fact this holds for any $\varepsilon < \varepsilon_0 := (\xi(0)/\xi(1))(1 - \sqrt{\alpha^{(1)}_{\rm UB}/\alpha})^2_+$.

Proof. As in the previous section, we write $\boldsymbol{F}(\boldsymbol{x}) = \boldsymbol{F}(\boldsymbol{0}) + \boldsymbol{F}_{>0}(\boldsymbol{x})$ and note that the two summands are independent. By Remark 2.1 (applied to $\boldsymbol{F}_{>0}$) and Eq. (13)

$$\operatorname*{p\text{-}lim}_{n,d\to\infty} \frac{1}{\sqrt{d}} \max_{\boldsymbol{x}\in\mathbb{S}^{d-1}} \left|\frac{\langle \boldsymbol{F}(\boldsymbol{0}), \boldsymbol{F}_{>0}(\boldsymbol{x})\rangle}{\|\boldsymbol{F}(\boldsymbol{0})\|_2}\right| = E_\star(\xi)\,. \tag{18}$$

Therefore,

$$\begin{aligned}\frac{1}{\sqrt{n}} \min_{\boldsymbol{x}\in\mathbb{S}^{d-1}} \|\boldsymbol{F}(\boldsymbol{x})\|_2 &\ge \frac{1}{\sqrt{n}}\|\boldsymbol{F}(\boldsymbol{0})\|_2 - \frac{1}{\sqrt{n}} \max_{\boldsymbol{x}\in\mathbb{S}^{d-1}} \left|\frac{\langle \boldsymbol{F}(\boldsymbol{0}), \boldsymbol{F}_{>0}(\boldsymbol{x})\rangle}{\|\boldsymbol{F}(\boldsymbol{0})\|_2}\right| &(19)\\ &\ge \sqrt{\xi(0)} - \frac{1}{\sqrt{\alpha}} E_\star(\xi) - o_P(1)\,, &(20)\end{aligned}$$

where $o_P(1)$ denotes a term which converges in probability to 0 as $n, d \to \infty$, which implies the claim. □

We next obtain a bound for pure models $\xi(t) = \xi_0 + t^p$. For $E \ge 2\sqrt{(p-1)/p}$, define

$$\Theta_p(E) := \frac{1}{2}\log(p-1) - \frac{p-2}{p-1}\frac{E^2}{4} - \sqrt{\frac{p}{p-1}}\frac{E}{4}\sqrt{\frac{p}{p-1}E^2 - 4} + \log\left(\sqrt{\frac{p}{p-1}\frac{E^2}{4} - 1} + \sqrt{\frac{p}{p-1}}\frac{E}{2}\right).$$

The meaning of this function is established in [CS95] (at a heuristic level) and in [ABA13, Auf13, Sub17a, SZ21] (rigorously), and it is worth reminding it here. Consider the process $F_{>0,1}(\,\cdot\,)$ (the first coordinate of $\boldsymbol{F}_{>0}(\boldsymbol{x}) = \boldsymbol{F}(\boldsymbol{x}) - \boldsymbol{F}(\boldsymbol{0})$), which is a Gaussian process with covariance $\mathbb{E}[F_{>0,1}(\boldsymbol{x}^1)F_{>0,1}(\boldsymbol{x}^2)] = \langle \boldsymbol{x}^1, \boldsymbol{x}^2\rangle^p$. Then, the number of local maxima $\boldsymbol{x}$ of this process with $F_{>0,1}(\boldsymbol{x}) \approx E\sqrt{d}$ concentrates around $\exp(d\Theta_p(E))$. In particular, the function $E \mapsto \Theta_p(E)$ is monotone decreasing for $E > 2\sqrt{(p-1)/p}$ and vanishes at $E = E_\star(\xi(t) = t^p)$. With an abuse of notation, we will write $E_\star(p) = E_\star(\xi(t) = t^p)$.

For $c \ge 0$, define

$$\begin{aligned}\varphi_1(c,p) &= \sup_{t,s\ge 0}\Big\{c(E_\star(p) + t + s) - \frac{1}{2}\frac{s^2}{\xi(0)} + \Theta_p(E_\star(p) + t)\Big\},\\ \varphi_2(c,\alpha) &= \sup_{0<t\le 1}\Big\{-c\sqrt{\alpha\xi(1)}t - \alpha\frac{t^2-1}{2} + \alpha\log t\Big\}\end{aligned}$$

$$= -c\sqrt{\alpha\xi(1)}t_* - \alpha\frac{t_*^2-1}{2} + \alpha\log t_*,$$

where $t_* = \frac{1}{2}(\sqrt{c^2\xi(1)/\alpha+4} - c\sqrt{\xi(1)}/\sqrt{\alpha})$. Note that $\varphi_2(c,\alpha)$ is decreasing in $\alpha > 0$.

Theorem 2. *Define*

$$\alpha_{\rm UB}^{(2)}(\xi_0,p) := \inf\left\{\alpha \ge 0 : \ \inf_{c>0}\left[\varphi_1(c,p) + \varphi_2(c,\alpha) - \frac{1}{2}c^2(1+\xi_0)\right] < 0\right\}.$$

For the pure model $\xi(q) = \xi_0 + q^p$ with $p \ge 2$, if $\alpha > \alpha_{\rm UB}^{(2)}(\xi_0,p)$, then there exists $\varepsilon > 0$ such that

$$\lim_{n,d\to\infty} \mathbb{P}\big(\mathsf{Sol}_{n,d}(\varepsilon) = \emptyset\big) = 1\,.$$

The proof of this result is presented in Appendix D.

3 Gradient descent

In this section we study the classical projected gradient descent algorithm on the sphere $\mathbb{S}^{d-1}$. We use the cost function $H(\boldsymbol{x}) = \|\boldsymbol{F}(\boldsymbol{x})\|_2^2/2$. This algorithm can be thought of as a discretization of the gradient flow dynamics

$$\dot{\boldsymbol{x}}(t) = -\mathsf{P}_{\mathsf{T},\boldsymbol{x}(t)}\nabla H(\boldsymbol{x}(t))\,, \tag{21}$$

where $\mathsf{P}_{\mathsf{T},\boldsymbol{x}}$ is the projection onto the tangent space to the sphere of radius $\|\boldsymbol{x}\|_2$ at $\boldsymbol{x}$, namely[1]

$$\mathsf{P}_{\mathsf{T},\boldsymbol{x}} := \boldsymbol{I}_d - \frac{\boldsymbol{x}\boldsymbol{x}^{\mathsf{T}}}{\|\boldsymbol{x}\|_2^2}\,. \tag{22}$$

Appendix E states results about gradient flow as well.

In gradient descent, at iteration k we take a step along the direction $-\mathsf{P}_{\mathsf{T},\boldsymbol{x}^k}\nabla H(\boldsymbol{x}^k)$, with stepsize η, and then project back on the sphere $\mathbb{S}^{d-1}$. The algorithm, is defined by the pseudocode of Algorithm 1.

Algorithm 1: Projected Gradient Descent

Data: Couplings $\{\boldsymbol{G}^{(k)}\}_{k\ge 0}$, stepsize η, number of iterations K
Result: approximate optimizer $\boldsymbol{x} \in \mathbb{S}^{d-1}$
Initialize $\boldsymbol{x}^0 \sim \mathrm{Unif}(\mathbb{S}^{d-1})$;
for $k \in \{0,\dots,K-1\}$ **do**
  $\boldsymbol{z}^{k+1} = \boldsymbol{x}^k - \eta\mathsf{P}_{\mathsf{T},\boldsymbol{x}^k}\nabla H(\boldsymbol{x}^k)$;
  $\boldsymbol{x}^{k+1} = \boldsymbol{z}^{k+1}/\|\boldsymbol{z}^{k+1}\|_2$;
end
return $\boldsymbol{x}^K$

Our analysis of this algorithm is based on a technique that became recently popular to analyze overparametrized neural networks in the so called 'neural tangent' or 'lazy' regime. A few pointers to this literature include [DZPS18, COB19, OS20, ADH+19, AZLL19, BMR21]. We deliberately

[1] In Eq. (21) we have $\|\boldsymbol{x}\|_2 = 1$ but we define $\mathsf{P}_{\mathsf{T},\boldsymbol{x}}$ more generally for future reference. By convention, we set $\boldsymbol{v}/\|\boldsymbol{v}\|_2 = \boldsymbol{0}$ if $\boldsymbol{v} = \boldsymbol{0}$.

follow this type of analysis as it provides a useful comparison point for the sharper techniques in the next sections.

Below, for $\boldsymbol{x} \in \mathbb{S}^{d-1}$, $\boldsymbol{U}_{\boldsymbol{x}}$ is an orthonormal basis for $\mathsf{T}_{\boldsymbol{x}}$ the orthogonal space to $\boldsymbol{x}$. Further, for $\boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathbb{R}^d$, define $\boldsymbol{U}_{\boldsymbol{x}_1,\boldsymbol{x}_2} := \boldsymbol{R}_{\boldsymbol{x}_1,\boldsymbol{x}_2}\boldsymbol{U}_{\boldsymbol{x}_1}$, where $\boldsymbol{R}_{\boldsymbol{x}_1,\boldsymbol{x}_2}$ is the rotation that keeps unchanged the space orthogonal to $\boldsymbol{x}_1$, $\boldsymbol{x}_2$, and maps $\boldsymbol{x}_1$ to $\boldsymbol{x}_2$. Finally, let $\boldsymbol{DF}(\boldsymbol{x}) \in \mathbb{R}^{n\times d}$ be the Jacobian of $\boldsymbol{F}$ at $\boldsymbol{x}$.

Lemma 3.1. *Consider the Gradient Descent Algorithm 1, and fix $\varepsilon_0 \in (0, 1/2)$. Let $\lambda_0 := \sigma_{\min}(\boldsymbol{DF}(\boldsymbol{x}^0)|_{\mathsf{T},\boldsymbol{x}^0})$, and J_n, L_n, M_n be given by*

$$J_n := \sup_{\boldsymbol{x}_1\neq\boldsymbol{x}_2\in\mathsf{B}^d(1+\varepsilon_0)} \frac{\|\boldsymbol{DF}(\boldsymbol{x}_1) - \boldsymbol{DF}(\boldsymbol{x}_2)\|_{\rm op}}{\|\boldsymbol{x}_1 - \boldsymbol{x}_2\|_2}, \tag{23}$$

$$L_n := \sup_{\boldsymbol{x}_1\neq\boldsymbol{x}_2\in\Omega} \frac{\|\boldsymbol{DF}(\boldsymbol{x}_1)\boldsymbol{U}_{x_1} - \boldsymbol{DF}(\boldsymbol{x}_2)\boldsymbol{U}_{\boldsymbol{x}_1,\boldsymbol{x}_2}\|_{\rm op}}{\|\boldsymbol{x}_1 - \boldsymbol{x}_2\|_2}, \tag{24}$$

$$M_n := \sup_{\boldsymbol{x}\in\mathsf{B}^d(1+\varepsilon_0)} \|\boldsymbol{DF}(\boldsymbol{x})\|_{\rm op}, \tag{25}$$

and assume they satisfy

$$L_n\|\boldsymbol{F}(\boldsymbol{x}^0)\|_2 < \frac{\lambda_0^2}{16}, \tag{26}$$

Further assume the step size η to be such that

$$\eta \le \min\left\{\frac{\varepsilon_0}{\max_{\boldsymbol{x}\in\mathbb{S}^{d-1}}\|\mathsf{P}_{\mathsf{T}}\boldsymbol{DF}(\boldsymbol{x})^{\mathsf{T}}\boldsymbol{F}(\boldsymbol{x})\|_2}; \frac{1}{M_n^2}; \frac{1}{10\sqrt{n}(M_n+J_n)\max_{\boldsymbol{x}\in\mathbb{S}^{d-1}}\|\boldsymbol{F}(\boldsymbol{x})\|_2}\right\}. \tag{27}$$

Then, for all $k \ge 0$

$$\|\boldsymbol{F}(\boldsymbol{x}^k)\|_2^2 \le \|\boldsymbol{F}(\boldsymbol{x}^0)\|_2^2\, e^{-\lambda_0^2\eta k/8}. \tag{28}$$

The proof of this statement is presented in Appendix E.

By evaluating the conditions in this statement, we obtain the following.

Theorem 3. *Consider the Gradient Descent Algorithm 1, with $\boldsymbol{F}$ the Gaussian process defined in Section 1, and initialization $\boldsymbol{x}^0$ independent of $\boldsymbol{F}$. Assume $n, d \to \infty$ with $n/d \to \alpha \in [0, 1)$. Define, for c_0 a sufficiently small absolute constant and*

$$\underline{\alpha}_{\rm GD}(\xi) := \frac{c_0\xi'(1)^2}{\xi''(1)\xi(1)\big(\log(\xi'''(1)/\xi''(1)) \vee 1\big)}. \tag{29}$$

If $\alpha < \underline{\alpha}_{\rm GD}(\xi)$, and $\eta < 1/(C_1 d)$ with C_1 a suitable constant depending on ξ, then the following happens with high probability. For all $k \ge 1$,

$$\|\boldsymbol{F}(\boldsymbol{x}^k)\|_2^2 \le 2n\xi(1)\,\exp\Big(-\frac{\xi'(1)\eta}{16}\big(\sqrt{d}-\sqrt{n}\big)^2\cdot k\Big). \tag{30}$$

Again, we refer to Appendix E for a proof.

Considering our running example $\xi(t) = \xi_0 + t^p$, $\underline{\alpha}_{\rm GD}(\xi)$ is equivalent, up to constants, to

$$\underline{\alpha}_{\rm GD}(\xi_0, p) := \frac{c_1}{\xi_0 \log p}. \tag{31}$$

It is instructive to compare this result with the lower bound on the threshold for existence of solutions, see Eq. (12). Roughly speaking, for large p, and $1/\log p\alpha\xi_0 \ll \log p$ we know that solutions exist with high probability, but the approach developed here does not guarantee that we can find them. We will see that this gap shrinks using the methods in next sections.

4 Hessian descent

We next consider the Hessian descent algorithm first introduced in [Sub21] to optimize the spherical spin glass Hamiltonian. We do not expect this algorithm to be optimal in general, and in particular not so unless $\xi'(0) = 0$ (i.e. $\boldsymbol{F}(\boldsymbol{x})$ does not contain terms linear in $\boldsymbol{x}$). In the next section we describe an extension that covers this case as well, but we think it is useful to first introduce Hessian descent in a simpler setting.

The Hessian descent algorithm presents two important differences with respect to gradient descent. First, we extend the objective function $H(\boldsymbol{x})$ to the unit ball $\mathsf{B}^d(1)$: iterates are initialized at $\boldsymbol{x}^0 = \boldsymbol{0}$, and after k iterations we have $\|\boldsymbol{x}^k\|_2^2 = k\delta$. Second, we disregard gradient information at $\boldsymbol{x}^k$ and instead optimize the Hessian contribution to the cost function $H(\boldsymbol{x})$.

The pseudocode for Hessian descent is specified in Algorithm 2 below.

Algorithm 2: Hessian Descent

Data: Couplings $\{\boldsymbol{G}^{(k)}\}_{k\ge 0}$, stepsize δ, with $1/\delta \in \mathbb{N}$
Result: Approximate optimizer $\boldsymbol{x}^{\mathrm{HD}} \in \mathbb{S}^{d-1}$
Initialize $\boldsymbol{x}^0 = \boldsymbol{0}$, $\boldsymbol{x}^1 \sim \sqrt{\delta}\cdot \mathrm{Unif}(\mathbb{S}^{d-1})$;
for $k \in \{1,\dots,K := 1/\delta - 1\}$ **do**
  Compute $\boldsymbol{v} = \boldsymbol{v}(\boldsymbol{x}^k) \in \mathsf{T}_{\boldsymbol{x}^k}$ such that $\|\boldsymbol{v}\|_2 = 1$ and

$$\langle \boldsymbol{v}, \nabla^2 H(\boldsymbol{x}^k)\boldsymbol{v}\rangle \le \lambda_{\min}(\nabla^2 H(\boldsymbol{x}^k)|_{\mathsf{T},\boldsymbol{x}^k}) + d\delta\,; \tag{32}$$

  $s_k := \mathrm{sign}(\langle \boldsymbol{v}(\boldsymbol{x}^k), \nabla H(\boldsymbol{x}^k)\rangle)$;
  $\boldsymbol{x}^{k+1} = \boldsymbol{x}^k - s_k\sqrt{\delta}\,\boldsymbol{v}(\boldsymbol{x}^k)$;
end
return $\boldsymbol{x}^{\mathrm{HD}} = \boldsymbol{x}^K$;

The next theorem bounds the value achieved by the Hessian descent algorithm. (We formally state an upper bound on the value achieved, but we expect the bound to be tight.)

Theorem 4. *For $\alpha \in (0,1)$, $a, b \in \mathbb{R}_{\ge 0}$, define*

$$Q(m;\alpha,a,b) := -\frac{1}{m} + \frac{\alpha b}{1+bm} - a^2 m\,, \tag{33}$$

$$z_*(\alpha,a,b) := -\sup_{m>0} Q(m;\alpha,a,b)\,. \tag{34}$$

Let $u(\,\cdot\,;\alpha,\xi) : [0,1] \to \mathbb{R}$ be the unique solution of the ordinary differential equation

$$\frac{\mathrm{d}u}{\mathrm{d}t}(t) = -\frac{1}{2\alpha} z_*\big(\alpha; \sqrt{2\alpha u(t)\xi''(t)}, \xi'(t)\big)\,, \qquad u(0) = \frac{1}{2}\xi(0)\,. \tag{35}$$

Then there exists constants $C_0 = C_0(\alpha,\xi)$, $\delta_0 = \delta_0(\alpha,\xi) > 0$ depending uniquely on α, ξ such that the Hessian descent algorithm, with stepsize parameter $\delta \le \delta_0$, outputs $\boldsymbol{x}^{\mathrm{HD}} \in \mathbb{S}^{d-1}$ such that, with probability converging to one as $n, d \to \infty$ ($n/d \to \alpha$)

$$\frac{1}{2n}\big\|\boldsymbol{F}(\boldsymbol{x}^{\mathrm{HD}})\big\|_2^2 \le u(1;\alpha,\xi) + C_0\delta\,. \tag{36}$$

Further, the algorithm has complexity at most $(C_0\chi_{n,d}/\delta)\log(1/\delta)$, where $\chi_{n,d}$ is the complexity of a single matrix vector multiplication by $\nabla^2 H(\boldsymbol{x})$ at a query point $\boldsymbol{x} \in \mathsf{B}^d(1)$.

The proof of this theorem is presented in Appendix F.

Remark 4.1. Theorem 4 implies that the Hessian descent algorithm achieves an approximate solution (in the sense that p-lim$_{n,d\to\infty} \|\boldsymbol{F}(\boldsymbol{x}^{\rm HD})\|_2^2/n = 0$) provided $\alpha < \alpha_{\rm HD}(\xi)$, where

$$\alpha_{\rm HD}(\xi) := \inf\left\{\alpha \ge 0 :\ u(1;\alpha,\xi) = 0\right\}. \tag{37}$$

Remark 4.2. The complexity $\chi_{n,d}$ of matrix-vector multiplication by $\nabla^2 H(\boldsymbol{x})$ at a query point $\boldsymbol{x}$ depends on the details of the computation model in use. In a model in which sums and products in $\mathbb{R}$ can be carried out in $O(1)$ time, if $\xi_k = 0$ for all $k > k_{\max}$ (i.e. $\boldsymbol{F}$ is a polynomial), then $\chi_{n,d} = O(d^{k_{\max}})$.

If $\xi_k \neq 0$ for infinitely many k (i.e. $\boldsymbol{F}$ is not a polynomial), we can truncate it at a large level $k_{\max}$, and hence approximate matrix-vector multiplication by using the truncated Hessian. Similarly, if only integer operations are allowed, we can use finite-precision approximations of the entries of $\nabla^2 H(\boldsymbol{x})$. It is easy to show that these modifications do not change the claim of Theorem 4.

It is clear that establishing Theorem 4 requires to analyze the eigen-structure of the Hessian $\nabla^2 H(\boldsymbol{x})$ at a point $\boldsymbol{x} \in \mathbb{R}^d$. In particular, the following simple lemma characterizes the joint distribution of $H(\boldsymbol{x})$, $\nabla H(\boldsymbol{x})$, $\nabla^2 H(\boldsymbol{x})$. In synthesis, the restriction of the Hessian on the tangent space is a linear combination of a Wigner and an independent Wishart matrix, with coefficients that depend on the energy level.

Lemma 4.1. *For a fixed $\boldsymbol{x} \in \mathbb{R}^d$ with $\|\boldsymbol{x}\|_2^2 = q$, we have $\boldsymbol{F}(\boldsymbol{x}) = \sqrt{\xi(q)}\,\boldsymbol{g}$, $\boldsymbol{D}\boldsymbol{F}(\boldsymbol{x})\boldsymbol{U}_{\boldsymbol{x}} = \sqrt{\xi'(q)}\,\boldsymbol{Z}$, $\boldsymbol{U}_{\boldsymbol{x}}^{\mathsf{T}}\nabla^2 F_\ell(\boldsymbol{x})\boldsymbol{U}_{\boldsymbol{x}} = \sqrt{\xi''(q)}\,\boldsymbol{W}_\ell$, where $\boldsymbol{g}, (\boldsymbol{W}_\ell)_{\ell\le n}, \boldsymbol{Z}$ are mutually independent with*

$$\boldsymbol{g} \sim \mathsf{N}(0, \boldsymbol{I}_n)\,, \quad \boldsymbol{W}_\ell \sim \mathsf{GOE}(d-1)\,, \quad \boldsymbol{Z} \sim \mathsf{GOE}(n, d-1)\,. \tag{38}$$

As a consequence, letting $\boldsymbol{\mathcal{H}}(\boldsymbol{x}) := \boldsymbol{U}_{\boldsymbol{x}}^{\mathsf{T}}\nabla^2 H(\boldsymbol{x})\boldsymbol{U}_{\boldsymbol{x}}$ be the restriction of the Hessian to the tangent space, we have

$$\boldsymbol{\mathcal{H}}(\boldsymbol{x}) = \sqrt{\xi(q)\xi''(q)}\,\|\boldsymbol{g}\|_2 \boldsymbol{W} + \xi'(q)\,\boldsymbol{Z}^{\mathsf{T}}\boldsymbol{Z}\,, \tag{39}$$

$$H(\boldsymbol{x}) = \frac{1}{2}\xi(q)\|\boldsymbol{g}\|_2^2\,, \tag{40}$$

where $(\boldsymbol{g}, \boldsymbol{W}, \boldsymbol{Z}) \sim \mathsf{N}(0, \boldsymbol{I}_n) \otimes \mathsf{GOE}(d-1) \otimes \mathsf{GOE}(n, d-1)$.

This lemma suggests to estimate the energy decrease at step k of Hessian descent, by computing the minimum eigenvalue of $\lambda_{\min}(\boldsymbol{\mathcal{H}}(\boldsymbol{x}))$ at a point $\boldsymbol{x}$ with $\|\boldsymbol{x}\|_2^2 = k\delta$ and $H(\boldsymbol{x})/n = u$. If $\boldsymbol{x}$ is a point independent of the Gaussian process $\boldsymbol{F}(\,\cdot\,)$, $\lambda_{\min}(\boldsymbol{\mathcal{H}}(\boldsymbol{x}))$ turns out to concentrate around $-z_\#(t = k\delta)d$ where $z_\#(t) := z_*(\alpha; \sqrt{2\alpha u\xi''(t)}, \xi'(t))$. By summing this energy decrement over $k \in \{, \dots, \lfloor 1/\delta \rfloor - 1\}$ and letting δ be small, this calculation yields the value $u(1;\alpha,\xi)$ of Theorem 4.

At first sight, such a derivation might seem incorrect because $\boldsymbol{x}^k$ is not independent of $\boldsymbol{\mathcal{H}}(\boldsymbol{x}^k)$. However, the fast decay of the probability of upper deviations of the minimum eigenvalue allows to establish the claim nevertheless.

Remark 4.3 (Algorithm complexity)**.** Recall that $\lambda_1(\boldsymbol{A}) \ge \lambda_2(\boldsymbol{A}) \ge \cdots$ denote the eigenvalues of $\boldsymbol{A}$ in decreasing order. Then the proof of Theorem 4 outlined above (together with the fact that the matrix $\boldsymbol{\mathcal{H}}(\boldsymbol{x})$ of Eq. (39) has with high probability $c_0\varepsilon^{3/2}\cdot d$ eigenvalues in $[-z_\# d, (-z_\# + \varepsilon)d]$) implies that condition (32) can be replaced by

$$\langle \boldsymbol{v}, \nabla^2 H(\boldsymbol{x}^k)\boldsymbol{v}\rangle \le (-z_\#(k\delta) + c_*\delta)d\,, \tag{41}$$

which is essentially equivalent. Such a vector $\boldsymbol{v}$ can be computed with $O(\log(1/\delta))$ matrix vector multiplications by $\nabla^2 H(\boldsymbol{x}^k)|_{\mathsf{T},\boldsymbol{x}^k}$, via Chebyshev approximation [Saa11]. Also notice that computing $\lambda_{\min}(\nabla^2 H(\boldsymbol{x}^k)|_{\mathsf{T},\boldsymbol{x}^k})$ is not needed.

Summing up, the total complexity of one step Hessian descent algorithm is $O(\log(1/\delta))$ matrix vector multiplications by the Hessian.

The formula $u(1;\alpha,\xi)$ for the energy achieved by Hessian descent, cf. Theorem 4, is somewhat implicit. The next corollary provides user-friendly upper and lower bounds. Its proof follows immediately from Theorem 4 using the bounds on z_* given in Lemma F.1, part 3.

Corollary 4.2. *Define*

$$u_{\rm LB}(\alpha,\xi) := \frac{1}{2}\left(\sqrt{\xi(0)} - \sqrt{\frac{1}{\alpha}}\int_0^1 \sqrt{\xi''(s)}\right)_+^2, \tag{42}$$

$$u_{\rm UB}(\alpha,\xi) := \frac{1}{2}\left(\sqrt{\xi(0)} - \sqrt{\frac{1-\alpha}{\alpha}}\int_0^1 \sqrt{\xi''(s)}\right)_+^2. \tag{43}$$

Then the energy achieved by Hessian descent satisfies

$$u_{\rm LB}(\alpha,\xi) \le \operatorname*{p-lim}_{d,n\to\infty} \frac{1}{2n}\|\boldsymbol{F}(\boldsymbol{x}^{\rm HD})\|_2^2 \le u_{\rm UB}(\alpha,\xi)\,. \tag{44}$$

In particular, the critical point of the algorithm satisfies

$$\frac{A(\xi)}{1+A(\xi)} \le \alpha_{\rm HD}(\xi) \le A(\xi)\,, \qquad A(\xi) := \left(\int_0^1 \sqrt{\frac{\xi''(t)}{\xi(0)}}\,{\rm d}t\right)^2. \tag{45}$$

For our running example of a pure model $\xi(t) = \xi_0 + t^p$ (denoting the corresponding threshold by $\alpha^{\rm HD}(\xi_0,p)$), the last bounds yields

$$\frac{4(p-1)}{p\xi_0 + 4(p-1)} \le \alpha_{\rm HD}(\xi_0,p) \le \frac{4(p-1)}{p\xi_0}\,. \tag{46}$$

In particular, for large p, we obtain $\alpha_{\rm HD}(\xi_0,p) \ge (4/(4+\xi_0))(1+o_p(1))$. This is substantially better than the guarantee $\underline{\alpha}_{\rm GD}(\xi_0,p) \asymp 1/(\xi_0\log p)$ that we obtained in the previous section for gradient descent, but still far from the maximum value of α in which we know that solutions exist, $\alpha < \alpha_{\rm LB}(\xi_0,p) \asymp (\log p)/\xi_0$, cf. Eq. (12).

5 Two-phase algorithm

By analogy with the case of the spherical p-spin glass model [Sub21], we expect Hessian descent to be optimal (among algorithms with comparable complexity) if $\xi'(0) = 0$. On the other hand, it is easy to construct examples in which it is suboptimal if $\xi'(0) \neq 0$, as discussed in the next remark.

Remark 5.1. Consider the case $\xi(q) = \xi_0 + q$. Namely $\boldsymbol{F}$ is a linear function $\boldsymbol{F}(\boldsymbol{x}) = \sqrt{\xi_0}\boldsymbol{h} + \boldsymbol{G}\boldsymbol{x}$ for $\boldsymbol{h}\in\mathbb{R}^n$, $\boldsymbol{G}\in\mathbb{R}^{n\times d}$ with i.i.d. standard normal entries. In this case the Hessian of $H(\boldsymbol{x})$ is always positive semidefinite and Hessian descent does not yield any improvement over random guessing, namely $H(\boldsymbol{x}^{\rm HD})/n = \xi_0/2 + o_P(1)$.

On the other hand, it is simple to find a solution by linear algebra. More precisely, we can find the minimum norm solution of $\boldsymbol{G}\boldsymbol{x} = -\sqrt{\xi_0}\boldsymbol{h}$, namely $\boldsymbol{x}_0 = -\boldsymbol{G}^{\mathsf{T}}(\boldsymbol{G}\boldsymbol{G}^{\mathsf{T}})^{-1}\sqrt{\xi_0}\boldsymbol{h}$ and add a vector in the null space of $\boldsymbol{G}$ of suitable length to get a solution of unit norm. A simple random matrix calculation yields

$$\|\boldsymbol{x}_0\|_2^2 = \frac{\xi_0\alpha}{1-\alpha} + o_P(1)\,. \tag{47}$$

We thus can efficiently construct a solution $\boldsymbol{x}_* \in \mathbb{S}^{d-1}$ provided $\alpha < 1/(\xi_0+1)$.

From the point of view of optimization, the reason for suboptimality is that Hessian descent does not exploit gradient information in a neighborhood of $\boldsymbol{x} = \boldsymbol{0}$. Spin-glass theory offers a more refined picture. If $\xi'(0) \neq 0$, the set of solutions of the system $\boldsymbol{F}(\boldsymbol{x}) = \boldsymbol{0}$ is centered at $\boldsymbol{m}_*$, with $\|\boldsymbol{m}_*\|_2$ bounded away from 0.

In order to find solutions in this situation, we proceed in two phases: (i) We first find a proxy $\boldsymbol{m}^L$ for $\boldsymbol{m}_*$ using an approximate message passing (AMP) algorithm. (ii) We use the vector $\boldsymbol{m}^L$ as initialization for the next phase. We restrict to the hyperplane orthogonal to this vector and run Hessian descent to optimize the energy $H(\boldsymbol{x})$.

The AMP iteration in the first phase is defined by letting, for $\ell \geq 0$

$$\boldsymbol{h}^{\ell+1} = \frac{1}{\sqrt{n}}\boldsymbol{F}(\boldsymbol{m}^\ell) - \gamma B_\ell \boldsymbol{h}^{\ell-1}\,, \tag{48}$$

$$\boldsymbol{m}^{\ell+1} = \frac{\gamma}{\sqrt{d}}\boldsymbol{D}\boldsymbol{F}(\boldsymbol{m}^\ell)^{\mathsf{T}}\boldsymbol{h}^\ell - \gamma C_\ell \boldsymbol{m}^{\ell-1} - \gamma^2 D_\ell \boldsymbol{m}^{\ell-1}\,. \tag{49}$$

with initialization

$$\boldsymbol{m}^0 = \boldsymbol{m}^{-1} = \boldsymbol{h}^0 = \boldsymbol{0}\,. \tag{50}$$

Further, γ is a (non-random) constant to be fixed in the course of the proof and B_ℓ, C_ℓ, D_ℓ are given by

$$B_\ell := \frac{1}{\sqrt{\alpha}}\xi'\big(\langle \boldsymbol{m}^\ell, \boldsymbol{m}^{\ell-1}\rangle\big)\,, \tag{51}$$

$$C_\ell := \sqrt{\alpha}\,\xi'\big(\langle \boldsymbol{m}^\ell, \boldsymbol{m}^{\ell-1}\rangle\big)\,, \tag{52}$$

$$D_\ell := \xi'\Big(\langle \boldsymbol{m}^\ell, \boldsymbol{m}^{\ell-1}\rangle\Big)\,\langle \boldsymbol{h}^\ell, \boldsymbol{h}^{\ell-1}\rangle\,. \tag{53}$$

The pseudocode the algorithm is given in Algorithm 3.

Algorithm 3: Two-Phase Algorithm

Data: Couplings $\{\boldsymbol{G}^{(k)}\}_{0\le k\le k_{\max}}$, iteration number L, stepsize δ, AMP parameter γ
Result: Approximate optimizer $\boldsymbol{x}^{\rm TP}\in\mathbb{S}^{d-1}$
Initialize $\boldsymbol{m}^0=\boldsymbol{m}^{-1}=\boldsymbol{h}^{-1}=\boldsymbol{0}$;
for $\ell\in\{0,\dots,L-1\}$ **do**
  Compute $\boldsymbol{h}^\ell$, $\boldsymbol{m}^{\ell+1}$ via Eqs. (48), (49);
end
Set $\boldsymbol{x}^0=\boldsymbol{m}^L$, $V_L:=\{\boldsymbol{x}\in\mathbb{R}^d:\langle\boldsymbol{x},\boldsymbol{m}^L\rangle=0\}$;
Set $K=(1-\|\boldsymbol{m}^L\|^2)/\delta$;
for $k\in\{0,\dots,K-1\}$ **do**
  Compute $\boldsymbol{v}=\boldsymbol{v}(\boldsymbol{x}^k)\in\mathsf{T}_{\boldsymbol{x}^k}\cap V_L$ such that $\|\boldsymbol{v}\|_2=1$ and

$$\langle\boldsymbol{v},\nabla^2 H(\boldsymbol{x}^k)\boldsymbol{v}\rangle\le\lambda_{\min}(\nabla^2 H(\boldsymbol{x}^k)|_{\mathsf{T}_{\boldsymbol{x}^k}\cap V_L})+d\delta\,;$$

  Set $s_k:=\mathrm{sign}(\langle\boldsymbol{v}(\boldsymbol{x}^k),\nabla H(\boldsymbol{x}^k)\rangle)$;
  $\boldsymbol{x}^{k+1}=\boldsymbol{x}^k-s_k\sqrt{\delta}\,\boldsymbol{v}(\boldsymbol{x}^k)$;
end
return $\boldsymbol{x}^{\rm TP}=\boldsymbol{x}^K/\|\boldsymbol{x}^K\|_2$;

We begin by stating a theorem that characterizes the first phase of the algorithm.

Theorem 5. *Assume $\xi(0),\xi'(0)>0$, $\alpha\in(0,1)$. Let $q_{\rm RS}(\xi):=\mathrm{argmin}_{q>0}[\xi(q)\xi'(q)/q]$ and $q_0(\alpha)=q_0(\alpha,\xi)$ to be the unique positive solution of $\alpha=q\xi'(q)/\xi(q)$. For $q\in(0,q_{\rm RS}(\xi))$, define $\gamma_*(q,\alpha,\xi)$ via*

$$\gamma_*(q,\alpha,\xi)=-\sqrt{\frac{\underline{q}}{\xi(\underline{q})\xi'(\underline{q})}}\,,\quad \underline{q}:=q\wedge q_0(\alpha)\,. \tag{54}$$

(If $\xi(t)=\xi_0+\xi_1 t$ we set $q_{\rm RS}(\xi)=\infty$.)

Then the first phase of Algorithm 3, with input parameter $\gamma=\gamma_*(q,\alpha,\xi)$, outputs $\boldsymbol{m}^L$ such that

$$\Big|\operatorname*{p-lim}_{n,d\to\infty}\frac{1}{d}\big\|\boldsymbol{m}^L\big\|^2-q\wedge q_0(\alpha)\Big|\le Ce^{-L/C}\,, \tag{55}$$

$$\Big|\operatorname*{p-lim}_{n,d\to\infty}\frac{1}{2n}\big\|\boldsymbol{F}(\boldsymbol{m}^L)\big\|^2-u_{\rm RS}(q,\alpha,\xi)\Big|\le Ce^{-L/C}\,, \tag{56}$$

$$u_{\rm RS}(q,\alpha,\xi):=\frac{1}{2}\Big(\sqrt{\xi(q)}-\sqrt{\frac{1}{\alpha}q\xi'(q)}\Big)_+^2\,. \tag{57}$$

Note that for our problem statement, we should choose $q\in(0,q_{\rm RS}(\xi)\wedge 1)$. However, the statement above makes sense and holds also for $q\ge 1$.

Remark 5.2. Note that $q\mapsto V(q):=\xi(q)\xi'(q)/q$ is strictly convex on $(0,q_{\max})$ with $q_{\max}$ the maximum radius of convergence of ξ (see Appendix G), hence $q_{\rm RS}(\xi)$ is well defined, with $q_{\rm RS}(\xi)=\infty$ if and only if ξ is linear. Further $q\mapsto g(q):=q\xi'(q)/\xi(q)$ is strictly increasing with $g(0)=0$, $\lim_{q\to\infty}g(q)=\infty$. Hence $q_0(\alpha,\xi)$ is well defined as well. In general neither $q_{\rm RS}(\xi)$ nor $q_0(\alpha,\xi)$ need to take values in $[0,1]$.

We use this result together with a generalization of the proof of Theorem 4 to prove the next statement.

Theorem 6. *For $\alpha \in (0,1)$, $a, b \in \mathbb{R}_{\ge 0}$, define $z_*(\alpha, a, b)$ as in Theorem 4. Further, define $q_{\rm RS}(\xi)$, $q_0(\alpha,\xi)$, $\gamma_*(q,\alpha,\xi)$ as in Theorem 5 and set $q_* := q_{\rm RS}(\xi) \wedge q_0(\alpha,\xi)$. Assume $q_* < 1$ (see Remark 5.3 for the case $q_* \ge 1$.)*

Let $u(\,\cdot\,;\alpha,\xi,q) : [0, 1-q] \to \mathbb{R}$ be the unique solution of the ordinary differential equation

$$\frac{\mathrm{d}u}{\mathrm{d}t}(t) = -\frac{1}{2\alpha} z_*\big(\alpha; \sqrt{2\alpha u(t)\xi''(q+t)}, \xi'(q+t)\big)\,, \qquad u(0) = u_{\rm RS}(q,\alpha,\xi)\,. \tag{58}$$

Then for any $\varepsilon > 0$, there exists a constant $\delta_0 = \delta_0(\varepsilon,\alpha,\xi) > 0$ depending uniquely on α, ξ such that the two phase algorithm, with stepsize parameter $\delta \le \delta_0$ and $\gamma = \gamma_*(q_* - \delta_0, \alpha, \xi)$ outputs $\boldsymbol{x}^{\rm TP} \in \mathbb{S}^{d-1}$ such that

$$\operatorname*{p-lim}_{d,n\to\infty} \frac{1}{2n} \big\|\boldsymbol{F}(\boldsymbol{x}^{\rm TP})\big\|_2^2 \le u(1-q_*;\alpha,\xi,q_*) + \varepsilon\,. \tag{59}$$

Further the algorithm has the complexity of $C(\delta)$ matrix vector multiplications by $\nabla^2 H$ and $\boldsymbol{DF}$.

The proofs of this theorem and of Theorem 5 are presented in Appendix G.

Remark 5.3. Theorem 6 only covers the case $q_* < 1$.

If $q_* \ge 1$, the first phase of the algorithm (with input parameter $\gamma = \gamma_*(q=1,\alpha,\xi)$) achieves, by Theorem 5

$$\operatorname*{p-lim}_{n,d\to\infty} \frac{1}{2n} \big\|\boldsymbol{F}(\boldsymbol{m}^L)\big\|^2 = \frac{1}{2}\Big(\sqrt{\xi(1)} - \sqrt{\frac{1}{\alpha} 2\xi'(1)}\Big)_+^2 + O(e^{-L/C})\,. \tag{60}$$

We expect this value cannot be improved cannot be improved by other choices of the algorithm parameters.

6 Interpolating random data: An empirical comparison

Throughout the paper, we studied the problem of solving the system of nonlinear equations $\boldsymbol{F}(\boldsymbol{W}) = \boldsymbol{0}$ with respect to unknowns $\boldsymbol{W} \in \mathbb{S}^{d-1}$, when $\boldsymbol{F}$ is a Gaussian process. (Throughout this section, the unknowns will be denoted by $\boldsymbol{W}$ to match the applied literature.)

It is natural to wonder whether the theory developed in this setting can provide any guidance towards understanding more complex cases in which the functions $F_1, \dots, F_n$ are random but non-Gaussian.

In this section we consider the interpolation problem introduced in Section 1 and describe a simulation study comparing the Gaussian theory of the previous section to empirical results. More precisely, we consider the problem of interpolating random data $(y_i, \boldsymbol{z}_i) \in \{+1, -1\} \times \mathbb{R}^D$ using a two-layer neural network with weights $\boldsymbol{W}$.

The setup for these simulations is defined in Section 6.1. Section 6.2 investigates the robustness of our results with respect to various choices in the simulations. Finally, in Section 6.3 we compare empirical results in the interpolation problem with the Gaussian theory of the previous sections. Namely, we compare empirical results with theoretical ones within a Gaussian model that approximately matches the covariance structure of the interpolation model.

In general, we observe a gap between the Gaussian theory and the neural network model. However, the two appear to be in rough qualitative agreement and, in certain cases, surprisingly close to each other.

6.1 Setup and definitions

We are interested in the ability of large neural networks to interpolate completely unstructured (pure noise) data, a phenomenon that has attracted considerable attention over the last few years [ZBH+21, BHMM19, BMR21]. In order to capture the essence of this problem in a simple setting, we assume to be given i.i.d. data $\{(y_i, \boldsymbol{z}_i)\}_{i\le n}$ with $y_i \sim \mathrm{Unif}(\{+1,-1\})$ independent of $\boldsymbol{z}_i \sim \mathsf{N}(\boldsymbol{0}, \boldsymbol{I}_D)$. We consider a two-layer neural network with D inputs and m hidden units:

$$f(\boldsymbol{z}; \boldsymbol{W}) = \frac{a}{\sqrt{m}} \sum_{j=1}^{m} s_i \, \sigma(\langle \boldsymbol{w}_j, \boldsymbol{z}\rangle) \, . \tag{61}$$

Here $a \in \mathbb{R}_{>0}$ is a scale parameter that will be fixed independently of the data. The signs $s_1, \dots, s_m \in \{+1,-1\}$ are also fixed (independent of the data) and uniformly random subject to $\#\{i \in [m] : \ s_i = +1\} = \#\{i \in [m] : \ s_i = -1\} = m/2$ (for simplicity we assume m even). The weights $\boldsymbol{W} = (\boldsymbol{w}_1, \dots, \boldsymbol{w}_m)^{\mathsf{T}} \in \mathbb{R}^{m\times D}$, are instead fit to the data as to minimize the empirical risk

$$\hat{R}_n(\boldsymbol{W}) := \frac{1}{2n} \sum_{i=1}^{n} \big(y_i - f(\boldsymbol{z}_i; \boldsymbol{W})\big)^2 \, , \tag{62}$$

subject to the norm constraints

$$\|\boldsymbol{W}\|_F^2 = \sum_{i=1}^{m} \|\boldsymbol{w}_i\|_2^2 \le m \, . \tag{63}$$

Note that $n \cdot \hat{R}_n(\boldsymbol{W})$ is nothing but the Hamiltonian $H(\boldsymbol{x}) = \|\boldsymbol{F}(\boldsymbol{x})\|_2^2/2$ of the previous pages, with the replacement $\boldsymbol{x} = \boldsymbol{W}/\sqrt{m}$ and $F_i(\boldsymbol{x}) = y_i - f(\boldsymbol{z}_i; \boldsymbol{W})$. We thus identify the dimension of the optimization problem as $d = mD$.

We attempt to minimize the cost $\hat{R}_n(\boldsymbol{W})$ using stochastic gradient descent (SGD). Below are some specifics of our experiments.

Activation function. We use the standard ELU activation:

$$\sigma(x) = \begin{cases} x & \text{if } x \ge 0, \\ e^x - 1 & \text{if } x < 0. \end{cases} \tag{64}$$

Optimization algorithm. We use SGD with batch size $|B(k)| = 4$:

$$\tilde{\boldsymbol{W}}^{k+1} = \boldsymbol{W}^k - \frac{\eta_k}{2} \sum_{i\in B(k)} \nabla_{\boldsymbol{w}} \big(y_i - f(\boldsymbol{z}_i; \boldsymbol{W}^k)\big)^2 \, , \tag{65}$$

$$\boldsymbol{W}^{k+1} = \Big(1 \wedge \frac{\sqrt{m}}{\|\tilde{\boldsymbol{W}}^{k+1}\|_F}\Big) \tilde{\boldsymbol{W}}^{k+1} \, , \tag{66}$$

and stepsize $\eta_k = \mathrm{lr}/(1 + E(k))^{1/2}$, where $E(k)$ is the epoch index. We will vary the learning rate lr and number of epochs.

In order to satisfy the ℓ_2 constraint (63), we project back onto this set at each iteration, as per Eq. (66).

Initialization. We initialize the weights to $\boldsymbol{W} \sim \mathsf{N}(0, \varepsilon^2 \boldsymbol{I}_{mD}/D)$, and use $\varepsilon = 0.03$ in simulations. This corresponds to initializing close to the center of the ball (63), since $\|\boldsymbol{W}\|_F/\sqrt{m} = \varepsilon + o_P(1)$.

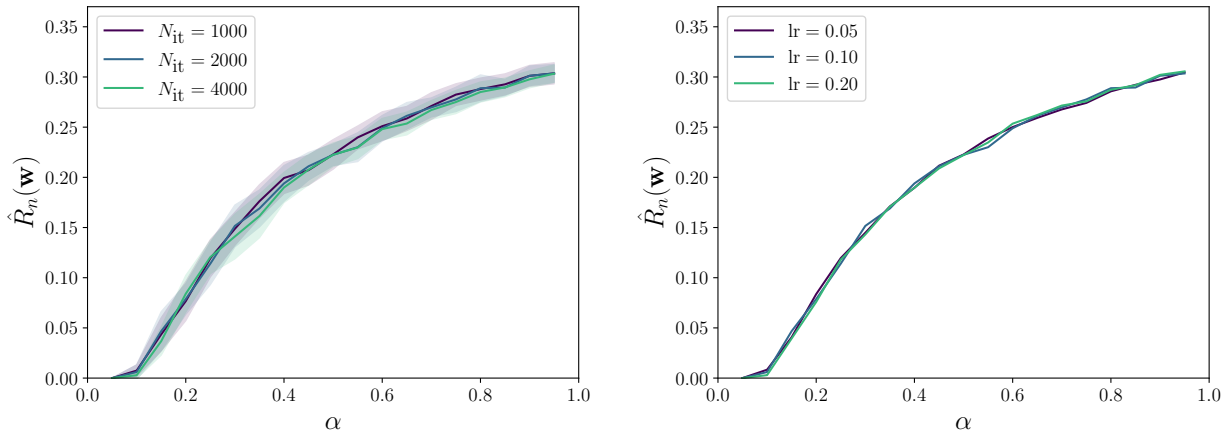

Figure 2: Fitting random binary labels using a 2-layer ELU network (61), using SGD and weights bounded as per Eq. (63), as a function of the number of samples per parameter $n/(mD)$. Left: Dependence of training error on the number of epochs $N_{\rm it}$. Right: Dependence on the learning rate lr.

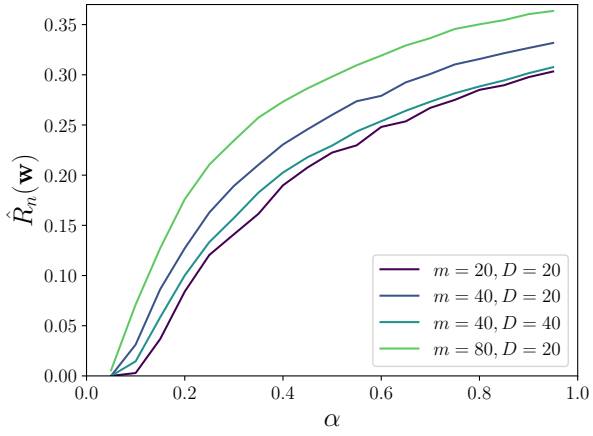

Figure 3: Same as in Fig. 2. Training error as a function of the number of samples per parameter $n/(mD)$. Here we vary the input dimension D and number of neurons m.

6.2 Dependence on the simulations parameters

In Figures 2, 3 we investigated the dependence of training error (energy) achieved over various parameters of the model and training algorithm. All of these figures are obtained by setting $a = 1$, and report the average training error over 20 realizations of the data, and independent runs of the algorithm. When present, bands represent an interval of plus/minus one standard deviation over these 20 realizations.

In more detail, we consider the following settings:

- Figure 2, left frame: we use $m = D = 20$, lr $= 0.1$, and increase the number of epochs $N_{\rm it}$ from 1000 to 4000.
- Figure 2, right frame: we use $m = D = 20$, $N_{\rm it} = 2000$, and vary the learning rate lr between 0.05 and 0.20.

Our theory does not capture dependence on learning rate or number of iterations. In this respect, the empirical observation that this dependence is mild in the current setting is reassuring.

Next, the Gaussian theory only depends on the ratio $\alpha = n/d$, where d is the number of decision variables. In the current setting $d = mD$, and is therefore interesting to investigate the dependence on m, D while keeping $\alpha = n/(mD)$ fixed:

- In Figure 3 we use lr $= 0.1$, $N_{\rm it} = 4000$, and vary the number of neurons m and input dimension D.

We observe that the training error achieved *does* depend on the architecture, and most noticeably on the 'aspect ratio' m/D. On the other hand, this dependence is not as dramatic as one might expect. A fourfold increase in aspect ratio only changes the training error by 20%.

6.3 Comparison with Gaussian theory

We next compare simulation results with the analytical prediction for a Gaussian model with matching covariance (see Section 6.4 for a description of this prediction). More precisely we compute the prediction for the two-phase algorithms of Theorem 6. We expect this to correspond to the optimal energy achieved (asymptotically as $n, d \to \infty$) by a broad class of efficient algorithms. Hence should provide an upper bound on the energy achieved by SGD (and an upper bound on the threshold.)

In Figure 4 we consider number of neurons $m = 20$, and input dimension $D = 20$, and vary $N_{\rm it}$, lr as indicated in legends. We change $a \in \{1, 2, 5\}$, which of course impacts the Gaussian prediction. The agreement is surprisingly good for $a = 1$, and worsens for $a \in \{2, 5\}$. Despite the worse agreement for $a \gtrsim 2$, the Gaussian theory still has the correct qualitative behavior, and appears to be a good starting point to analyze the more complex model (61).

6.4 Covariance matching

We conclude this section by discussing the "covariance matching" used in to compute the theoretical predictions in Fig. 4. We begin by observing that, *in general*, we cannot expect the Gaussian model to capture the behavior of the system of nonlinear equations $y_i - f(\boldsymbol{z}_i; \boldsymbol{W}) = 0$, with $f(\,\cdot\,; \boldsymbol{W})$ given by Eq. (61).

To see this, consider the case of a linear activation $\sigma(x) = x$. Then the system of equations reads

$$\Big(\frac{a}{\sqrt{m}} \sum_{j=1}^{m} s_j \, \boldsymbol{w}_j\Big)^{\mathsf{T}} \boldsymbol{z}_i = y_i \qquad \forall i \le n \, . \tag{67}$$

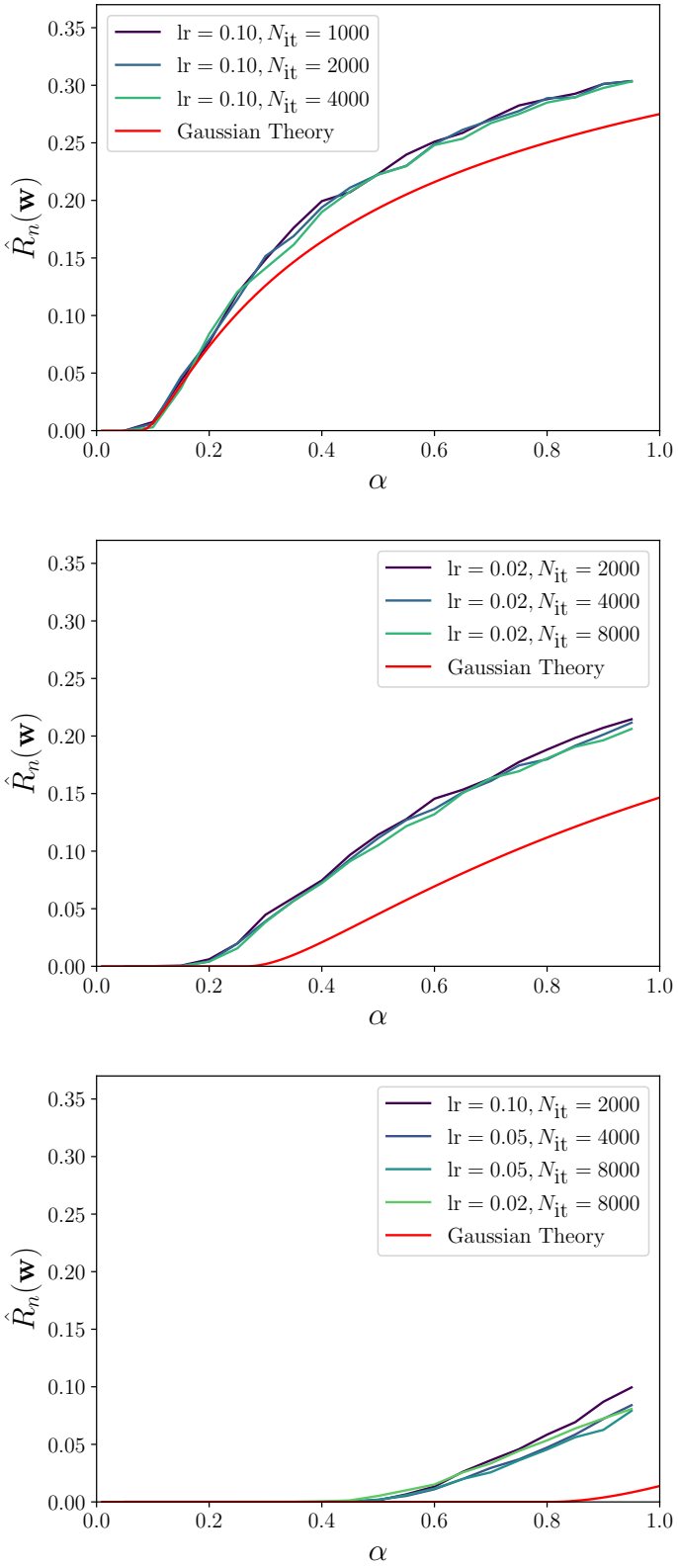

Figure 4: Training error od SGD as a function of the number of samples per parameter $\alpha = n/(mD)$. From top to bottom: $a \in \{1, 2, 5\}$. Red lines report the theoretical prediction of Theorem 6 for the optimal 'two-phase' algorithm.

In other words, the model depends on the weight vectors only through the D-dimensional projection $\boldsymbol{u} := am^{-1/2}\sum_{j=1}^m s_j\,\boldsymbol{w}_j$. In particular, a solution exists if and only if $D \ge n$ (for generic data $(\boldsymbol{z}_i)_{i\le n}$), and the total number of parameters is irrelevant.

In order to take into account the fact that the linear component of σ has a much smaller number of degrees of freedom than in the corresponding Gaussian model, we project out the linear component of activations before matching covariances. Namely, we define $\sigma_\#(x) := \sigma(x) - \sigma_1 \cdot x$, where $\sigma_1 := \mathbb{E}_{G\sim\mathsf{N}(0,1)}[G\sigma(G)]$ is the linear coefficient in the Hermite expansion of σ, and

$$f_\#(\boldsymbol{z};\boldsymbol{W}) = \frac{a}{\sqrt{m}}\sum_{j=1}^m s_j\,\sigma_\#(\langle \boldsymbol{w}_j,\boldsymbol{z}\rangle)\,, \tag{68}$$

$$F_i(\boldsymbol{W}) := y_i - f_\#(\boldsymbol{z};\boldsymbol{W})\,. \tag{69}$$

We now compute the covariance of the process $F_i(\boldsymbol{w})$ with respect to the random data $\boldsymbol{z}$. Letting $C_F(\boldsymbol{W};\boldsymbol{W}') := \mathbb{E}_{\boldsymbol{z}_i,y_i}[F_i(\boldsymbol{W})F_i(\boldsymbol{W}')]$, we have

$$C_F(\boldsymbol{W};\boldsymbol{W}') = 1 + \mathbb{E}_{\boldsymbol{z}}\big\{f_\#(\boldsymbol{z};\boldsymbol{W})f_\#(\boldsymbol{z};\boldsymbol{W}')\big\} \tag{70}$$

$$= 1 + \frac{a^2}{m}\sum_{j,l=1}^m s_j s_l\,\mathbb{E}_{\boldsymbol{z}}\sigma_\#(\langle \boldsymbol{w}_j,\boldsymbol{z}\rangle)\sigma_\#(\langle \boldsymbol{w}'_l,\boldsymbol{z}\rangle)\big\} \tag{71}$$

$$= 1 + \frac{a^2}{m}\sum_{j,k=1}^m s_j s_l\,\widehat{K}\big(\langle \boldsymbol{w}_j,\boldsymbol{w}'_l\rangle, \|\boldsymbol{w}_j\|_2^2, \|\boldsymbol{w}'_l\|_2^2\big)\,, \tag{72}$$

where we introduced the kernel $\widehat{K}$ defined by

$$\widehat{K}(r_{12},r_{11},r_{22}) := \mathbb{E}\big[\sigma_\#(G_1)\sigma_\#(G_2)\big]\,, \tag{73}$$

$$(G_1,G_2)\sim\mathsf{N}(\boldsymbol{0},\boldsymbol{R})\quad \boldsymbol{R} := \begin{pmatrix} r_{11} & r_{12}\\ r_{12} & r_{22}\end{pmatrix}\,. \tag{74}$$

We observe that the covariance (72) is not invariant under the orthogonal group $\mathcal{O}(mD)$ but only under $\mathfrak{S}_{m/2}\otimes\mathfrak{S}_{m/2}\otimes\mathcal{O}(D)$ (with $\mathfrak{S}_{m/2}$ denoting the group of permutations in the space of neurons, for neurons with same s_j, and $\mathcal{O}(D)$ denoting rotations in the space of weights). In order to replace it by an orthogonally invariant covariance (depending only on $\langle \boldsymbol{W},\boldsymbol{W}'\rangle$), we make the following approximations:

1. Since the signs $(s_j)_{j\le m}$ in Eq. (72) are random, we keep only diagonal terms in the sum, and thus replace $C_F(\boldsymbol{W};\boldsymbol{W}')$ by $C_F^{(1)}(\boldsymbol{W};\boldsymbol{W}')$ whereby

$$C_F^{(1)}(\boldsymbol{W};\boldsymbol{W}') = 1 + \frac{a^2}{m}\sum_{j=1}^m \widehat{K}\big(\langle \boldsymbol{w}_j,\boldsymbol{w}'_j\rangle, \|\boldsymbol{w}_j\|_2^2, \|\boldsymbol{w}_j\|_2^2\big)\,.$$

   Notice that this is accurate at a fixed $\boldsymbol{W}$, $\boldsymbol{W}'$ but not uniformly over $\boldsymbol{W}$, $\boldsymbol{W}'$.

2. We approximate the value of $C_F^{(1)}(\boldsymbol{W};\boldsymbol{W}')$ by the one taken by $C_F^{(1)}(\boldsymbol{W};\boldsymbol{W}')$ for a 'typical' pair $\boldsymbol{W}$, $\boldsymbol{W}'$ with a given inner product $\langle \boldsymbol{W},\boldsymbol{W}'\rangle$. More formally define the set

$$\mathcal{W}_{m,n}(q) := \Big\{\boldsymbol{W},\boldsymbol{W}'\in\mathbb{R}^{m\times D}:\ \ \|\boldsymbol{W}\|_F^2 = m, \|\boldsymbol{W}'\|_F^2 = m, \langle \boldsymbol{W},\boldsymbol{W}'\rangle = mq\Big\}\,. \tag{75}$$

   If we draw $(\boldsymbol{W},\boldsymbol{W}')\sim\mathrm{Unif}(\mathcal{W}_{m,n}(q))$, then for a fixed $j\in[m]$, $\|\boldsymbol{w}_j\|_2 = 1+o_P(1)$, $\|\boldsymbol{w}'_j\|_2 = 1+o_P(1)$ and $\langle \boldsymbol{w}_j,\boldsymbol{w}'_j\rangle = q + o_P(1)$. As a consequence, we will have

$$C_F^{(1)}(\boldsymbol{W};\boldsymbol{W}') = 1 + a^2\,\widehat{K}\big(q,1,1\big) + o_P(1)\,. \tag{76}$$

Summarizing, and simplifying notations, the above approximation suggests to use a Gaussian process $\boldsymbol{F}^G$ in $d = mD$ dimensions with

$$\mathbb{E}[F_i^G(\boldsymbol{W})F_j^G(\boldsymbol{W}')] = \xi\big(\langle \boldsymbol{W}, \boldsymbol{W}'\rangle/m\big)\,, \tag{77}$$

$$\xi(q) := 1 + a^2\, K_0(q) := 1 + a^2[K(q) - \sigma_1^2 q]\,, \tag{78}$$

$$K(q) := \mathbb{E}\big[\sigma(G_1)\sigma(G_2)\big]\,, \qquad (G_1, G_2) \sim \mathsf{N}\left(\boldsymbol{0}, \begin{pmatrix} 1 & q \\ q & 1 \end{pmatrix}\right)\,. \tag{79}$$

This is exactly the function $\xi(\,\cdot\,)$ used to compare simulations and theory in the previous sections.

We conclude this section with a warning. *We do not expect the theoretical predictions based on the covariance matching here to be asymptotically exact as $m, D, n \to \infty$.* In particular, as emphasized above, the actual covariance structure is not invariant under $\mathcal{O}(mD)$.

Nevertheless, the rough agreement between theory and empirical results suggests that the Gaussian model (possibly with a more complex covariance structure) might be a useful starting point.

7 Discussion

In this paper, presented some simple upper and lower bounds on the threshold for existence of solutions of a set of overparametrized nonlinear equations $\boldsymbol{F}(\boldsymbol{x}) = \boldsymbol{0}$ when $\boldsymbol{F} : \mathbb{S}^{d-1} \to \mathbb{R}^n$ is a smooth Gaussian process.

Our main focus was to analyze polynomial-time algorithms to construct approximate solutions, i.e. $\boldsymbol{x} \in \mathbb{S}^{d-1}$ such that $\|\boldsymbol{F}(\boldsymbol{x})\|_2^2 \le \varepsilon\mathbb{E}\{\|\boldsymbol{F}(\boldsymbol{x}_0)\|_2^2\}$ (here $\boldsymbol{x}_0 \in \mathbb{S}^{d-1}$ is uniformly random). In particular, we presented a two-phase algorithm which we expect to be near-optimal, and characterized the value it achieves in Section 5. For cases in which $\boldsymbol{DF}(\boldsymbol{0}) = \boldsymbol{0}$ (i.e. $\xi'(0) = 0$), this algorithm reduces to the Hessian descent algorithm of [Sub21].

As shown in Figure 1, there exist cases in which we know that solutions exist but we are not able to find any in polynomial time.

These results naturally suggest a number of interesting questions. Among others: (*i*) Can we provide evidence of hardness for cases in which the algorithms presented here fail? (*ii*) Can we construct exact solutions? (*iii*) What happens when the system of equations is only modestly overparametrized? (*iv*) Can we characterize sharply the conditions under which gradient flow finds solutions?

We plan to report on questions (*i*)-(*iii*) in the near future [MS23a, MS23b].

Acknowledgements

AM was supported by the NSF through award DMS-2031883, the Simons Foundation through Award 814639 for the Collaboration on the Theoretical Foundations of Deep Learning, the NSF grant CCF-2006489 and the ONR grant N00014-18-1-2729. Part of this work was carried out while Andrea Montanari was on partial leave from Stanford and a Chief Scientist at Ndata Inc dba Project N. The present research is unrelated to AM's activity while on leave. ES was supported by the Israel Science Foundation (Grant Agreement No. 2055/21) and a research grant from the Center for Scientific Excellence at the Weizmann Institute of Science. ES is the incumbent of the Skirball Chair in New Scientists.

A Basic estimates

This appendix contains some basic estimates of the process $(\boldsymbol{F}(\boldsymbol{x}))_{\boldsymbol{x}\in\mathbb{R}^d}$, which will be useful in the following. Throughout $(F_\ell)_{\ell\ge 0}$ are i.i.d. Gaussian processes with $\mathbb{E}[F_\ell(\boldsymbol{x})F_\ell(\boldsymbol{y})] = \xi(\langle \boldsymbol{x},\boldsymbol{y}\rangle)$.

We begin by bounding the expected maximum. Notice that the limiting value of this quantity is exactly given by Parisi's formula, cf. Eq. (16). However, we find it convenient to derive some explicit estimates.

Proposition A.1. *There exist absolute constants C_1, C_2 such that (writing $\log_+(t) := \max(1,\log t)$)*

$$C_1\sqrt{\frac{\xi'(1)^2}{\xi(1)\xi''(1)}\log_+\frac{\xi''(1)}{\xi'(1)}} \le \frac{1}{\sqrt{d\xi(1)}}\mathbb{E}\max_{\boldsymbol{x}\in\mathbb{S}^{d-1}} F_1(\boldsymbol{x}) \le C_2\sqrt{\log_+\frac{\xi'(1)}{\xi(1)}}\,. \tag{80}$$

The same bounds hold if the maximum is taken over $\boldsymbol{x}\in\mathsf{B}^d(1)$.

Proof. Given $\boldsymbol{x},\boldsymbol{y}\in\mathbb{S}^{d-1}$, we denote their Euclidean and rescaled canonical distances by r and u respectively

$$r := \|\boldsymbol{x}-\boldsymbol{y}\|_2 = \sqrt{2(1-\langle \boldsymbol{x},\boldsymbol{y}\rangle)}\,, \tag{81}$$

$$u := \sqrt{\frac{1}{\xi(1)}\mathbb{E}\big\{\big[F_1(\boldsymbol{x})-F_1(\boldsymbol{y})\big]^2\big\}} = \sqrt{2\Big(1-\frac{\xi(\langle \boldsymbol{x},\boldsymbol{y}\rangle)}{\xi(1)}\Big)}\,. \tag{82}$$

We denote by $r(u)$ the (strictly increasing) function that maps u to r, namely the unique solution of the equation

$$1-\frac{u^2}{2} = \frac{\xi(1-r^2/2)}{\xi(1)}\,. \tag{83}$$

Since $\xi'(t)\le\xi'(1)$ on $[0,1]$,

$$\frac{\xi(1-r^2/2)}{\xi(1)} \ge 1-\frac{\xi'(1)}{\xi(1)}\cdot\frac{r^2}{2}\,, \tag{84}$$

whence

$$r(u) \ge \sqrt{\frac{\xi(1)}{\xi'(1)}}\,u\,. \tag{85}$$

Let $N_d(r)\le[(10/r)\vee 1]^d$ (respectively, $N_d^{F_1}(r)$) be the covering number of $\mathbb{S}^{d-1}$ w.r.t. the Euclidean distance (respectively, the canonical distance of F_1). Noting the $1/\xi(1)$ factor in (82), we have that

$$N_d^{F_1}(\sqrt{\xi(1)}u) = N_d(r(u)) \le N_d\Big(\sqrt{\frac{\xi(1)}{\xi'(1)}}u\Big) \le \Big(\sqrt{\frac{\xi'(1)}{\xi(1)}}\frac{10}{u}\vee 1\Big)^d.$$

Finally notice that, under the canonical distance, $\mathrm{diam}_{F_1}(\mathbb{S}^{d-1})\le 2\sqrt{\xi(1)}$. By Dudley's inequality, we have

$$\begin{aligned}\mathbb{E}\max_{\boldsymbol{x}\in\mathbb{S}^{d-1}} F_1(\boldsymbol{x}) &\le 24\int_0^{2\sqrt{\xi(1)}}\sqrt{\log N_d^{F_1}(u)}\,\mathrm{d}u\\ &\le 24\sqrt{\xi(1)}\int_0^{2}\sqrt{\log N_d^{F_1}(\sqrt{\xi(1)}u)}\,\mathrm{d}u\end{aligned}$$

$$\le 24\sqrt{d\xi(1)}\int_0^2 \sqrt{\log_+\Big(\sqrt{\frac{\xi'(1)}{\xi(1)}}\cdot\frac{10}{u}\Big)}\,\mathrm{d}u$$
$$\le C\sqrt{d\xi(1)\log_+\frac{\xi'(1)}{\xi(1)}}\,.$$

This proves the desired upper bound.

In order to prove the lower bound, we use Sudakov's inequality. Recall that the r-packing number is lower bounded by the $2r$-covering number. For any $0 < r < 1$, letting $t = r^2/2$, we have

$$\begin{aligned}\mathbb{E}\max_{\boldsymbol{x}\in\mathbb{S}^{d-1}} F_1(\boldsymbol{x}) &\ge C'\sqrt{\big(\xi(1)-\xi(1-(r^2/2)\big)\log N_d(2r)}\\ &\ge C'\sqrt{d\big(\xi(1)-\xi(1-t)\big)\log_+\frac{1}{t}}\\ &\ge C'\sqrt{d\big(\xi'(1)t-\xi''(1)t^2/2\big)\log_+\frac{1}{t}}\,,\end{aligned}$$

where $C' > 0$ is a constant that may change from line to line and where in the last step we used the intermediate value theorem, and the fact that $\xi''(s) \le \xi''(1)$ for $s \le 1$. The desired inequality follows by choosing $t = \xi'(1)/(2\xi''(1))$. □

Remark A.1. Let L be an integer-valued random variable with probability distribution $\mathbb{P}(L = \ell) = \xi_\ell/\xi(1)$. Then the upper and lower bounds of the last proposition are within a constant provided $\mathbb{E}[L^2] \le C\mathbb{E}[L]^2$.

We will next establish upper bounds on the derivatives of the process $F_\ell(\,\cdot\,)$. Before doing it, we state a useful lemma.

Lemma A.2. *Let $(Z(\boldsymbol{x},\boldsymbol{v}) :\ \boldsymbol{x}\in\mathsf{B}^d(1), \boldsymbol{v}\in\mathbb{S}^{d-1})$ be a continuous centered Gaussian process, and assume that, for some $r_0 > 0$, the following holds for all sets $A \subseteq \mathsf{B}^d(1)$:*

$$\operatorname{diam}(A) \le r_0 \ \Rightarrow\ \mathbb{E}[\sup_{\boldsymbol{x}\in A,\boldsymbol{v}\in\mathbb{S}^{d-1}} Z(\boldsymbol{x},\boldsymbol{v})] \le M(r_0)\,. \tag{86}$$

Then there exists a universal constant C such that

$$\mathbb{E}[\sup_{\boldsymbol{x}\in\mathsf{B}^d(1),\boldsymbol{v}\in\mathbb{S}^{d-1}} Z(\boldsymbol{x},\boldsymbol{v})] \le M(r_0) + C\sqrt{dV_*\log_+(1/r_0)}\,, \tag{87}$$

$$V_* := \sup_{\boldsymbol{x}\in\mathsf{B}^d(1),\boldsymbol{v}\in\mathbb{S}^{d-1}} \operatorname{Var}(Z(\boldsymbol{x},\boldsymbol{v}))\,. \tag{88}$$

Proof. Let $(A_\ell :\ \ell \le N)$ be a covering of $\mathsf{B}^d(1)$ with sets of diameter $\operatorname{diam}(A_\ell) \le r_0$, such that $\log N \le Cd\log_+(1/r_0)$. Define

$$Z_\ell := \sup_{\boldsymbol{x}\in A_\ell,\boldsymbol{v}\in\mathbb{S}^{d-1}} Z(\boldsymbol{x},\boldsymbol{v})\,. \tag{89}$$

We then have

$$\mathbb{P}\Big(\sup_{\boldsymbol{x}\in\mathsf{B}^d(1),\boldsymbol{v}\in\mathbb{S}^{d-1}} Z(\boldsymbol{x},\boldsymbol{v}) \ge M(r_0)+t\Big) = \mathbb{P}\Big(\max_{\ell\le N} Z_\ell \ge M(r_0)+t\Big)$$

$$\le N \cdot \max_{\ell \le N} \mathbb{P}\Big(Z_\ell \ge \mathbb{E} Z_\ell + t \Big)$$
$$\le \exp\Big\{ Cd \log_+(1/r_0) - \frac{t^2}{2V_*} \Big\},$$

where the last step follows from the Borell-TIS inequality. This proves the claim. □

Lemma A.3. *Recall that $\mathsf{P}_{\mathsf{T},\boldsymbol{x}}$ denotes the projector onto the tangent space to the sphere of radius $\|\boldsymbol{x}\|_2$ at $\boldsymbol{x}$. Write, as above, $\log_+(t) = \max(\log t; 1)$ Then there exist an absolute constant C such that, letting $\hat{\boldsymbol{x}} := \boldsymbol{x}/\|\boldsymbol{x}\|_2$, we have*

$$\mathbb{E} \max_{\boldsymbol{x} \in \mathsf{B}^d(1)} \|\mathsf{P}_{\mathsf{T},\boldsymbol{x}} \nabla F_1(\boldsymbol{x})\|_2 \le C \sqrt{d \xi'(1) \log_+ \frac{\xi''(1)}{\xi'(1)}}\,, \tag{90}$$

$$\mathbb{E} \max_{\boldsymbol{x} \in \mathsf{B}^d(1)} \big|\langle \hat{\boldsymbol{x}}, \nabla F_1(\boldsymbol{x})\rangle\big| \le C \sqrt{d \xi''(1) \log_+ \frac{\xi'''(1)}{\xi''(1)}}\,,, \tag{91}$$

$$\mathbb{E} \max_{\boldsymbol{x} \in \mathsf{B}^d(1)} \|\mathsf{P}_{\mathsf{T},\boldsymbol{x}} \nabla^2 F_1(\boldsymbol{x}) \mathsf{P}_{\mathsf{T},\boldsymbol{x}}\|_{\rm op} \le C \sqrt{d \xi''(1) \log_+ \frac{\xi^{(3)}(1)}{\xi''(1)}}\,, \tag{92}$$

$$\mathbb{E} \max_{\boldsymbol{x} \in \mathsf{B}^d(1)} \|\mathsf{P}_{\mathsf{T},\boldsymbol{x}} \nabla^2 F_1(\boldsymbol{x}) \hat{\boldsymbol{x}}\|_{\rm op} \le C \sqrt{d \xi^{(3)}(1) \log_+ \frac{\xi^{(4)}(1)}{\xi^{(3)}(1)}}\,, \tag{93}$$

$$\mathbb{E} \max_{\boldsymbol{x} \in \mathsf{B}^d(1)} \langle \hat{\boldsymbol{x}}, \nabla^2 F_1(\boldsymbol{x}) \hat{\boldsymbol{x}}\rangle \le C \sqrt{d \xi^{(4)}(1) \log_+ \frac{\xi^{(5)}(1)}{\xi^{(4)}(1)}}\,. \tag{94}$$

(Here $\xi^{(\ell)}$ denotes the ℓ-th derivative of ξ.)

Proof. All these inequalities are analogous to the upper bound in Proposition A.1. We will repeatedly use Lemma A.2 to restrict ourselves to sets of some small diameter r_0.

Consider Eq. (90). Define the following Gaussian process indexed by $\boldsymbol{x} \in \mathsf{B}^d(1)$, $\boldsymbol{v} \in \mathsf{T}_{\boldsymbol{x}}$, $\|\boldsymbol{v}\|_2 = 1$

$$G_\perp(\boldsymbol{v}, \boldsymbol{x}) := \langle \boldsymbol{v}, \nabla F_1(\boldsymbol{x})\rangle\,. \tag{95}$$

We have

$$\mathbb{E}\big\{ G_\perp(\boldsymbol{v}_1, \boldsymbol{x}_1) G_\perp(\boldsymbol{v}_2, \boldsymbol{x}_2)\big\} = \langle \boldsymbol{v}_1, \boldsymbol{x}_2\rangle \langle \boldsymbol{x}_1, \boldsymbol{v}_2\rangle\, \xi''(\langle \boldsymbol{x}_1, \boldsymbol{x}_2\rangle) + \langle \boldsymbol{v}_1, \boldsymbol{v}_2\rangle\, \xi'(\langle \boldsymbol{x}_1, \boldsymbol{x}_2\rangle)\,. \tag{96}$$

This, and other covariance calculations below, can be checked either by writing the derivatives of F_1 explicitly or using that the order of taking derivatives and expectations can be interchanged. Letting $d_{G,\perp}(\boldsymbol{v}_1, \boldsymbol{x}_1; \boldsymbol{v}_2, \boldsymbol{x}_2)$ denote the canonical distance associated to this process, we have

$$\begin{aligned}
d_{G,\perp}(\boldsymbol{v}_1, \boldsymbol{x}_1; \boldsymbol{v}_2, \boldsymbol{x}_2)^2 &= \xi'(\langle \boldsymbol{x}_1, \boldsymbol{x}_2\rangle) \|\boldsymbol{v}_1 - \boldsymbol{v}_2\|_2^2 + \xi'(\|\boldsymbol{x}_1\|_2^2) - 2\xi'(\langle \boldsymbol{x}_1, \boldsymbol{x}_2\rangle) + \xi'(\|\boldsymbol{x}_2\|_2^2) \\
&\quad + (\langle \boldsymbol{v}_1, \boldsymbol{x}_2\rangle - \langle \boldsymbol{v}_2, \boldsymbol{x}_1\rangle)^2 \xi''(\langle \boldsymbol{x}_1, \boldsymbol{x}_2\rangle) \\
&\quad - \langle \boldsymbol{v}_1, \boldsymbol{x}_2\rangle^2 \xi''(\langle \boldsymbol{x}_1, \boldsymbol{x}_2\rangle) - \langle \boldsymbol{v}_1, \boldsymbol{x}_2\rangle^2 \xi''(\langle \boldsymbol{x}_1, \boldsymbol{x}_2\rangle) \\
&\le \xi'(1) \|\boldsymbol{v}_1 - \boldsymbol{v}_2\|_2^2 + \xi'(\|\boldsymbol{x}_1\|_2^2) - 2\xi'(\langle \boldsymbol{x}_1, \boldsymbol{x}_2\rangle) + \xi'(\|\boldsymbol{x}_2\|_2^2) \\
&\quad + \xi''(1) \|\boldsymbol{x}_1 - \boldsymbol{x}_2\|^2 \|\boldsymbol{v}_1 - \boldsymbol{v}_2\|^2 \\
&\le 2\xi'(1) \|\boldsymbol{v}_1 - \boldsymbol{v}_2\|_2^2 + \xi'(\|\boldsymbol{x}_1\|_2^2) - 2\xi'(\langle \boldsymbol{x}_1, \boldsymbol{x}_2\rangle) + \xi'(\|\boldsymbol{x}_2\|_2^2)\,,
\end{aligned}$$

where the last inequality follows for $\|\boldsymbol{x}_1-\boldsymbol{x}_2\|_2 \le r_0 := \xi'(1)/\xi''(1)$. Therefore, for any set $A \subseteq \mathsf{B}^d(1)$, $\mathrm{diam}(A) \le r_0$, using the Sudakov-Fernique inequality we have

$$\mathbb{E}\max_{\boldsymbol{x}\in A}\max_{\boldsymbol{v}\in\mathsf{T}_{\boldsymbol{x}},\|\boldsymbol{v}\|_2=1} G_\perp(\boldsymbol{v},\boldsymbol{x}) \le C\mathbb{E}\max_{\|\boldsymbol{v}\|_2=1} G_1(\boldsymbol{v}) + \mathbb{E}\max_{\|\boldsymbol{x}\|_2\le 1} G_2(\boldsymbol{x})\,, \tag{97}$$

where G_1 is a process with canonical distance $\xi'(1)\|\boldsymbol{v}_1-\boldsymbol{v}_2\|_2^2$ and G_2 is a process with canonical distance $\xi'(\|\boldsymbol{x}_1\|_2^2) - 2\xi'(\langle \boldsymbol{x}_1,\boldsymbol{x}_2\rangle) + \xi'(\|\boldsymbol{x}_2\|_2^2)$. By applying the previous proposition to G_2 (with ξ replaced by ξ'), we obtain

$$\mathbb{E}\max_{\boldsymbol{x}\in A}\max_{\boldsymbol{v}\in\mathsf{T}_{\boldsymbol{x}},\|\boldsymbol{v}\|_2=1} G_\perp(\boldsymbol{v},\boldsymbol{x}) \le C\sqrt{s\xi'(1)} + C\sqrt{d\xi'(1)\log_+\frac{\xi''(1)}{\xi'(1)}}\,. \tag{98}$$

Finally, using Lemma A.2, we see that

$$\mathbb{E}\max_{\boldsymbol{x}\in\mathsf{B}^d(1)}\max_{\boldsymbol{v}\in\mathsf{T}_{\boldsymbol{x}},\|\boldsymbol{v}\|_2=1} G_\perp(\boldsymbol{v},\boldsymbol{x}) \le C\sqrt{d\xi'(1)\log_+\frac{\xi''(1)}{\xi'(1)}} + C\sqrt{d\xi'(1)\log_+(1/r_0)}\,, \tag{99}$$

which yields the bound of Eq. (90).

Next consider Eq. (91). Recalling the representation (5), we have

$$\langle \hat{\boldsymbol{x}}, \nabla F_1(\boldsymbol{x})\rangle = \frac{1}{\|\boldsymbol{x}\|_2}\sum_{k\ge 0} k\sqrt{\xi_k}\sum_{j_1,\dots,j_k=1}^{d} G^{(k)}_{1,j_1\dots j_k} x_{j_1}\cdots x_{j_k}\,, \tag{100}$$

and therefore

$$\mathbb{E}\{\langle \hat{\boldsymbol{x}}_1,\nabla F_1(\boldsymbol{x}_1)\rangle\langle \hat{\boldsymbol{x}}_2,\nabla F_1(\boldsymbol{x}_2)\rangle\} = \frac{1}{\|\boldsymbol{x}_1\|_2\|\boldsymbol{x}_2\|_2}[\langle \boldsymbol{x}_1,\boldsymbol{x}_2\rangle^2\xi''(\langle \boldsymbol{x}_1,\boldsymbol{x}_2\rangle) + \langle \boldsymbol{x}_1,\boldsymbol{x}_2\rangle\xi'(\langle \boldsymbol{x}_1,\boldsymbol{x}_2\rangle)]\,. \tag{101}$$

In other words, $\langle \hat{\boldsymbol{x}}, \nabla F_1(\boldsymbol{x})\rangle$ has the same structure as $F_1(\boldsymbol{x})$ provided we replace ξ by $\xi''+\xi'$. Therefore the claim follows by argument in Proposition A.1.

The bounds (92), (93), (94) follow from similar arguments, and we limit ourselves to defining the relevant Gaussian process and computing its covariance:

- For Eq. (92), we define $H_\perp(\boldsymbol{v},\boldsymbol{x})$ indexed by $\boldsymbol{x}\in\mathsf{B}^d(1)$ and $\boldsymbol{v}\in\mathsf{T}_{\boldsymbol{x}}$ via

$$H_\perp(\boldsymbol{v},\boldsymbol{x}) := \langle \boldsymbol{v}, \nabla^2 F_1(\boldsymbol{x})\boldsymbol{v}\rangle\,. \tag{102}$$

  Its covariance is (denoting by $\xi^{(\ell)}$ the ℓ-th derivative of ξ and letting $q := \langle \boldsymbol{x}_1,\boldsymbol{x}_2\rangle$)

$$\begin{aligned}\mathbb{E}\big\{H_\perp(\boldsymbol{v}_1,\boldsymbol{x}_1)H_\perp(\boldsymbol{v}_2,\boldsymbol{x}_2)\big\} =& \xi^{(4)}(q)\,\langle \boldsymbol{x}_1,\boldsymbol{v}_2\rangle^2\langle \boldsymbol{v}_1,\boldsymbol{x}_2\rangle^2 + 4\,\xi^{(3)}(q)\langle \boldsymbol{x}_1,\boldsymbol{v}_2\rangle\langle \boldsymbol{v}_1,\boldsymbol{x}_2\rangle\langle \boldsymbol{v}_1,\boldsymbol{v}_2\rangle\\ &+ 2\,\xi^{(2)}(q)q^2\langle \boldsymbol{v}_1,\boldsymbol{v}_2\rangle^2\,.\end{aligned}$$

  The associated canonical distance is given by

$$d_{H,\perp}(\boldsymbol{v}_1,\boldsymbol{x}_1;\boldsymbol{v}_2,\boldsymbol{x}_2)^2 = \Delta_{1,2} + 2\xi^{(2)}(\|\boldsymbol{x}_1\|_2^2) - 4\xi^{(2)}(\langle \boldsymbol{x}_1,\boldsymbol{x}_2\rangle) + 2\xi^{(2)}(\|\boldsymbol{x}_2\|_2^2)\,,$$

  where $\Delta_{1,2}$ can be bounded as follows for $\|\boldsymbol{x}_1-\boldsymbol{x}_2\|_2\le c_0$ with c_0 a small absolute constant (recall that $\mathsf{P}_{\mathsf{T},\boldsymbol{x}_i}$ is the projector orthogonal to $\boldsymbol{x}_i$)

$$\Delta_{1,2} = -2\xi^{(4)}(q)\,\langle \boldsymbol{x}_1,\boldsymbol{v}_2\rangle^2\langle \boldsymbol{v}_1,\boldsymbol{x}_2\rangle^2 - 8\,\xi^{(3)}(q)\langle \boldsymbol{x}_1,\boldsymbol{v}_2\rangle\langle \boldsymbol{v}_1,\boldsymbol{x}_2\rangle\langle \boldsymbol{v}_1,\boldsymbol{v}_2\rangle$$

$$
\begin{aligned}
&+ 4\,\xi^{(2)}(q)q^2(1-\langle \boldsymbol{v}_1,\boldsymbol{v}_2\rangle^2)\\
&\le 8\xi^{(3)}(q)\big|\langle \mathsf{P}_{\mathsf{T},\boldsymbol{x}_2}\boldsymbol{x}_1,\boldsymbol{v}_1-\boldsymbol{v}_2\rangle\langle \mathsf{P}_{\mathsf{T},\boldsymbol{x}_1}\boldsymbol{x}_2,\boldsymbol{v}_1-\boldsymbol{v}_2\rangle\big|\\
&\quad + 8\xi^{(2)}(q)\|\boldsymbol{v}_1-\boldsymbol{v}_2\|_2^2\\
&\le 8\big(\xi^{(3)}(q)\|\boldsymbol{x}_1-\boldsymbol{x}_2\|_2^2+\xi^{(2)}(q)\big)\|\boldsymbol{v}_1-\boldsymbol{v}_2\|_2^2\\
&\le 8\big(\xi^{(3)}(q)(1-q)+\xi^{(2)}(q)\big)\|\boldsymbol{v}_1-\boldsymbol{v}_2\|_2^2\\
&\le 8\xi^{(2)}(1)\|\boldsymbol{v}_1-\boldsymbol{v}_2\|_2^2\,.
\end{aligned}
$$

Hence the proof of Eq. (92) follows by applying Propopsition A.1 and Lemma A.2.

- For Eq. (93), we define $H_2(\boldsymbol{v},\boldsymbol{x})$ indexed by $\boldsymbol{x}\in\mathsf{B}^d(1)$ and $\boldsymbol{v}\in\mathsf{T}_{\boldsymbol{x}}$ via

$$
H_2(\boldsymbol{v},\boldsymbol{x}) := \langle \hat{\boldsymbol{x}}, \nabla^2 F_1(\boldsymbol{x})\boldsymbol{v}\rangle\,. \tag{103}
$$

Its covariance is (here $q:=\langle \boldsymbol{x}_1,\boldsymbol{x}_2\rangle$)

$$
\begin{aligned}
\mathbb{E}\big\{H_2(\boldsymbol{v}_1,\boldsymbol{x}_1)H_2(\boldsymbol{v}_2,\boldsymbol{x}_2)\big\} =& \frac{\langle \boldsymbol{v}_1,\boldsymbol{v}_2\rangle}{\|\boldsymbol{x}_1\|_2\|\boldsymbol{x}_2\|_2}\big[\xi^{(3)}(q)q^2+\xi^{(2)}(q)q\big]\\
&+\frac{\langle \boldsymbol{v}_1,\boldsymbol{x}_2\rangle\langle \boldsymbol{x}_1,\boldsymbol{v}_2\rangle}{\|\boldsymbol{x}_1\|_2\|\boldsymbol{x}_2\|_2}\big[\xi^{(4)}(q)q^2+3\xi^{(3)}(q)q+\xi^{(2)}(q)\big]\,.
\end{aligned}
$$

- Finally, for (94) we define

$$
H_3(\boldsymbol{x}) := \langle \hat{\boldsymbol{x}}, \nabla^2 F_1(\boldsymbol{x})\hat{\boldsymbol{x}}\rangle\,,
$$

and note that (again $q:=\langle \boldsymbol{x}_1,\boldsymbol{x}_2\rangle$)

$$
\mathbb{E}\big\{H_3(\boldsymbol{x}_1)H_3(\boldsymbol{x}_2)\big\} = \frac{1}{\|\boldsymbol{x}_1\|_2^2\|\boldsymbol{x}_2\|_2^2}\big[\xi^{(4)}(q)q^4+4\xi^{(3)}(q)q^3+2\xi^{(2)}(q)q^2\big]\,.
$$

□

Proposition A.4. *Let $\boldsymbol{DF}(\boldsymbol{x})\in\mathbb{R}^{n\times d}$ be the Jacobian of $\boldsymbol{F}$ at $\boldsymbol{x}$ and $\boldsymbol{D}^2\boldsymbol{F}(\boldsymbol{x})\in\mathbb{R}^{n\times d\times d}$ the tensor of second derivatives. We equivalently view the latter as a linear operator $\boldsymbol{D}^2\boldsymbol{F}(\boldsymbol{x}):\mathbb{R}^{d\times d}\to\mathbb{R}^n$, and write $\boldsymbol{D}^2\boldsymbol{F}(\boldsymbol{x})|_{\mathsf{T}\otimes\mathsf{T}}$ for its restriction to $\mathsf{T}_{\boldsymbol{x}}\otimes\mathsf{T}_{\boldsymbol{x}}$ and similarly for $\boldsymbol{D}^2\boldsymbol{F}(\boldsymbol{x})|_{\mathsf{T}\otimes\hat{\boldsymbol{x}}}$ and $\boldsymbol{D}^2\boldsymbol{F}(\boldsymbol{x})|_{\hat{\boldsymbol{x}}\otimes\hat{\boldsymbol{x}}}$.*

If $n\le d$, there exists an absolute constant C such that (using the notation $\log_+(t):=max(\log(t);1)$ introduced above)

$$
\mathbb{E}\big[\max_{\boldsymbol{x}\in\mathsf{B}^d(1)}\|\boldsymbol{DF}(\boldsymbol{x})|_{\mathsf{T}}\|_{\rm op}\big]\le C\sqrt{d\,\xi'(1)\log_+\frac{\xi''(1)}{\xi'(1)}}\,, \tag{104}
$$

$$
\mathbb{E}\big[\max_{\boldsymbol{x}\in\mathsf{B}^d(1)}\|\boldsymbol{DF}(\boldsymbol{x})|_{\hat{\boldsymbol{x}}}\|_{\rm op}\big]\le C\sqrt{d\,\xi''(1)\log_+\frac{\xi^{(3)}(1)}{\xi''(1)}}\,, \tag{105}
$$

$$
\mathbb{E}\big[\max_{\boldsymbol{x}\in\mathsf{B}^d(1)}\|\boldsymbol{D}^2\boldsymbol{F}(\boldsymbol{x})|_{\mathsf{T}\otimes\mathsf{T}}\|_{\rm op}\big]\le C\sqrt{d\,\xi''(1)\log_+\frac{\xi^{(3)}(1)}{\xi''(1)}}\,, \tag{106}
$$

$$
\mathbb{E}\big[\max_{\boldsymbol{x}\in\mathsf{B}^d(1)}\|\boldsymbol{D}^2\boldsymbol{F}(\boldsymbol{x})|_{\mathsf{T}\otimes\hat{\boldsymbol{x}}}\|_{\rm op}\big]\le C\sqrt{d\,\xi^{(3)}(1)\log_+\frac{\xi^{(4)}(1)}{\xi^{(3)}(1)}}\,, \tag{107}
$$

$$
\mathbb{E}\big[\max_{\boldsymbol{x}\in\mathsf{B}^d(1)}\|\boldsymbol{D}^2\boldsymbol{F}(\boldsymbol{x})|_{\hat{\boldsymbol{x}}\otimes\hat{\boldsymbol{x}}}\|_{\rm op}\big]\le C\sqrt{d\,\xi^{(4)}(1)\log_+\frac{\xi^{(5)}(1)}{\xi^{(4)}(1)}}\,. \tag{108}
$$

Proof. All of these bounds follow from Lemma A.3 via the same argument. We will focus to be definite on the bound (104). We define the following Gaussian process indexed by $\boldsymbol{u} \in \mathbb{R}^n$, $\|\boldsymbol{u}\|_2 = 1$, $\boldsymbol{x} \in \mathsf{B}(1)$, $\boldsymbol{v} \in \mathsf{T}_{\boldsymbol{x}}$, $\|\boldsymbol{v}\|_2 = 1$:

$$Z(\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{x}) := \langle \boldsymbol{u}, \boldsymbol{DF}(\boldsymbol{x})\boldsymbol{v} \rangle \,, \tag{109}$$

and notice that of course

$$\max_{\boldsymbol{x} \in \mathsf{B}^d(1)} \|\boldsymbol{DF}(\boldsymbol{x})|_{\mathsf{T}}\|_{\rm op} = \max_{\boldsymbol{u} \in \mathbb{R}^n, \|\boldsymbol{u}\|_2 = 1} \max_{\boldsymbol{v} \in \mathsf{T}_{\boldsymbol{x}}, \|\boldsymbol{v}\|_2 = 1} \max_{\boldsymbol{x} \in \mathsf{B}^d(1)} Z(\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{x}) \,. \tag{110}$$

We next compute the canonical distance of Z, to get

$$d_Z(\boldsymbol{u}_1, \boldsymbol{v}_1, \boldsymbol{x}_1; \boldsymbol{u}_2, \boldsymbol{v}_2, \boldsymbol{x}_2)^2 = \frac{1}{2}\|\boldsymbol{u}_1 - \boldsymbol{u}_2\|^2 (\xi'(\|\boldsymbol{x}_1\|^2) + \xi'(\|\boldsymbol{x}_2\|^2)) + \langle \boldsymbol{u}_1, \boldsymbol{u}_2 \rangle d_{G,\perp}(\boldsymbol{v}_1, \boldsymbol{x}_1; \boldsymbol{v}_2, \boldsymbol{x}_2)^2 \,, \tag{111}$$

where $d_{G,\perp}(\boldsymbol{v}_1, \boldsymbol{x}_1; \boldsymbol{v}_2, \boldsymbol{x}_2)^2$ is the canonical distance of the process $G_\perp(\boldsymbol{v}, \boldsymbol{x}) := \langle \boldsymbol{v}, \nabla F_1(\boldsymbol{x}) \rangle$ (this was derived in the proof of Lemma A.3 but will not be needed here). We bound this distance as

$$d_Z(\boldsymbol{u}_1, \boldsymbol{v}_1, \boldsymbol{x}_1; \boldsymbol{u}_2, \boldsymbol{v}_2, \boldsymbol{x}_2)^2 \le \xi'(1)\|\boldsymbol{u}_1 - \boldsymbol{u}_2\|^2 + d_{G,\perp}(\boldsymbol{v}_1, \boldsymbol{x}_1; \boldsymbol{v}_2, \boldsymbol{x}_2)^2 \,, \tag{112}$$

Therefore by Sudakov-Fernique, letting $\boldsymbol{g} \sim \mathsf{N}(0, \boldsymbol{I}_n)$,

$$\mathbb{E}\big[\max_{\boldsymbol{x} \in \mathsf{B}^d(1)} \|\boldsymbol{DF}(\boldsymbol{x})|_{\mathsf{T}}\|_{\rm op}\big] \le \mathbb{E}\big[\max_{\boldsymbol{u} \in \mathbb{S}^{n-1}} \sqrt{\xi'(1)}\langle \boldsymbol{g}, \boldsymbol{u} \rangle\big] + \mathbb{E} \max_{\boldsymbol{x} \in \mathsf{B}^d(1)} \|\mathsf{P}_{\mathsf{T},\boldsymbol{x}} \nabla F_1(\boldsymbol{x})\|_2 \,. \tag{113}$$

The bound (104) then follows from Lemma A.3. The other bounds are proved analogously. □

We next use the above estimates to control the Lipschitz constant of F and its derivatives. Recall that, for $\boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathbb{R}^d$, $\boldsymbol{x}_i \neq \boldsymbol{0}$, we let $\boldsymbol{R}_{\boldsymbol{x}_1,\boldsymbol{x}_2} \in \mathbb{R}^{d \times d}$ be the rotation in the plane $\boldsymbol{x}_1, \boldsymbol{x}_2$ such that $\boldsymbol{R}_{\boldsymbol{x}_1,\boldsymbol{x}_2}\boldsymbol{x}_1/\|\boldsymbol{x}_1\| = \boldsymbol{x}_2/\|\boldsymbol{x}_2\|$. Among the two rotations that leave the orthogonal complement of $\boldsymbol{x}_1, \boldsymbol{x}_2$ unchanged, choose the one with smallest angle, and break ties arbitrarily. (In other words, $\boldsymbol{R}_{\boldsymbol{x}_1,\boldsymbol{x}_2}$ is the parallel transport on $\mathbb{S}^{d-1}$.)

Remark A.2. By Borell-TIS inequality, the bounds in Proposition A.1, Lemma A.3 and Proposition A.4 with modified constants also hold with probability at least $1 - \exp(-C(\xi)d)$ for some $C(\xi) > 0$.

Definition A.5. *For $\boldsymbol{x} \in \mathbb{R}^d$, we choose $\boldsymbol{U}_{\boldsymbol{x}} \in \mathbb{R}^{d \times (d-1)}$ such that $\boldsymbol{U}_{\boldsymbol{x}}^{\mathsf{T}}\boldsymbol{U}_{\boldsymbol{x}} = \boldsymbol{I}_{d-1}$, $\boldsymbol{U}_{\boldsymbol{x}}^{\mathsf{T}}\boldsymbol{x} = \boldsymbol{0}$ (i.e. $\boldsymbol{U}_{\boldsymbol{x}}$ is an orthonormal basis for the tangent space $\mathsf{T}_{\boldsymbol{x}}$.) Further, for $\boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathbb{R}^d$, define $\boldsymbol{U}_{\boldsymbol{x}_1,\boldsymbol{x}_2} := \boldsymbol{R}_{\boldsymbol{x}_1,\boldsymbol{x}_2}\boldsymbol{U}_{\boldsymbol{x}_1}$.*

For $\Omega \subseteq \mathbb{R}^d$, we define the following Lipschitz constants

$$\mathrm{Lip}(\boldsymbol{F}; S) := \sup_{\boldsymbol{x}_1 \neq \boldsymbol{x}_2 \in \Omega} \frac{\|\boldsymbol{F}(\boldsymbol{x}_1) - \boldsymbol{F}(\boldsymbol{x}_2)\|}{\|\boldsymbol{x}_1 - \boldsymbol{x}_2\|_2} \,, \tag{114}$$

$$\mathrm{Lip}_\perp(\boldsymbol{DF}; \Omega) := \sup_{\boldsymbol{x}_1 \neq \boldsymbol{x}_2 \in \Omega} \frac{\|\boldsymbol{DF}(\boldsymbol{x}_1)\boldsymbol{U}_{x_1} - \boldsymbol{DF}(\boldsymbol{x}_2)\boldsymbol{U}_{\boldsymbol{x}_1,\boldsymbol{x}_2}\|_{\rm op}}{\|\boldsymbol{x}_1 - \boldsymbol{x}_2\|_2} \,, \tag{115}$$

$$\mathrm{Lip}_\perp(\nabla^2 \boldsymbol{F}; \Omega) := \sup_{\boldsymbol{x}_1 \neq \boldsymbol{x}_2 \in \Omega} \max_{\ell \le n} \frac{\|\boldsymbol{U}_{\boldsymbol{x}_1}^{\mathsf{T}} \nabla^2 F_\ell(\boldsymbol{x}_1)\boldsymbol{U}_{\boldsymbol{x}_1} - \boldsymbol{U}_{\boldsymbol{x}_1,\boldsymbol{x}_2}^{\mathsf{T}} \nabla^2 F_\ell(\boldsymbol{x}_2)\boldsymbol{U}_{\boldsymbol{x}_1,\boldsymbol{x}_2}\|_{\rm op}}{\|\boldsymbol{x}_1 - \boldsymbol{x}_2\|_2} \,. \tag{116}$$

Lemma A.6. *Assume $n \le d$, and recall the notation $\log_+(t) := \max(\log(t); 1)$. Then there exists a universal constant C_1, and a ξ-dependent constant $C_*(\xi)$ such that the following hold with probability at least $1 - \exp(-d/C_*(\xi))$:*

$$\mathrm{Lip}(\boldsymbol{F}; \mathsf{B}^d(1)) \le C_1 \sqrt{d\, \xi''(1) \log_+ \frac{\xi^{(3)}(1)}{\xi''(1)}}\,, \tag{117}$$

$$\mathrm{Lip}_\perp(\boldsymbol{DF}; \mathsf{B}^d(1)) \le C_1 \sqrt{d\, \xi^{(4)}(1) \log_+ \frac{\xi^{(5)}(1)}{\xi^{(4)}(1)}}\,, \tag{118}$$

$$\mathrm{Lip}_\perp(\nabla^2 \boldsymbol{F}; \mathsf{B}^d(1)) \le C_*(\xi)\sqrt{d}\,. \tag{119}$$

Further, the following tighter Lipshitz constant holds on $\mathbb{S}^{d-1}$:

$$\mathrm{Lip}(\boldsymbol{F}; \mathbb{S}^{d-1}) \le C_1 \sqrt{d\, \xi'(1) \log_+ \frac{\xi''(1)}{\xi'(1)}}\,, \tag{120}$$

$$\mathrm{Lip}_\perp(\boldsymbol{DF}; \mathbb{S}^{d-1}) \le C_1 \sqrt{d\, \xi''(1) \log_+ \frac{\xi^{(3)}(1)}{\xi''(1)}}\,. \tag{121}$$

Proof. Throughout, we make use of Remark A.2.

- The bound (117) follows from Eqs. (104), (105).
- The bound (118) follows from Eqs. (106), (107), (108).
- The proof of Eq. (119) follows by bounding $\max_{\boldsymbol{x}} \|\nabla^3 F_\ell(\boldsymbol{x})\|$, which in turn can be done by the same technique as in the proof of Lemma A.3. Since we are not seeking a precise characterization of the constant $C_*(\xi)$ it suffices to notice that the canoninical distance of the process $\langle \nabla^3 F_\ell(\boldsymbol{x}), \boldsymbol{v}^{\otimes 3}\rangle$ is bounded by a smooth function of the distances $\|\boldsymbol{v}_1 - \boldsymbol{v}_2\|$, $\|\boldsymbol{x}_1 - \boldsymbol{x}_2\|$.
- The bound (120) follows from Eqs. (104) by noting that, for $\boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathbb{S}^{d-1}$, letting $\boldsymbol{x}(t)$ denote a geodesic on the sphere parametrized by arclength, we have

  $$\boldsymbol{F}(\boldsymbol{x}_2) - \boldsymbol{F}(\boldsymbol{x}_1) = \int_0^{t_*} \boldsymbol{DF}(\boldsymbol{x}(t))\dot{\boldsymbol{x}}(t)\, \mathrm{d}t\,. \tag{122}$$

  Here $t_* := \arccos(\langle \boldsymbol{x}_1, \boldsymbol{x}_2\rangle)$ and we notice that $\dot{\boldsymbol{x}}(t) \in \mathsf{T}_{\boldsymbol{x}(t)}$.
- Finally, for the bound (121) we consider a geodesic path between $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ in the sphere. For any two vectors $\boldsymbol{v}_1 \in \mathsf{T}_{\boldsymbol{x}_1}$, $\boldsymbol{v}_2 = \boldsymbol{R}_{\boldsymbol{x}_1,\boldsymbol{x}_2}\boldsymbol{v}_1 \in \mathsf{T}_{\boldsymbol{x}_2}$, let $\boldsymbol{v}(t)$ be the parallel transport of $\boldsymbol{v}_1$ along the geodesic connecting $\boldsymbol{x}_1$ to $\boldsymbol{x}_2$. We then have

  $$\boldsymbol{DF}(\boldsymbol{x}_2)\boldsymbol{v}_2 - \boldsymbol{DF}(\boldsymbol{x}_1)\boldsymbol{v}_1 = \int_0^{t_*} \left\{\boldsymbol{D}^2\boldsymbol{F}(\boldsymbol{x}(t))\{\dot{\boldsymbol{x}}(t), \boldsymbol{v}(t)\} + \boldsymbol{DF}(\boldsymbol{x}(t))\dot{\boldsymbol{v}}(t)\right\} \mathrm{d}t\,. \tag{123}$$

  Hence the claim follows from Eqs. (105) and (106)

□

B Second moment method: Proof of Theorem 1

In this section we study the volume $\mathrm{V}(\boldsymbol{u})$ of the set of solutions for $\boldsymbol{F}(\boldsymbol{x}) = \boldsymbol{u}$ for a given vector $\boldsymbol{u} \in \mathbb{R}^n$, as defined in Eqs. (9) and (11). We will use the Kac-Rice formula [Kac43, Ric45, Sub23] to prove that, for any deterministic sequence $\delta_d > 0$ with $\lim_{d\to\infty} \delta_d = 0$,

$$\lim_{n,d\to\infty} \sup_{\|\boldsymbol{u}\|_2/\sqrt{n} \in [\xi_0^{1/2}-\delta_d, \xi_0^{1/2}+\delta_d]} \frac{\mathbb{E}\{\mathrm{V}(\boldsymbol{u})^2\}}{\mathbb{E}\{\mathrm{V}(\boldsymbol{u})\}^2} = 1\,. \tag{124}$$

This will complete the proof sketch given in Section 2.1.

We begin by computing the first moment. Recall that Vol_i is the Hausdorff measure of dimension i, or the counting measure when $i = 0$.

Proposition B.1. *For any $\boldsymbol{u} \in \mathbb{R}^n$,*

$$\mathbb{E}\,\mathrm{V}(\boldsymbol{u}) = \mathcal{V}_{d-n-1}\Big(\frac{\xi'(1)}{\xi(1)}\Big)^{\frac{n}{2}} e^{-\frac{\|\boldsymbol{u}\|^2}{2\xi(1)}}, \tag{125}$$

where $\mathcal{V}_i := \mathrm{Vol}_i(\{\boldsymbol{x} \in \mathbb{R}^{i+1} : \|\boldsymbol{x}\| = 1\}) = 2\pi^{\frac{i+1}{2}}/\Gamma(\frac{i+1}{2})$ is the volume of the sphere.

Proof. The proof here is identical to the case $\boldsymbol{u} = 0$ treated in [Sub23]. We repeat it for completeness, using the notation of the current paper.

Recall that, for $\boldsymbol{x} \in \mathbb{R}^d$ on the sphere $\boldsymbol{U}_{\boldsymbol{x}} \in \mathbb{R}^{d\times(d-1)}$ denotes a matrix whose columns form a basis of the tangent space $\mathsf{T}_{\boldsymbol{x}}$ to the sphere of radius $\|\boldsymbol{x}\|_2$ at $\boldsymbol{x}$ (equivalently, to the orthogonal complement of $\boldsymbol{x}$ in $\mathbb{R}^d$). We write $\boldsymbol{D}_\perp \boldsymbol{F}(\boldsymbol{x}) := \boldsymbol{D}\boldsymbol{F}(\boldsymbol{x})\boldsymbol{U}_{\boldsymbol{x}} \in \mathbb{R}^{n\times d-1}$.

By a variant of the Kac-Rice formula in [AW09, Theorem 6.8],

$$\mathbb{E}\mathrm{V}(u) = \int_{S^{d-1}} \mathbb{E}\Big[J(\boldsymbol{D}_\perp \boldsymbol{F}(\boldsymbol{x}))\,\Big|\,\boldsymbol{F}(\boldsymbol{x}) = \boldsymbol{u}\Big]\, p_{\boldsymbol{F}(\boldsymbol{x})}(\boldsymbol{u})\,\mathrm{dVol}_{d-1}(\boldsymbol{x}), \tag{126}$$

where $J(\boldsymbol{A}) = \sqrt{\det \boldsymbol{A}\boldsymbol{A}^{\mathsf{T}}}$, $\mathrm{dVol}_{d-1}(\boldsymbol{x})$ is the $(d-1)$-dimensional volume element on the sphere and

$$p_{\boldsymbol{F}(\boldsymbol{x})}(\boldsymbol{u}) = \big(2\pi\xi(1)\big)^{-\frac{n}{2}} e^{-\frac{\|\boldsymbol{u}\|^2}{2\xi(1)}}$$

is the density of $\boldsymbol{F}(\boldsymbol{x})$ at $\boldsymbol{u}$. Using the Kac-Rice formula requires checking that certain conditions are satisfied, this was done in Remark 3 of [Sub23].

Recall that by Lemma 4.1, $\boldsymbol{D}_\perp \boldsymbol{F}(\boldsymbol{x}) \overset{\mathrm{d}}{=} \sqrt{\xi'(q)}\,\boldsymbol{Z}$ where $q = \|\boldsymbol{x}\|^2$, $\boldsymbol{Z} \sim \mathsf{GOE}(n, d-1)$ and $\boldsymbol{D}_\perp \boldsymbol{F}(\boldsymbol{x})$ is independent of $\boldsymbol{F}(\boldsymbol{x})$. Thus,

$$\mathbb{E}\mathrm{V}(\boldsymbol{u}) = \mathcal{V}_{d-1}\Big(\frac{\xi'(1)}{2\pi\xi(1)}\Big)^{n/2} e^{-\frac{\|\boldsymbol{u}\|^2}{2\xi(1)}}\mathbb{E}J(\boldsymbol{Z}). \tag{127}$$

In the case $\xi(t) = t$ and $\boldsymbol{u} = 0$, the above is the volume of the set of points on the sphere orthogonal to n independent Gaussian vectors, hence equal to $\mathcal{V}_{d-n-1}$. Hence,

$$\mathcal{V}_{d-n-1} = \mathcal{V}_{d-1}\Big(\frac{1}{2\pi}\Big)^{n/2}\mathbb{E}\,J(\boldsymbol{Z}) \tag{128}$$

and (125) follows. □

The next theorem establishes the desired second moment bound. To verify this note that the function $\Phi(r)$ defined here matches $\Psi(r;\alpha,\xi)$ of Eq. (7), if we replace γ^2 by $\alpha\xi(0)$ and $\xi(t)$ by $\xi_{>0}(t) = \xi(t) - \xi(0)$, and that condition (129) is implied by $\alpha < \alpha_{\mathrm{LB}}(\xi)$.

Theorem 7. *Assume that* $\xi(0)=\xi'(0)=\xi''(0)=0$. *Let* $\gamma\geq 0$ *and* $\alpha\in(0,1]$ *and define*

$$\Phi(r):=\frac{1}{2}\log(1-r^2)-\frac{\alpha}{2}\log\Big(1-\frac{\xi(r)^2}{\xi(1)^2}\Big)+\frac{\gamma^2}{\xi(1)}-\frac{\gamma^2}{\xi(1)+\xi(r)}\,.$$

If for any $\varepsilon>0$,

$$\Phi(0)>\sup_{r\in(\varepsilon,1)}\Phi(r)\,, \tag{129}$$

then for any deterministic sequence $\delta_d>0$ *such that* $\lim_{d\to\infty}\delta_d=0$,

$$\lim_{d,n\to\infty}\frac{\mathbb{E}\big(\mathrm{V}(\boldsymbol{u})^2\big)}{\big(\mathbb{E}\mathrm{V}(\boldsymbol{u})\big)^2}=1 \tag{130}$$

uniformly in $n=n(d)<d-1$ *and* $\boldsymbol{u}=\boldsymbol{u}^{(n,d)}\in\mathbb{R}^n$ *such that* $\lim_{d\to\infty}n(d)\to\alpha$ *and* $\sqrt{d}(\gamma-\delta_d)\leq\|\boldsymbol{u}\|\leq\sqrt{d}(\gamma+\delta_d)$.

Proof. To simplify the proof, throughout we will assume that $n\leq d-2$. The case in which $n=d-1$ infinitely often (in particular $\alpha=1$) was treated for $\boldsymbol{u}=\boldsymbol{0}$ in [Sub23] and the argument can be adapted to our situation with $\boldsymbol{u}\neq 0$, but to save space we will omit the proof. We refer the interested reader to Lemmas 8 and 9 in [Sub23].

For any union of intervals $I\subset[-1,1]$, define

$$\mathrm{V}^{(2)}(\boldsymbol{u},I):=\mathrm{Vol}_{2(d-n-1)}\Big(\Big\{(\boldsymbol{x},\boldsymbol{y}):\,\langle\boldsymbol{x},\boldsymbol{y}\rangle\in I,\,\boldsymbol{F}(\boldsymbol{x})=\boldsymbol{F}(\boldsymbol{y})=\boldsymbol{u},\boldsymbol{x},\boldsymbol{y}\in\mathbb{S}^{d-1}\Big\}\Big).$$

As in the proof of Proposition B.1, we may apply the Kac-Rice formula to the random field $(\boldsymbol{x},\boldsymbol{y})\to(\boldsymbol{F}(\boldsymbol{x}),\boldsymbol{F}(\boldsymbol{y}))$ defined on the domain

$$\mathcal{D}(I):=\{(\boldsymbol{x},\boldsymbol{y}):\,\langle\boldsymbol{x},\boldsymbol{y}\rangle\in I,\|\boldsymbol{x}\|=\|\boldsymbol{y}\|=1\} \tag{131}$$

to compute the expectation of $\mathrm{V}^{(2)}(\boldsymbol{u},I)$, see Section 3 of [Sub23] for the details. This yields that

$$\begin{aligned}\mathbb{E}\,\mathrm{V}^{(2)}(\boldsymbol{u},I)&=\big(\xi'(1)\big)^n\int_{(\boldsymbol{x},\boldsymbol{y})\in\mathcal{D}(I)}p_{\boldsymbol{F}(\boldsymbol{x}),\boldsymbol{F}(\boldsymbol{y})}(\boldsymbol{u},\boldsymbol{u})\,T(\boldsymbol{u},\langle\boldsymbol{x},\boldsymbol{y}\rangle)\,\mathrm{dVol}_{d-1}(\boldsymbol{x})\mathrm{dVol}_{d-1}(\boldsymbol{y})\\&=\mathcal{V}_{d-1}\cdot\big(\xi'(1)\big)^n\int_{\{\boldsymbol{y}\in\mathbb{S}^{d-1}:\,\langle\boldsymbol{x},\boldsymbol{y}\rangle\in I\}}p_{\boldsymbol{F}(\boldsymbol{x}),\boldsymbol{F}(\boldsymbol{y})}(\boldsymbol{u},\boldsymbol{u})T(\boldsymbol{u},\langle\boldsymbol{x},\boldsymbol{y}\rangle)d\mathrm{Vol}_{d-1}(\boldsymbol{y}),\end{aligned} \tag{132}$$

where in the second line $\boldsymbol{x}$ is an arbitrary point on the sphere, and writing $\langle\boldsymbol{x},\boldsymbol{y}\rangle=r$, we have

$$p_{\boldsymbol{F}(\boldsymbol{x}),\boldsymbol{F}(\boldsymbol{y})}(\boldsymbol{u},\boldsymbol{u})=(2\pi)^{-n}(\xi(1)^2-\xi(r)^2)^{-\frac{n}{2}}e^{-\frac{\|\boldsymbol{u}\|^2}{\xi(1)+\xi(r)}}$$

is the density of $(\boldsymbol{F}(\boldsymbol{x}),\boldsymbol{F}(\boldsymbol{y}))$ at $(\boldsymbol{u},\boldsymbol{u})$ since the covariance matrix of $(F_i(\boldsymbol{x}),F_i(\boldsymbol{y}))$ is

$$\begin{pmatrix}\xi(1)&\xi(r)\\\xi(r)&\xi(1)\end{pmatrix},$$

and we define (always for $\langle\boldsymbol{x},\boldsymbol{y}\rangle=r$)

$$T(\boldsymbol{u},r):=\mathbb{E}\left[J\Big(\frac{\boldsymbol{D}_\perp\boldsymbol{F}(\boldsymbol{x})}{\sqrt{\xi'(1)}}\Big)J\Big(\frac{\boldsymbol{D}_\perp\boldsymbol{F}(\boldsymbol{y})}{\sqrt{\xi'(1)}}\Big)\,\middle|\,\boldsymbol{F}(\boldsymbol{x})=\boldsymbol{F}(\boldsymbol{y})=\boldsymbol{u}\right]. \tag{133}$$

By rotational invariance $T(\boldsymbol{u}, r)$ does not depend on the choice of $\boldsymbol{U}_{\boldsymbol{x}}$ and $\boldsymbol{U}_{\boldsymbol{y}}$ and of $\boldsymbol{x}$ and $\boldsymbol{y}$ as long as $\langle \boldsymbol{x}, \boldsymbol{y}\rangle = r$.

Using the co-area formula with the function $\rho(\boldsymbol{y}) = \langle \boldsymbol{x}, \boldsymbol{y}\rangle$, we may express the integral w.r.t. $\boldsymbol{y}$ in Eq. (132) as a one-dimensional integral over a parameter r (the volume of the inverse-image $\rho^{-1}(r)$ and the inverse of the Jacobian are given by $\mathcal{V}_{d-2}(1-r^2)^{(d-2)/2}$ and $(1-r^2)^{-1/2}$, respectively). This yields

$$\mathbb{E}\,\mathrm{V}^{(2)}(\boldsymbol{u}, I) = \mathcal{V}_{d-1}\mathcal{V}_{d-2}(2\pi)^{-n}\int_I \left(\frac{\xi'(1)^2}{\xi(1)^2 - \xi(r)^2}\right)^{n/2} (1-r^2)^{\frac{d-3}{2}} T(\boldsymbol{u}, r) e^{-\frac{\|\boldsymbol{u}\|^2}{\xi(1)+\xi(r)}}\,\mathrm{d}r\,. \tag{134}$$

From Eqs. (127) and (134),we get

$$\frac{\mathbb{E}\,\mathrm{V}^{(2)}(\boldsymbol{u}, I)}{\big(\mathbb{E}\,\mathrm{V}(\boldsymbol{u})\big)^2} = \frac{\mathcal{V}_{d-2}}{\mathcal{V}_{d-1}}\int_I \frac{T(\boldsymbol{u}, r)}{\big(\mathbb{E}J(\boldsymbol{Z})\big)^2}\left(1 - \frac{\xi(r)^2}{\xi(1)^2}\right)^{-n/2} (1-r^2)^{(d-3)/2} e^{\frac{\|\boldsymbol{u}\|^2}{\xi(1)} - \frac{\|\boldsymbol{u}\|^2}{\xi(1)+\xi(r)}}\,\mathrm{d}r. \tag{135}$$

The next lemma provide a useful bound on the integrand.

Lemma B.2. *Assume that $\xi(0) = \xi'(0) = 0$. Define*

$$\kappa = \kappa(r) = \frac{\xi'(r)}{\xi(1) + \xi(r)}\sqrt{\frac{1-r^2}{\xi'(1)}}\,. \tag{136}$$

Then for some universal constant c and sequence $\tau_d \to 0$, for any $r \in (-1, 1)$,

$$\frac{T(\boldsymbol{u}, r)}{\big(\mathbb{E}J(\boldsymbol{Z})\big)^2} \le \big(1+\tau_d\big)\big(1 + 2r^2\sqrt{d}\big)\big(1 + c\kappa(r)\|\boldsymbol{u}\| + c\kappa(r)^2\|\boldsymbol{u}\|^2\big)\,.$$

Before proving this lemma, let us complete the proof of the theorem. Note that $\frac{\mathcal{V}_{d-2}}{\mathcal{V}_{d-1}} = \sqrt{\frac{d}{2\pi}}(1 + o_d(1))$. Hence, we have that

$$\frac{1}{d}\log\frac{\mathbb{E}\,\mathrm{V}^{(2)}(\boldsymbol{u}, I)}{\big(\mathbb{E}\,\mathrm{V}(\boldsymbol{u})\big)^2} \le \frac{1}{d}\log\int_I \nu(\boldsymbol{u}, r)\,\mathrm{d}r + o_d(1)\,,$$

where the $o_d(1)$ term is independent of $\boldsymbol{u}$ and n, and

$$\nu(\boldsymbol{u}, r) := \left(1 - \frac{\xi(r)^2}{\xi(1)^2}\right)^{-n/2} (1-r^2)^{(d-3)/2}\big(1 + c\kappa(r)\|\boldsymbol{u}\| + c\kappa(r)^2\|\boldsymbol{u}\|^2\big) e^{\frac{\|\boldsymbol{u}\|^2}{\xi(1)} - \frac{\|\boldsymbol{u}\|^2}{\xi(1)+\xi(r)}}\,.$$

Note that $\nu(\boldsymbol{u}, r) \le (1-r^2)^{-1/2} e^{d\Phi(r) + o(d)}$, uniformly in $n \le d-2$, $\sqrt{d}(\gamma - \delta_d) \le \|\boldsymbol{u}\| \le \sqrt{d}(\gamma + \delta_d)$ and $r \in (-1, 1)$. Assuming that $n = n(d) \to \alpha$, since $(1-r^2)^{-1/2}$ is integrable on $(-1, 1)$, we have that for any $\varepsilon > 0$, uniformly in $\sqrt{d}(\gamma - \delta_d) \le \|\boldsymbol{u}\| \le \sqrt{d}(\gamma + \delta_d)$,

$$\limsup_{d\to\infty}\frac{1}{d}\log\frac{\mathbb{E}\mathrm{V}^{(2)}(\boldsymbol{u}, [-1,1]\setminus(-\varepsilon, \varepsilon))}{\big(\mathbb{E}\mathrm{V}(\boldsymbol{u})\big)^2} \le \sup_{r\in[\varepsilon,1]}\Phi(r) < \Phi(0) = 0\,,$$

where the second inequality follows by assumption.

Of course, the same bound holds with some sequence $\varepsilon_d \to 0$ instead of ε. Since the ratio of moments in (130) is lower bounded by 1, to prove the same equation it remains to show that, uniformly,

$$\limsup_{d\to\infty}\frac{\mathbb{E}\mathrm{V}^{(2)}(\boldsymbol{u}, (-\varepsilon_d, \varepsilon_d))}{\big(\mathbb{E}\mathrm{V}(\boldsymbol{u})\big)^2} \le 1\,. \tag{137}$$

Note that $\frac{\xi(r)}{\xi(1)} = O(r^3)$ and $\kappa(r) = O(r^2)$ for small r, since we assume that $\xi(0) = \xi'(0) = \xi''(0) = 0$, and recall that $\frac{\mathcal{V}_{d-2}}{\mathcal{V}_{d-1}} = \sqrt{\frac{d}{2\pi}}(1 + o_d(1))$. From (135) we obtain that the ratio in (137) is equal to

$$(1+o_d(1))\sqrt{\frac{d}{2\pi}}\int_{-\varepsilon_d}^{\varepsilon_d}\frac{T(\boldsymbol{u},r)}{\big(\mathbb{E}J(\boldsymbol{Z})\big)^2}\big(1-O(r^6)\big)^{-n/2}(1-r^2)^{d/2}e^{\frac{\|\boldsymbol{u}\|^2}{\xi(1)}O(r^3)}dr$$
$$=(1+o_d(1))\sqrt{\frac{d}{2\pi}}\int_{-\varepsilon_d}^{\varepsilon_d}\exp\Big(-\frac{d}{2}r^2+2\sqrt{d}r^2+\frac{\gamma^2 d}{\xi(1)}O(r^3)+dO(r^6)+c\gamma\sqrt{d}O(r^2)+c\gamma^2 dO(r^4)\Big)dr$$

where we used Lemma B.2 and that $\log(1+t) \le t$. This proves (137) and completes the proof. □

Proof of Lemma B.2. Fix some $r \in (-1,1)$. Let $\boldsymbol{M}^1$ and $\boldsymbol{M}^2$ be jointly Gaussian, centered random matrices such that

$$\operatorname{cov}(\boldsymbol{M}^t_{ij},\boldsymbol{M}^s_{kl}) = \delta_{ik}\delta_{jl}\cdot\begin{cases}1 & t=s,\ j<d-1\\ 1-\frac{\xi(1)}{\xi'(1)}\frac{\xi'(r)^2(1-r^2)}{\xi(1)^2-\xi(r)^2} & t=s,\ j=d-1\\ \frac{\xi'(r)}{\xi'(1)} & t\neq s,\ j<d-1\\ r\frac{\xi'(r)}{\xi'(1)}-\frac{\xi''(r)}{\xi'(1)}(1-r^2)-\frac{\xi(r)}{\xi'(1)}\frac{\xi'(r)^2(1-r^2)}{\xi(1)^2-\xi(r)^2} & t\neq s,\ j=d-1\\ 0 & \text{otherwise.}\end{cases} \tag{138}$$

Suppose $\boldsymbol{x}$ and $\boldsymbol{y}$ are such that $\langle \boldsymbol{x},\boldsymbol{y}\rangle = r \in (-1,1)$. In Lemma 6 of [Sub23] it was shown that for an appropriate choice of $\boldsymbol{U}_{\boldsymbol{x}}$ and $\boldsymbol{U}_{\boldsymbol{y}}$ (bases of the tangent spaces $\mathsf{T}_{\boldsymbol{x}}$, $\mathsf{T}_{\boldsymbol{y}}$),

$$\Big(\frac{\boldsymbol{D}_\perp\boldsymbol{F}(\boldsymbol{x})}{\sqrt{\xi'(1)}},\frac{\boldsymbol{D}_\perp\boldsymbol{F}(\boldsymbol{y})}{\sqrt{\xi'(1)}}\Big)\overset{\mathrm{d}}{=}\Big(\boldsymbol{M}^1-\kappa\boldsymbol{u}\cdot\boldsymbol{e}_{d-1}^{\mathsf{T}},\ \boldsymbol{M}^2+\kappa\boldsymbol{u}\cdot\boldsymbol{e}_{d-1}^{\mathsf{T}}\Big),$$

where $\boldsymbol{e}_{d-1}\in\mathbb{R}^{d-1}$ is the last (column) vector in the standard basis and κ is defined in (136). (To be precise, only the case $\boldsymbol{u}=0$ was treated there, but only the conditional expectation depends on $\boldsymbol{u}$ and is easily computed in the same way.)

Let $\boldsymbol{O}$ be some orthogonal matrix such that $\boldsymbol{O}\boldsymbol{u} = (0,\dots,0,\|\boldsymbol{u}\|)^{\mathsf{T}}$. Since $J(\boldsymbol{A}) = J(\boldsymbol{O}\boldsymbol{A})$ and $(\boldsymbol{O}\boldsymbol{M}^1,\boldsymbol{O}\boldsymbol{M}^2)\overset{\mathrm{d}}{=}(\boldsymbol{M}^1,\boldsymbol{M}^2)$,

$$\Big(J\Big(\frac{\boldsymbol{D}_\perp\boldsymbol{F}(\boldsymbol{x})}{\sqrt{\xi'(1)}}\Big),J\Big(\frac{\boldsymbol{D}_\perp\boldsymbol{F}(\boldsymbol{y})}{\sqrt{\xi'(1)}}\Big)\Big)\overset{\mathrm{d}}{=}\Big(J\big(\hat{\boldsymbol{M}}^1\big),J\big(\hat{\boldsymbol{M}}^2\big)\Big),$$

where we denote $\hat{\boldsymbol{M}}^1 := \boldsymbol{M}^1-\kappa\|\boldsymbol{u}\|\boldsymbol{e}_n\boldsymbol{e}_{d-1}^{\mathsf{T}}$ and $\hat{\boldsymbol{M}}^2 := \boldsymbol{M}^2+\kappa\|\boldsymbol{u}\|\boldsymbol{e}_n\boldsymbol{e}_{d-1}^{\mathsf{T}}$. Recall that for any matrix $\boldsymbol{A}$, $J(\boldsymbol{A}) = \prod_{i=1}^n\Theta_i(\boldsymbol{A})$, where we define $\Theta_i(\boldsymbol{A})$ as the norm of the projection of the i-th row of $\boldsymbol{A}$ to the orthogonal space to its first $i-1$ rows. Note that $\Theta_i(\hat{\boldsymbol{M}}^j) = \Theta_i(\boldsymbol{M}^j)$ for $i<n$ and $\Theta_n(\hat{\boldsymbol{M}}^j)\le\Theta_n(\boldsymbol{M}^j)+\kappa\|\boldsymbol{u}\|$. Therefore, using independence of the rows anf the orthogonal invariance of Gaussians,

$$\begin{aligned}T(\boldsymbol{u},r)&\le\mathbb{E}\Big[\prod_{i=1}^{n-1}\Theta_i(\boldsymbol{M}^1)\big(\Theta_n(\boldsymbol{M}^1)+\kappa\|\boldsymbol{u}\|\big)\prod_{i=1}^{n-1}\Theta_i(\boldsymbol{M}^2)\big(\Theta_n(\boldsymbol{M}^2)+\kappa\|\boldsymbol{u}\|\big)\Big]\\&=\mathbb{E}\big[J(\boldsymbol{M}^1)J(\boldsymbol{M}^2)\big]\Big(1+\frac{2\kappa\|\boldsymbol{u}\|\mathbb{E}\Theta_n(\boldsymbol{M}^1)}{\mathbb{E}\big[\Theta_n(\boldsymbol{M}^1)\Theta_n(\boldsymbol{M}^2)\big]}+\frac{\kappa^2\|\boldsymbol{u}\|^2}{\mathbb{E}\big[\Theta_n(\boldsymbol{M}^1)\Theta_n(\boldsymbol{M}^2)\big]}\Big).\end{aligned}$$

Using that $\Theta_n(\boldsymbol{M}^i) \sim \chi_{n-d}$, we have that, for some universal constant c,

$$T(\boldsymbol{u}, r) \le \mathbb{E}\big[J(\boldsymbol{M}^1)J(\boldsymbol{M}^2)\big]\big(1 + c\kappa(r)\|\boldsymbol{u}\| + c\kappa(r)^2\|\boldsymbol{u}\|^2\big)\,.$$

Lemma 10 of [Sub23] bounds $T(\boldsymbol{u}, r)$ for $\boldsymbol{u} = 0$, which is exactly equal to $\mathbb{E}\big[J(\boldsymbol{M}^1)J(\boldsymbol{M}^2)\big]$. Precisely, it states that

$$T(\boldsymbol{0}, r) = \mathbb{E}\big[J(\boldsymbol{M}^1)J(\boldsymbol{M}^2)\big] \le \big(1 + \tau_d\big)\big(1 + 2r^2\sqrt{d}\big)\,,$$

for some universal $\tau_d \to 0$. This completes the proof. □

C Large-p Asymptotics of the lower bound: Eq. (12)

For $\xi(t) = \xi_0 + t^p$, the function $\Psi(r) = \Psi(r; \alpha, \xi_0, p)$ of Eq. (7) reads

$$\Psi(r; \alpha, \xi_0, p) := \frac{1}{2}\log(1 - r^2) - \frac{\alpha}{2}\log\big(1 - r^{2p}\big) + \alpha\xi_0\frac{r^p}{1 + r^p}\,, \tag{139}$$

and since $\Psi''(0; \alpha, \xi_0, p) = -1$ for all α, we have

$$\alpha_{\rm LB}(\xi_0, p) := \inf\Big\{\alpha \ge 0 : \sup_{r\in[0,1]} \Psi(r; \alpha, \xi_0, p) > 0\Big\}\,. \tag{140}$$

To get an upper bound on $\alpha_{\rm LB}(\xi_0, p)$, we choose $r = 1 - M/p$ for some constant $M > 0$. Recalling that $\xi_0 = \gamma_0 \log p$ for some constant $\gamma_0 > 1$, it is easy to compute that, as $p \to \infty$,

$$\Psi\Big(r = 1 - \frac{M}{p}\Big) = -\frac{1}{2}\log p + \alpha\xi_0\frac{e^{-M}}{1 + e^{-M}} + O(1) \tag{141}$$

$$= \Big(-\frac{1}{2} + \frac{\alpha\gamma_0}{1 + e^M}\Big)\log p + O(1)\,. \tag{142}$$

Therefore, for all p large enough

$$\alpha_{\rm LB}(\xi_0 = \gamma_0 \log p, p) \le \frac{e^M + 1}{2\gamma_0}\,. \tag{143}$$

Since $M > 0$ is arbitrary, we get $\limsup_{p\to\infty} \alpha_{\rm LB}(\xi_0 = \gamma_0 \log p, p) \le 1/\gamma_0$.

Next we prove the lower bound. For that purpose, we will fix $\gamma_0 > 1$, assume that $\xi_0 = \gamma_0 \log p$, $\alpha \le (1 - \varepsilon)\gamma_0^{-1}$ and prove that $\sup_{r\in[0,1]} \Psi(r; \alpha, \xi_0, p) \le 0$ for all p large enough. Note that, under these definitions, $\Psi(r; \alpha, \xi_0, p) \le \Psi_*(r)$, where

$$\Psi_*(r) := \frac{1}{2}\log(1 - r^2) - \frac{1}{2}\log\big(1 - r^{2p}\big) + (1 - \varepsilon)\frac{r^p}{1 + r^p}\,\log p\,. \tag{144}$$

We split the maximization over r in two regions (depending on a constant Δ to be fixed below):

1. $0 \le r \le 1 - p^{-1+\Delta}$. For an absolute constant C, we have

$$\Psi_*(r) \le -\frac{1}{2}\,r^2 + C(\log p)r^p \tag{145}$$

$$\le \Big(-\frac{1}{2} + C'(\log p)e^{-p^\Delta}\Big)r^2\,. \tag{146}$$

2. $1 - p^{-1+\Delta} < r < 1$. Setting $r = 1 - x/p$, $x \in (0, p^{\Delta})$, we have

$$\begin{aligned}
\Psi_*(r) &\le \frac{1}{2}\log\Big(\frac{2x}{p}\Big) - \frac{1}{2}\log(1-e^{-2x}) + \frac{1-\varepsilon}{1+e^x}\log p \\
&\le -\frac{\varepsilon}{2}\log p + \frac{1}{2}\log(2x) - \frac{1}{2}\log(1-e^{-2x}) \\
&\le -\frac{\varepsilon}{2}\log p + \frac{1}{2}\log(2p^{\Delta}) + 1 \\
&\le \Big(\Delta - \frac{\varepsilon}{2}\Big)\log p + 2\,.
\end{aligned}$$

Therefore the claim follows by choosing $\Delta = \varepsilon/4$.

D Upper bounds on the existence threshold: Proof of Theorem 2

Note that

$$\mathsf{Sol}_{n,d}(\varepsilon) \neq \emptyset \iff \min_{\boldsymbol{x}\in\mathbb{S}^{d-1}} \max_{\boldsymbol{v}\in\mathbb{S}^{n-1}} \langle \boldsymbol{F}(\boldsymbol{x}), \boldsymbol{v}\rangle \le \sqrt{n\xi(1)\varepsilon}\,. \tag{147}$$

We define the two Gaussian processes on $\mathbb{S}^{d-1}\times\mathbb{S}^{n-1}$,

$$f_1(\boldsymbol{x},\boldsymbol{v}) := \langle \boldsymbol{F}(\boldsymbol{x}), \boldsymbol{v}\rangle + \sqrt{\xi(1)}\, z, \quad f_2(\boldsymbol{x},\boldsymbol{v}) := F_1(\boldsymbol{x}) + \sqrt{\xi(1)}\,\langle \boldsymbol{v}, \boldsymbol{g}\rangle\,,$$

where $\boldsymbol{g}\sim \mathsf{N}(0,\boldsymbol{I}_n)$ and $z\sim\mathsf{N}(0,1)$, independent of each other and $\boldsymbol{F}(\boldsymbol{x})$. Clearly,

$$\begin{aligned}
\mathbb{E} f_1(\boldsymbol{x}^1,\boldsymbol{v}^1) f_1(\boldsymbol{x}^2,\boldsymbol{v}^2) &= \xi(\langle \boldsymbol{x}^1,\boldsymbol{x}^2\rangle)\,\langle \boldsymbol{v}^1,\boldsymbol{v}^2\rangle + \xi(1)\,, \\
\mathbb{E} f_2(\boldsymbol{x}^1,\boldsymbol{v}^1) f_2(\boldsymbol{x}^2,\boldsymbol{v}^2) &= \xi(\langle \boldsymbol{x}^1,\boldsymbol{x}^2\rangle) + \xi(1)\langle \boldsymbol{v}^1,\boldsymbol{v}^2\rangle\,.
\end{aligned}$$

For $\boldsymbol{x}^1 = \boldsymbol{x}^2$ the two covariance functions coincide and for general $\boldsymbol{x}^1,\boldsymbol{x}^2\in\mathbb{S}^{d-1}$,

$$\mathbb{E} f_1(\boldsymbol{x}^1,\boldsymbol{v}^1) f_1(\boldsymbol{x}^2,\boldsymbol{v}^2) \ge \mathbb{E} f_2(\boldsymbol{x}^1,\boldsymbol{v}^1) f_2(\boldsymbol{x}^2,\boldsymbol{v}^2)\,.$$

Hence, by Gordon's Gaussian comparison inequality [Gor85], we have

$$\min_{\boldsymbol{x}\in\mathbb{S}^{d-1}} \max_{\boldsymbol{v}\in\mathbb{S}^{n-1}} f_1(\boldsymbol{x}^2,\boldsymbol{v}^2) \succeq \min_{\boldsymbol{x}\in\mathbb{S}^{d-1}} \max_{\boldsymbol{v}\in\mathbb{S}^{n-1}} f_2(\boldsymbol{x}^2,\boldsymbol{v}^2)\,,$$

where $\succeq$ denotes stochastic domination. In particular, for any $c>0$,

$$\begin{aligned}
e^{\frac{1}{2}c^2 d\xi(1)}\cdot \mathbb{E}\exp\Big\{-c\sqrt{d}\min_{\boldsymbol{x}\in\mathbb{S}^{d-1}}\max_{\boldsymbol{v}\in\mathbb{S}^{n-1}}\langle \boldsymbol{F}(\boldsymbol{x}),\boldsymbol{v}\rangle\Big\} &= \mathbb{E}\exp\Big\{-c\sqrt{d}\min_{\boldsymbol{x}\in\mathbb{S}^{d-1}}\max_{\boldsymbol{v}\in\mathbb{S}^{n-1}} f_1(\boldsymbol{x},\boldsymbol{v})\Big\} \\
&\le \mathbb{E}\exp\Big\{-c\sqrt{d}\Big(\min_{\boldsymbol{x}\in\mathbb{S}^{d-1}} F_1(\boldsymbol{x}) + \sqrt{\xi(1)}\|\boldsymbol{g}\|\Big)\Big\}\,.
\end{aligned}$$

Using Eq. (147) and Markov's inequality,

$$\begin{aligned}
\mathbb{P}\Big(\mathsf{Sol}_{n,d}(\varepsilon)\neq\emptyset\Big) &\le e^{c\sqrt{dn\xi(1)\varepsilon}}\mathbb{E}\exp\Big\{-c\sqrt{d}\min_{\boldsymbol{x}\in\mathbb{S}^{d-1}}\max_{\boldsymbol{v}\in\mathbb{S}^{n-1}}\langle \boldsymbol{F}(\boldsymbol{x}),\boldsymbol{v}\rangle\Big\} \\
&\le e^{d\varepsilon' - \frac{1}{2}c^2 d\xi(1)}\,\mathbb{E}e^{-c\sqrt{d\xi(1)}\|\boldsymbol{g}\|}\,\mathbb{E}\exp\Big\{c\sqrt{d}\max_{\boldsymbol{x}\in\mathbb{S}^{d-1}} F_1(\boldsymbol{x})\Big\}\,,
\end{aligned} \tag{148}$$

where in the second line we defined $\varepsilon' := c\sqrt{\alpha\xi(1)\varepsilon}$.

By Cramer's theorem, for $t \in (0,1]$,

$$\mathbb{P}(\|\boldsymbol{g}\| < t\sqrt{n}) = \mathbb{P}(\|\boldsymbol{g}\|^2 < t^2 n) = \exp\Big(-n\frac{t^2-1}{2} + n\log t + o(n)\Big)$$

and therefore

$$\lim_{n,d\to\infty} \frac{1}{d}\log \mathbb{E} e^{-c\sqrt{d\xi(1)}\|\boldsymbol{g}\|} = \varphi_2(c,\alpha)\,. \tag{149}$$

Let $F_1^p(\boldsymbol{x})$ be the random process as defined in (3) with $\xi(t) = t^p$ and let $z \sim \mathsf{N}(0,1)$ be independent of $F_1^p(\boldsymbol{x})$. Then, $F_1(\boldsymbol{x}) \stackrel{\mathrm{d}}{=} F_1^p(\boldsymbol{x}) + \xi_0 z$ as processes. By Markov's inequality and [Auf13, Theorem 2.8], for $t \geq 0$,

$$\mathbb{P}\Big(\max_{\boldsymbol{x}\in\mathbb{S}^{d-1}} F_1^p(\boldsymbol{x}) \geq \sqrt{d}(E+t)\Big) \leq e^{d\Theta_p(E+t)+o(d)}\,.$$

For $s \geq 0$, of course, $\mathbb{P}(\xi_0 z \geq \sqrt{d}t) \leq e^{-\frac{1}{2}\frac{dt^2}{\xi_0^2}+o(d)}$. Combining these facts, one easily sees that

$$\lim_{n,d\to\infty} \frac{1}{d}\log \mathbb{E}\exp\Big\{c\sqrt{d}\max_{\boldsymbol{x}\in\mathbb{S}^{d-1}} F_1(\boldsymbol{x})\Big\} = \varphi_1(c,p)\,. \tag{150}$$

The lemma follows from (148), (149) and (150).

E Analysis of gradient descent: Proof of Theorem 3

Since calculations are more transparent in the case of gradient flow, we will first treat this case, and then outline the modifications that arise for discrete time. Both arguments are standard in the machine learning literature [DZPS18, COB19, OS20, ADH+19, AZLL19, BMR21] and we therefore present them succinctly.

E.1 Gradient flow

For gradient flow the time $t \in \mathbb{R}_{\geq 0}$ is continuous and the state is updated according to

$$\dot{\boldsymbol{x}}(t) = -\mathsf{P}_{\mathsf{T},\boldsymbol{x}(t)}\nabla H(\boldsymbol{x}(t))\,. \tag{151}$$

(Recall the definition of cost function $H(\boldsymbol{x}) = \|\boldsymbol{F}(\boldsymbol{x})\|_2^2/2$.)

Lemma E.1. *Let $\lambda_0 := \sigma_{\min}(\boldsymbol{DF}(\boldsymbol{x}(0))|_{\mathsf{T},\boldsymbol{x}(0)})$ and $L_n := \mathrm{Lip}_\perp(\boldsymbol{DF};\mathbb{S}^{d-1})$, with the Lipschitz constant of Definition A.5. If*

$$L_n\|\boldsymbol{F}(\boldsymbol{x}(0))\|_2 < \frac{\lambda_0^2}{4}\,, \tag{152}$$

then for all $t \geq 0$,

$$\|\boldsymbol{F}(\boldsymbol{x}(t))\|_2^2 \leq \|\boldsymbol{F}(\boldsymbol{x}(0))\|_2^2\, e^{-\lambda_0^2 t/4}\,. \tag{153}$$

Proof. We define $\boldsymbol{K}(\boldsymbol{x}) := \boldsymbol{DF}(\boldsymbol{x})\mathsf{P}_{\mathsf{T},\boldsymbol{x}}\boldsymbol{DF}(\boldsymbol{x})^{\mathsf{T}}$. Then the gradient flow equation implies

$$\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{F}(\boldsymbol{x}(t)) = -\boldsymbol{K}(\boldsymbol{x}(t))\cdot\boldsymbol{F}(\boldsymbol{x}(t))\,. \tag{154}$$

Let

$$B := \Big\{ \boldsymbol{x} \in \mathbb{S}^{d-1} : \ \sigma_{\min}(\boldsymbol{DF}(\boldsymbol{x})|_{\mathsf{T},\boldsymbol{x}}) \le \frac{\lambda_0}{2} \Big\} . \tag{155}$$

Obviously $\boldsymbol{x}(0) \in B^c$. Further, letting $d_{\mathbb{S}^{d-1}}(\boldsymbol{x}_1, \boldsymbol{x}_2) := \arccos(\langle \boldsymbol{x}_1, \boldsymbol{x}_2 \rangle)$ denote the geodesic distance on the unit sphere, we have

$$d_{\mathbb{S}^{d-1}}(\boldsymbol{x}(0), B) := \inf \big\{ \, d_{\mathbb{S}^{d-1}}(\boldsymbol{x}(0), \boldsymbol{x}) : \ \boldsymbol{x} \in B \, \big\} \ge \frac{\lambda_0}{2L_n} . \tag{156}$$

Define $t_* := \inf\{t : \ \boldsymbol{x}(t) \in B\}$. For all $t \le t_*$, we have

$$\frac{\mathrm{d}}{\mathrm{d}t} \|\boldsymbol{F}(\boldsymbol{x}(t))\|_2^2 \le -\frac{\lambda_0^2}{4} \big\|\boldsymbol{F}(\boldsymbol{x}(t))\big\|_2^2 , \tag{157}$$

which implies Eq. (153) for all $t \le t_*$

We next prove that $t_* = \infty$. Indeed, note that for $t \le t_*$,

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t} \|\boldsymbol{F}(\boldsymbol{x}(t))\|_2 &= -\frac{1}{\|\boldsymbol{F}(\boldsymbol{x}(t))\|_2} \big\| \mathsf{P}_{\mathsf{T},\boldsymbol{x}(t)} \boldsymbol{DF}(\boldsymbol{x}(t))^{\mathsf{T}} \boldsymbol{F}(\boldsymbol{x}(t)) \big\|_2^2 && (158) \\
&\le -\sigma_{\min}(\boldsymbol{DF}(\boldsymbol{x}(t)) \mathsf{P}_{\mathsf{T},\boldsymbol{x}(t)}) \big\| \mathsf{P}_{\mathsf{T},\boldsymbol{x}(t)} \boldsymbol{DF}(\boldsymbol{x}(t))^{\mathsf{T}} \boldsymbol{F}(\boldsymbol{x}(t)) \big\|_2 && (159) \\
&\le -\frac{\lambda_0}{2} \big\| \mathsf{P}_{\mathsf{T},\boldsymbol{x}(t)} \boldsymbol{DF}(\boldsymbol{x}(t))^{\mathsf{T}} \boldsymbol{F}(\boldsymbol{x}(t)) \big\|_2 . && (160)
\end{aligned}$$

Further, denoting by $\boldsymbol{u}(t) \in \mathsf{T}_{\boldsymbol{x}(t)}$ the unit vector that is tangent to the geodesic between $\boldsymbol{x}(0)$ and $\boldsymbol{x}(t)$ at $\boldsymbol{x}(t)$, we have

$$\frac{\mathrm{d}}{\mathrm{d}t} d_{\mathbb{S}^{d-1}}(\boldsymbol{x}(t), \boldsymbol{x}(0)) = -\langle \mathsf{P}_{\mathsf{T},\boldsymbol{x}(t)} \boldsymbol{DF}(\boldsymbol{x}(t))^{\mathsf{T}} \boldsymbol{F}(\boldsymbol{x}(t)), \boldsymbol{u}(t) \rangle . \tag{161}$$

Therefore,

$$\frac{\mathrm{d}}{\mathrm{d}t} \Big\{ d_{\mathbb{S}^{d-1}}(\boldsymbol{x}(t), \boldsymbol{x}(0)) + \frac{2}{\lambda_0} \|\boldsymbol{F}(\boldsymbol{x}(t))\|_2 \Big\} \le 0 . \tag{162}$$

whence, for all $t \le t_*$, we have $d_{\mathbb{S}^{d-1}}(\boldsymbol{x}(t), \boldsymbol{x}(0)) \le 2\|\boldsymbol{F}(\boldsymbol{x}(0))\|_2/\lambda_0$. Recalling Eq. (152), we get $t_* = \infty$ as claimed. □

We next prove a version of Theorem 3 for gradient flow.

Theorem 8. *Consider gradient flow, as defined in Eq. (151), with respect to the energy function $H(\boldsymbol{x}) = \|\boldsymbol{F}(\boldsymbol{x})\|_2^2/2$ with $\boldsymbol{F}$ the Gaussian process defined in Section 1, and initialization $\boldsymbol{x}(0)$ independent of $\boldsymbol{F}$. Assume $n, d \to \infty$ with $n/d \to \alpha \in [0, 1)$. Define, for c_0 a sufficiently small absolute constant,*

$$\underline{\alpha}_{\mathrm{GF}}(\xi) := \frac{c_0 \xi'(1)^2}{\xi''(1)\xi(1)\big(\log(\xi^{(3)}(1)/\xi''(1)) \vee 1\big)} . \tag{163}$$

If $\alpha < \underline{\alpha}_{\mathrm{GF}}(\xi)$, then the following happens with high probability. For all $t \ge 0$

$$\|\boldsymbol{F}(\boldsymbol{x}(t))\|_2^2 \le 2n\xi(1) \, \exp\Big(-\frac{3\xi'(1)}{16} \big(\sqrt{d} - \sqrt{n}\big)^2 \cdot t \Big) . \tag{164}$$

Proof. This is an immediate application of Lemma E.1, whereby we note that

$$\|\boldsymbol{F}(\boldsymbol{x}_0)\|_2^2 = n\xi(1) + o_P(n)\,, \tag{165}$$

$$\sigma_{\min}(\boldsymbol{DF}(\boldsymbol{x}(0))|_{\mathsf{T},\boldsymbol{x}(0)}) = \sqrt{\xi'(1)}\big(\sqrt{d}-\sqrt{n}\big) + o_P(\sqrt{n})\,, \tag{166}$$

$$\mathrm{Lip}_\perp(\boldsymbol{DF};\mathbb{S}^{d-1}) \le C\sqrt{d\,\xi''(1)\log\frac{\xi^{(3)}(1)}{\xi''(1)}}\,. \tag{167}$$

where the last inequality holds with high probability for a universal constant C. The first estimate is by the law of large numbers, the second by Lemma 4.1, which implies $\boldsymbol{DF}(\boldsymbol{x})\boldsymbol{U}_{\boldsymbol{x}} = \sqrt{\xi'(q)}\,\boldsymbol{Z}$ with $\boldsymbol{Z}\sim\mathsf{GOE}(n,d-1)$ and $q=\|\boldsymbol{x}\|_2^2$ and the Bai-Yin law, and for the last one, we refer to Lemma A.6. □

E.2 Gradient descent

The analogue of Lemma E.1 for gradient descent is stated in the main text as Lemma 3.1.

Proof of Lemma 3.1. The key step is to prove that an inequality analogous to the upper bound (157) holds for all $k \le k_\star := \min\{k : \boldsymbol{x}^k \in B\}$ with B defined as per Eq. (155), which we copy here for the reader's convenience

$$B := \Big\{\boldsymbol{x}\in\mathbb{S}^{d-1} :\ \sigma_{\min}(\boldsymbol{DF}(\boldsymbol{x})|_{\mathsf{T},\boldsymbol{x}}) \le \frac{\lambda_0}{2}\Big\}\,. \tag{168}$$

To this end first note that

$$\big\|\boldsymbol{z}^{k+1}-\boldsymbol{x}^k\big\| = \eta\big\|\mathsf{P}_{\mathsf{T}}\boldsymbol{DF}(\boldsymbol{x}^k)^{\mathsf{T}}\boldsymbol{F}(\boldsymbol{x}^k)\big\|_2 \le \varepsilon_0\,. \tag{169}$$

(In what follows we omit the reference to the point on the sphere from the projector P_{T}.)

Further there exist $\boldsymbol{\xi}_{(i)}^k \in [\boldsymbol{x}^k, \boldsymbol{z}^{k+1}]$ (here $[\boldsymbol{u},\boldsymbol{v}]$ denotes the segment between $\boldsymbol{u}$ and $\boldsymbol{v}\in\mathbb{R}^d$) such that

$$\begin{aligned} F_i(\boldsymbol{z}^{k+1}) &= F_i(\boldsymbol{x}^k) - \eta\nabla F_i(\boldsymbol{x}^k)^{\mathsf{T}}\mathsf{P}_{\mathsf{T}}\boldsymbol{DF}(\boldsymbol{x}^k)^{\mathsf{T}}\boldsymbol{F}(\boldsymbol{x}^k) - \eta[\nabla F_i(\boldsymbol{\xi}_{(i)}^k) - \nabla F_i(\boldsymbol{x}^k)]^{\mathsf{T}}\mathsf{P}_{\mathsf{T}}\boldsymbol{DF}(\boldsymbol{x}^k)^{\mathsf{T}}\boldsymbol{F}(\boldsymbol{x}^k) \\ &=: F_i(\boldsymbol{x}^k) - \eta\nabla F_i(\boldsymbol{x}^k)^{\mathsf{T}}\mathsf{P}_{\mathsf{T}}\boldsymbol{DF}(\boldsymbol{x}^k)^{\mathsf{T}}\boldsymbol{F}(\boldsymbol{x}^k) + \overline{\Delta_i^k}\,. \end{aligned}$$

We have the estimate

$$|\overline{\Delta_i^k}| \le \eta J_n\cdot\|\boldsymbol{z}^{k+1}-\boldsymbol{x}^k\|_2\cdot\|\mathsf{P}_{\mathsf{T}}\boldsymbol{DF}(\boldsymbol{x}^k)^{\mathsf{T}}\boldsymbol{F}(\boldsymbol{x}^k)\|_2 \tag{170}$$

$$\le J_n\eta^2\|\mathsf{P}_{\mathsf{T}}\boldsymbol{DF}(\boldsymbol{x}^k)^{\mathsf{T}}\boldsymbol{F}(\boldsymbol{x}^k)\|_2^2\,. \tag{171}$$

Further by Pythagoras' theorem $\|\boldsymbol{z}^{k+1}\|_2^2 = 1+\|\boldsymbol{z}^{k+1}-\boldsymbol{x}^k\|^2$, whence, for some $\boldsymbol{\zeta}_{(i)}^k\in[\boldsymbol{z}^{k+1},\boldsymbol{x}^{k+1}]$,

$$\begin{aligned} \big|F_i(\boldsymbol{x}^{k+1}) - F_i(\boldsymbol{z}^{k+1})\big| &= \big|\langle\nabla F_i(\boldsymbol{\zeta}_{(i)}^k), \boldsymbol{x}^{k+1}\rangle\big|\cdot\big|\|\boldsymbol{z}^{k+1}\|_2 - 1\big| \\ &\le \sup_{\boldsymbol{x}\in\mathsf{B}^d(1+\varepsilon_0)}\|\nabla F_i(\boldsymbol{x})\|_2\cdot\|\boldsymbol{z}^{k+1}-\boldsymbol{x}^k\|_2^2 \\ &\le M_n\eta^2\big\|\mathsf{P}_{\mathsf{T}}\boldsymbol{DF}(\boldsymbol{x}^k)^{\mathsf{T}}\boldsymbol{F}(\boldsymbol{x}^k)\big\|_2^2\,. \end{aligned}$$

We therefore obtain that

$$\boldsymbol{F}(\boldsymbol{x}^{k+1}) = \boldsymbol{F}(\boldsymbol{x}^k) - \eta\boldsymbol{DF}(\boldsymbol{x}^k)\mathsf{P}_{\mathsf{T}}\boldsymbol{DF}(\boldsymbol{x}^k)^{\mathsf{T}}\boldsymbol{F}(\boldsymbol{x}^k) + \boldsymbol{\Delta}^k\,, \tag{172}$$

$$\begin{aligned}\|\boldsymbol{\Delta}^k\|_2 &\le \sqrt{n}(J_n+M_n)\eta^2\big\|\mathsf{P}_{\mathsf{T}}\boldsymbol{DF}(\boldsymbol{x}^k)^{\mathsf{T}}\boldsymbol{F}(\boldsymbol{x}^k)\big\|_2^2\\ &\le \frac{\eta}{10\max_{\boldsymbol{x}\in\mathbb{S}^{d-1}}\|\boldsymbol{F}(\boldsymbol{x})\|_2}\big\|\mathsf{P}_{\mathsf{T}}\boldsymbol{DF}(\boldsymbol{x}^k)^{\mathsf{T}}\boldsymbol{F}(\boldsymbol{x}^k)\big\|_2^2\,,\end{aligned}\tag{173}$$

where the last inequality follows from Eq. (27). Further

$$\begin{aligned}\big\|\boldsymbol{F}(\boldsymbol{x}^k)-\eta\boldsymbol{DF}(\boldsymbol{x}^k)\mathsf{P}_{\mathsf{T}}\boldsymbol{DF}(\boldsymbol{x}^k)^{\mathsf{T}}\boldsymbol{F}(\boldsymbol{x}^k)\big\|_2 &\le \|\boldsymbol{F}(\boldsymbol{x}^k)\|_2+\eta M_n\|\mathsf{P}_{\mathsf{T}}\boldsymbol{DF}(\boldsymbol{x}^k)^{\mathsf{T}}\boldsymbol{F}(\boldsymbol{x}^k)\big\|_2\\ &\le (1+\eta M_n^2)\|\boldsymbol{F}(\boldsymbol{x}^k)\|_2\\ &\le 2\|\boldsymbol{F}(\boldsymbol{x}^k)\|_2\,.\end{aligned}\tag{174}$$

Again, in the last step, we used condition (27). Also, note that by Eq. (173) and condition (27), we have

$$\|\boldsymbol{\Delta}^k\|_2 \le \frac{\eta}{8\max_{\boldsymbol{x}\in\mathbb{S}^{d-1}}\|\boldsymbol{F}(\boldsymbol{x})\|_2}M_n^2\|\boldsymbol{F}(\boldsymbol{x}^k)\|_2^2 \le \frac{1}{10}\|\boldsymbol{F}(\boldsymbol{x}^k)\|_2\,.\tag{175}$$

Using Eqs. (172) and (174), we get

$$\|\boldsymbol{F}(\boldsymbol{x}^{k+1})\|_2^2 \le \big\|\boldsymbol{F}(\boldsymbol{x}^k)-\eta\boldsymbol{DF}(\boldsymbol{x}^k)\mathsf{P}_{\mathsf{T}}\boldsymbol{DF}(\boldsymbol{x}^k)^{\mathsf{T}}\boldsymbol{F}(\boldsymbol{x}^k)\big\|_2^2+4\|\boldsymbol{\Delta}^k\|_2\|\boldsymbol{F}(\boldsymbol{x}^k)\|_2+\|\boldsymbol{\Delta}^k\|_2^2\,.$$

Using Eqs. (172) to (175), we get

$$\begin{aligned}\|\boldsymbol{F}(\boldsymbol{x}^{k+1})\|_2^2 &\le \big\|\boldsymbol{F}(\boldsymbol{x}^k)-\eta\boldsymbol{DF}(\boldsymbol{x}^k)\mathsf{P}_{\mathsf{T}}\boldsymbol{DF}(\boldsymbol{x}^k)^{\mathsf{T}}\boldsymbol{F}(\boldsymbol{x}^k)\big\|_2^2+5\|\boldsymbol{\Delta}^k\|_2\|\boldsymbol{F}(\boldsymbol{x}^k)\|_2\\ &\le \big\|\boldsymbol{F}(\boldsymbol{x}^k)-\eta\boldsymbol{DF}(\boldsymbol{x}^k)\mathsf{P}_{\mathsf{T}}\boldsymbol{DF}(\boldsymbol{x}^k)^{\mathsf{T}}\boldsymbol{F}(\boldsymbol{x}^k)\big\|_2^2+\frac{\eta}{2}\big\|\mathsf{P}_{\mathsf{T}}\boldsymbol{DF}(\boldsymbol{x}^k)^{\mathsf{T}}\boldsymbol{F}(\boldsymbol{x}^k)\big\|_2^2\\ &\le \langle\boldsymbol{F}(\boldsymbol{x}^k),\Big(\boldsymbol{I}-\frac{3}{2}\eta\boldsymbol{K}(\boldsymbol{x}^k)+\eta^2\boldsymbol{K}(\boldsymbol{x}^k)^2\Big)\boldsymbol{F}(\boldsymbol{x}^k)\rangle\,.\end{aligned}\tag{176}$$

Since $\eta\lambda_{\max}(\boldsymbol{K}(\boldsymbol{x}_k))\le\eta M_n^2\le 1$ by Eq. (27), we conclude that

$$\|\boldsymbol{F}(\boldsymbol{x}^{k+1})\|_2^2 \le \Big(1-\frac{1}{2}\eta\lambda_{\min}(\boldsymbol{K}(\boldsymbol{x}^k))\Big)\|\boldsymbol{F}(\boldsymbol{x}^k)\|_2^2\,,$$

and therefore, for any $k\le k_\star$,

$$\|\boldsymbol{F}(\boldsymbol{x}^k)\|_2^2 \le e^{-\eta\lambda_0^2 k/8}\|\boldsymbol{F}(\boldsymbol{x}^0)\|_2^2\,.$$

We are left with the task of proving that $k_\star=\infty$. To this end we proceed as in the case of gradient flow. Namely, we note that, for $k\le k_\star$ Eq. (176) implies (for $\eta\lambda_{\max}(\boldsymbol{K}(\boldsymbol{x}_k))\le 1$)

$$\|\boldsymbol{F}(\boldsymbol{x}^{k+1})\|_2^2 \le \|\boldsymbol{F}(\boldsymbol{x}^k)\|_2^2-\frac{\eta}{2}\big\|\mathsf{P}_{\mathsf{T}}\boldsymbol{DF}(\boldsymbol{x}^k)^{\mathsf{T}}\boldsymbol{F}(\boldsymbol{x}^k)\big\|_2^2\tag{177}$$

$$\le \|\boldsymbol{F}(\boldsymbol{x}^k)\|_2^2-\frac{\lambda_0\eta}{4}\big\|\mathsf{P}_{\mathsf{T}}\boldsymbol{DF}(\boldsymbol{x}^k)^{\mathsf{T}}\boldsymbol{F}(\boldsymbol{x}^k)\big\|_2\|\boldsymbol{F}(\boldsymbol{x}^k)\|_2\,,\tag{178}$$

and therefore

$$\|\boldsymbol{F}(\boldsymbol{x}^{k+1})\|_2 \le \|\boldsymbol{F}(\boldsymbol{x}^k)\|_2-\frac{\lambda_0\eta}{8}\big\|\mathsf{P}_{\mathsf{T}}\boldsymbol{DF}(\boldsymbol{x}^k)^{\mathsf{T}}\boldsymbol{F}(\boldsymbol{x}^k)\big\|_2\,.\tag{179}$$

Further

$$d_{\mathbb{S}^{d-1}}(\boldsymbol{x}^{k+1},\boldsymbol{x}^0) \le d_{\mathbb{S}^{d-1}}(\boldsymbol{x}^{k+1},\boldsymbol{x}^0)+\eta\big\|\mathsf{P}_{\mathsf{T}}\boldsymbol{DF}(\boldsymbol{x}^k)^{\mathsf{T}}\boldsymbol{F}(\boldsymbol{x}^k)\big\|_2\,.\tag{180}$$

Therefore, defining $\Psi(\boldsymbol{x}):=d_{\mathbb{S}^{d-1}}(\boldsymbol{x},\boldsymbol{x}^0)+(8/\lambda_0)\|\boldsymbol{F}(\boldsymbol{x})\|_2$, we have $\Psi(\boldsymbol{x}^{k+1})\le\Psi(\boldsymbol{x}^k)$, whence the proof follows. □

We are now in position to prove Theorem 3.

Proof of Theorem 3. The proof consists in using the estimates of Section A to check the assumptions of Lemma 3.1.

More precisely, the assumption in (26) holds with high probability under condition (29), using the estimates (165) to (167) that we obtained in the case of gradient flow. The convergence rate (30) follows from Eqs. (28) and (166).

Finally, the assumption in (27) on the stepsize, holds with high probability when $\eta < 1/(C_1 d)$ with C_1 large enough, because the following inequalities hold with high probability for $C = C(\xi)$ a sufficiently large constant:

$$\max_{\boldsymbol{x}\in\mathbb{S}^{d-1}} \left\|\boldsymbol{F}(\boldsymbol{x})\right\|_2 \le C\sqrt{d}\,, \tag{181}$$

$$\max_{\boldsymbol{x}\in\mathbb{S}^{d-1}} \left\|\mathsf{P}_{\mathsf{T}}\boldsymbol{D}\boldsymbol{F}(\boldsymbol{x})^{\mathsf{T}}\boldsymbol{F}(\boldsymbol{x})\right\|_2 \le C\,d\,, \tag{182}$$

$$J_n \vee M_n \le C\,\sqrt{d}\,. \tag{183}$$

These inequalities follow from Propositions A.1 and A.4, thus completing the proof. □

F Analysis of Hessian descent: Proof of Theorem 4

We will collect some random matrix theory results (mainly about the distribution of the Hessian) in Section F.1, and use them to prove Theorem 4 in Section F.2.

F.1 Random matrix theory

Throughout, for $M \le N$, we let $\boldsymbol{W} \sim \mathsf{GOE}(N)$ independent of $\boldsymbol{Z} \sim \mathsf{GOE}(M,N)$ and define

$$\boldsymbol{A} = \boldsymbol{A}_{M,N} := a\sqrt{N}\,\boldsymbol{W} + b\boldsymbol{Z}^{\mathsf{T}}\boldsymbol{Z}\,. \tag{184}$$

Lemma F.1. *Assume $a, b, \alpha \in \mathbb{R}_{\ge 0}$, and recall the definition of Q, z_* from Theorem 4, namely*

$$Q(m;\alpha,a,b) := -\frac{1}{m} + \frac{\alpha b}{1+bm} - a^2 m\,, \tag{185}$$

$$z_*(\alpha,a,b) := -\sup_{m>0} Q(m;\alpha,a,b)\,. \tag{186}$$

Further, let $S(\,\cdot\,;\alpha,a,b) : \mathbb{H} \to \mathbb{C}$ be the only analytic function on the upper half plane, such that: (i) $S(\,\cdot\,;\alpha,a,b) = -1/z + o(1/z)$ as $z \to i\infty$; (ii) $S(z;\alpha,a,b)$ solves $Q(S;\alpha,a,b) = z$.

Then the following hold almost surely in the limit $M, N \to \infty$ with $M/N \to \alpha \in (0,1]$.

1. $$\lim_{M,N\to\infty} \frac{1}{N}\lambda_{\min}(\boldsymbol{A}_{M,N}) = -z_*(\alpha,a,b)\,. \tag{187}$$

2. *For $z \in \mathbb{H}$ (the upper half complex plane)*
$$\lim_{M,N\to\infty} \frac{1}{N}\operatorname{Tr}\left((\boldsymbol{A}_{M,N}/N - z\boldsymbol{I})^{-1}\right) = S(z;\alpha,a,b)\,. \tag{188}$$

3. *If $a > 0$, $\alpha < 1$ then,*
$$2a\sqrt{1-\alpha} < z_*(\alpha,a,b) < 2a\,. \tag{189}$$

4. *Letting $\hat{\nu}_{M,N}(\,\cdot\,;a,b)$ denote the empirical spectral distribution of $\boldsymbol{A}_{M,N}/N$, there exists a non-decreasing, deterministic function $\nu_0(\,\cdot\,;a,b):\mathbb{R}\to\mathbb{R}_{\ge 0}$ such that $\nu_0(t;a,b)>0$ for all $a,b\ge 0$ and $z_*=z_*(\alpha,a,b)$, almost surely*

$$\lim_{M,N\to\infty}\hat{\nu}_{M,N}(-z_*+t;a,b)\ge\nu_0(t;a,b)\,. \tag{190}$$

5. *For any $a_0,b_0>0$, $t>t'>0$, we have*

$$\lim_{M,N\to\infty}\inf_{a\in[0,a_0],b\in[0,b_0]}\hat{\nu}_{M,N}(-z_*+t;a,b)\ge\nu_{\min}(t';a_0,b_0)\,, \tag{191}$$

$$\nu_{\min}(t';a_0,b_0):=\inf_{a\in[0,a_0],b\in[0,b_0]}\nu_0(t';a,b)>0\,. \tag{192}$$

Proof. The asymptotic R-transforms of the random matrices $\boldsymbol{X}:=\boldsymbol{W}/\sqrt{N}$, $\boldsymbol{Y}:=\boldsymbol{Z}^{\mathsf{T}}\boldsymbol{Z}/N$ are [MS17]

$$R_{\boldsymbol{X}}(z)=z\,,\qquad R_{\boldsymbol{Y}}(z)=\frac{\alpha}{1-z}\,. \tag{193}$$

Since $\boldsymbol{X}$, $\boldsymbol{Y}$ are asymptotically free,

$$R_{\boldsymbol{A}/N}(z)=aR_{\boldsymbol{X}}(az)+bR_{\boldsymbol{Y}}(bz) \tag{194}$$

$$=a^2z+\frac{\alpha b}{1-bz}\,. \tag{195}$$

Therefore, we have $Q(m;\alpha,a,b)=R_{\boldsymbol{A}/N}(-m)-m^{-1}$. Point 2 follows then from the standard connection between Stieltjes transform and S-transform.

Point 1 follows from [CDMFF11].

For part 3, Eq. (189) follows from the inequalities

$$-\frac{1}{m}-a^2m<Q(m;\alpha,a,b)<-\frac{1}{m}+\frac{\alpha}{m}-a^2m\,. \tag{196}$$

For part 4, note that $Q(S;\alpha,a,b)=z$ is a third order algebraic equation for the Stieltjes transform S, with coefficients that depend continuously on a,b,z. The equation reduces to a second order one if $a=0$ or $b=0$. By the definition of z_*, there exists $r>0$ such that, for $z\in(-z_*,-z_*+r)$ this equation has three solutions, of which two are complex conjugates. The imaginary part of these solutions gives (up to a constant factor) the asymptotic density of empirical spectral distribution, which is strictly positive, hence implying the claim (see e.g. [AGZ09, Theorem 2.4.3]).

Finally, we consider part 5. On the favorable event $\mathcal{G}:=\{\|\boldsymbol{W}\|_F^2\le 2N^2,\|\boldsymbol{Z}^{\mathsf{T}}\boldsymbol{Z}\|_F^2\le 2N^3\}$ (which holds with probability at least $1-\exp(-cN)$), we have

$$\Big\|\frac{1}{N}\boldsymbol{A}_{M,N}(a_1,b_1)-\frac{1}{N}\boldsymbol{A}_{M,N}(a_2,b_2)\Big\|_F\le 4\sqrt{N}\|(a_1,b_1)-(a_2,b_2)\|_2\,, \tag{197}$$

where we noted explicitly the dependence of $\boldsymbol{A}_{M,N}$ on parameters a,b. Hence, by the Wielandt-Hoffman inequality, for any Lipschitz function $f:\mathbb{R}\to\mathbb{R}$, the following holds on $\mathcal{F}$ for all $a_1,a_2,b_1,b_2\ge 0$:

$$\Big|\int f(x)\,\hat{\nu}_{M,N}(\mathrm{d}x;a_1,b_1)-\int f(x)\,\hat{\nu}_{M,N}(\mathrm{d}x;a_2,b_2)\Big|\le 4\|f\|_{\mathrm{Lip}}\cdot\|(a_1,b_1)-(a_2,b_2)\|_2\,. \tag{198}$$

Using the function

$$f_{t,\varepsilon}(x) = \begin{cases} 1 & \text{if } x \le -z_* + t - \varepsilon, \\ (-z_* + t - x)/\varepsilon & \text{if } -z_* + t - \varepsilon < x < -z_* + t, \\ 0 & \text{if } -z_* + t \le x, \end{cases} \tag{199}$$

we get, for all $a_1, a_2, b_1, b_2 \ge 0$:

$$\begin{aligned} \hat{\nu}_{M,N}(-z_* + t; a_1, b_1) &\ge \int f_{t,\varepsilon}(x)\, \hat{\nu}_{M,N}(\mathrm{d}x; a_1, b_1) \\ &\ge \int f_{t,\varepsilon}(x)\, \hat{\nu}_{M,N}(\mathrm{d}x; a_2, b_2) - \frac{4}{\varepsilon}\|(a_1, b_1) - (a_2, b_2)\|_2 \\ &\ge \hat{\nu}_{M,N}(-z_* + t - \varepsilon; a_2, b_2) - \frac{4}{\varepsilon}\|(a_1, b_1) - (a_2, b_2)\|_2 \,. \end{aligned}$$

Let S_δ be a finite δ-net in $[0, a_0] \times [0, b_0]$. Using the last inequality and the result at point 4 on S_δ, alongside Borel-Cantelli (which implies that $\mathcal{G}$ holds eventually almost surely), we get

$$\lim_{M,N\to\infty} \inf_{a\in[0,a_0], b\in[0,b_0]} \hat{\nu}_{M,N}(-z_* + t; a, b) \ge \min_{(a,b)\in S_\delta} \nu_0(t - \varepsilon; a, b) - \frac{4\delta}{\varepsilon} \,.$$

The lower bound (191) follows by taking $\delta \to 0$.

Finally to prove Eq. (192), i.e. $\nu_{\min}(t'; a, b) > 0$ strictly, we note that

$$\nu_0(t'; a, b) \ge \nu_0(f_{t',\varepsilon}; a, b) := \int f_{t',\varepsilon}(x)\, \nu_0(\mathrm{d}x; a, b) \,. \tag{200}$$

Further, $(a, b) \mapsto \nu_0(f_{t',\varepsilon}; a, b)$ is Lipschitz continuous by point 4 and the argument given above (with Lipschitz modulus $4/\varepsilon$), and $\nu_0(f_{t',\varepsilon}; a, b) \ge \nu_0(t' - \varepsilon; a, b) > 0$ for any a, b. Therefore, by taking $\varepsilon \in (0, t')$, we get $\nu_{\min}(t'; a, b) > \inf_{a,b\in[0,a_0]\times[0,b_0]} \nu_0(f_{t',\varepsilon}; a, b) > 0$. □

Lemma F.2. *Let $a_0, b_0 \in \mathbb{R}_{\ge 0}$, $\alpha \in (0, 1]$ and, for $(a, b) \in [0, a_0] \times [0, b_0]$ let $z_* = z_*(\alpha; a, b)$ be defined as in Lemma F.1. Let $M = M(N)$ be a sequence such that $M/N \to \alpha$ as $N \to \infty$. Then for any fixed $a_0, b_0, t, c > 0$, there exists $C(t) = C(t, c, a_0, b_0, \alpha)$ such that, if $(a, b) \in [0, a_0] \times [0, b_0]$ then for large enough N,*

$$\mathbb{P}\big(\lambda_{\min}(\boldsymbol{A}_{M,N}) \ge N(-z_* + t)\big) \le C(t)\, e^{-N^2/C(t)} \,. \tag{201}$$

Proof. Throughout the proof, we denote by C constants that depend on a_0, b_0, α, t, c, and possibly other quantities as indicated, but not on a, b.

Define $\boldsymbol{W}/\sqrt{N} := \frac{1}{\sqrt{2}}(\boldsymbol{G} + \boldsymbol{G}^\mathsf{T})$ and $\boldsymbol{S} := \boldsymbol{Z}/\sqrt{N}$ so that $N^{-1}\boldsymbol{A}_{M,N} = \frac{a}{\sqrt{2}}(\boldsymbol{G} + \boldsymbol{G}^\mathsf{T}) + b\boldsymbol{S}^\mathsf{T}\boldsymbol{S}$, and $\boldsymbol{G}, \boldsymbol{S}$ are matrices containing i.i.d. random variables $\mathsf{N}(0, 1/N)$. Further, for $\varepsilon > 0$, define

$$\boldsymbol{B}^\varepsilon := \frac{a}{\sqrt{2}}(\boldsymbol{G} + \boldsymbol{G}^\mathsf{T}) + b\boldsymbol{M}^\varepsilon(\boldsymbol{S}) \,, \tag{202}$$

$$\boldsymbol{M}^\varepsilon(\boldsymbol{S}) := \boldsymbol{S}^\mathsf{T}(\boldsymbol{I} + \varepsilon\boldsymbol{S}\boldsymbol{S}^\mathsf{T})^{-1}\boldsymbol{S} \,. \tag{203}$$

Since $N^{-1}\boldsymbol{A}_{M,N} \succeq \boldsymbol{B}^\varepsilon$, it is sufficient to prove the following claim:

> **Claim.** For any $a_0, b_0, \alpha, t, c > 0$, there exists $\varepsilon, C \ge 0$ such that $\mathbb{P}\big(\lambda_{\min}(\boldsymbol{B}^\varepsilon) \ge -z_* + t\big) \le C\, \exp(-N^2/C)$ for all $(a, b) \in [0, a_0] \times [0, b_0]$ and large N.

In order to prove this claim, we view $\boldsymbol{B}^\varepsilon$ as a function of the $N^2 + NM$ random variables (G_{ij}, S_{ij}). Note that

$$\|\boldsymbol{B}^\varepsilon(\boldsymbol{G}_1, \boldsymbol{S}_1) - \boldsymbol{B}^\varepsilon(\boldsymbol{G}_2, \boldsymbol{S}_2)\|_F \le 2a\|\boldsymbol{G}_1 - \boldsymbol{G}_2\|_F + b\|\boldsymbol{M}^\varepsilon(\boldsymbol{S}_1) - \boldsymbol{M}^\varepsilon(\boldsymbol{S}_2)\|_F\,. \tag{204}$$

Further

$$\begin{aligned}
\|\boldsymbol{M}^\varepsilon(\boldsymbol{S}_1) - \boldsymbol{M}^\varepsilon(\boldsymbol{S}_2)\|_F &\overset{(a)}{=} \frac{1}{\varepsilon}\big\|(\boldsymbol{I} + \varepsilon\boldsymbol{S}_1^\mathsf{T}\boldsymbol{S}_1)^{-1} - (\boldsymbol{I} + \varepsilon\boldsymbol{S}_2^\mathsf{T}\boldsymbol{S}_2)^{-1}\big\|_F \\
&\overset{(b)}{=} \big\|(\boldsymbol{I} + \varepsilon\boldsymbol{S}_1^\mathsf{T}\boldsymbol{S}_1)^{-1}(\boldsymbol{S}_1^\mathsf{T}\boldsymbol{S}_1 - \boldsymbol{S}_2^\mathsf{T}\boldsymbol{S}_2)(\boldsymbol{I} + \varepsilon\boldsymbol{S}_2^\mathsf{T}\boldsymbol{S}_2)^{-1}\big\|_F \\
&\overset{(c)}{\le} \big\|(\boldsymbol{I} + \varepsilon\boldsymbol{S}_1^\mathsf{T}\boldsymbol{S}_1)^{-1}\boldsymbol{S}_1^\mathsf{T}(\boldsymbol{S}_1 - \boldsymbol{S}_2)\big\|_F + \big\|(\boldsymbol{I} + \varepsilon\boldsymbol{S}_2^\mathsf{T}\boldsymbol{S}_2)^{-1}\boldsymbol{S}_2^\mathsf{T}(\boldsymbol{S}_1 - \boldsymbol{S}_2)\big\|_F \\
&\le \Big(\big\|(\boldsymbol{I} + \varepsilon\boldsymbol{S}_1^\mathsf{T}\boldsymbol{S}_1)^{-1}\boldsymbol{S}_1^\mathsf{T}\big\|_{\rm op} + \big\|(\boldsymbol{I} + \varepsilon\boldsymbol{S}_2^\mathsf{T}\boldsymbol{S}_2)^{-1}\boldsymbol{S}_2^\mathsf{T}\big\|_{\rm op}\Big)\big\|\boldsymbol{S}_1 - \boldsymbol{S}_2\big\|_F \\
&\overset{(d)}{\le} \frac{1}{\sqrt{\varepsilon}}\big\|\boldsymbol{S}_1 - \boldsymbol{S}_2\big\|_F\,.
\end{aligned}$$

The equality (a) can be proved by using the singular value decomposition of $\boldsymbol{S}_i$ to show that the two matrices whose Frobenius norm is computed define the same bilinear form. To verify (b) add and subtract $\boldsymbol{I}$ from the middle term. For (c) add and subtract $\boldsymbol{S}_1^\mathsf{T}\boldsymbol{S}_2$ from the middle term in the left-hand side of the inequality and use the fact that $\boldsymbol{I} + \varepsilon\boldsymbol{S}_i^\mathsf{T}\boldsymbol{S}_i \succeq \boldsymbol{I}$. Finally, for (d) we use the elementary inequality $x/(1+\varepsilon x^2) \le \varepsilon^{-1/2}/2$.

Recall that, by [AGZ09, Lemma 2.3.1], for any L-Lipschitz function $f : \mathbb{R} \to \mathbb{R}$, $N^{-1}\sum_{i=1}^N f(\lambda_i(\boldsymbol{B}^\varepsilon))$ is $\sqrt{2/N}\cdot L$ Lipschitz function of $\boldsymbol{B}^\varepsilon$. Therefore, by the above and Gaussian concentration, for any such function, any $u \ge 0$, and any $(a, b) \in [0, a_0] \times [0, b_0]$

$$\mathbb{P}\Big(\Big|\mathcal{F}(\boldsymbol{B}^\varepsilon) - \mathbb{E}\Big\{\mathcal{F}(\boldsymbol{B}^\varepsilon)\Big\}\Big| \ge u\Big) \le 2\,\exp\Big(-\frac{\varepsilon N^2 u^2}{C(a_0, b_0)L^2}\Big)\,, \tag{205}$$

$$\mathcal{F}(\boldsymbol{B}^\varepsilon) := \frac{1}{N}\sum_{i=1}^N f(\lambda_i(\boldsymbol{B}^\varepsilon))\,, \tag{206}$$

where we can take $C(a_0, b_0) = C_0(a_0^2 + b_0^2)$ for a suitable numerical constant $C_0 > 0$.

We next take

$$f(x) = \begin{cases} 0 & \text{if } -z_* + t \le x, \\ (2/t)(-x - z_* + t) & \text{if } -z_* + t/2 < x < -z_* + t, \\ 1 & \text{if } x \le -z_* + t/2. \end{cases} \tag{207}$$

We have

$$\begin{aligned}
\frac{1}{N}\sum_{i=1}^N f(\lambda_i(\boldsymbol{B}^\varepsilon)) &\ge \frac{1}{N}\sum_{i=1}^N f(\lambda_i(\boldsymbol{B}^0)) - \frac{2}{t}\cdot\frac{1}{N}\sum_{i=1}^N |\lambda_i(\boldsymbol{B}^0) - \lambda_i(\boldsymbol{B}^\varepsilon)| && (208)\\
&\overset{(a)}{\ge} \frac{1}{N}\sum_{i=1}^N \mathbf{1}\big\{\lambda_i(\boldsymbol{B}^0) \le -z_* + t/2\big\} - \frac{2b_0}{t\sqrt{N}}\|\boldsymbol{B}^0 - \boldsymbol{B}^\varepsilon\|_F && (209)\\
&\overset{(b)}{\ge} \nu_{\min}(t/3; a_0, b_0) - \frac{2b_0}{t\sqrt{N}}\big\|\varepsilon(\boldsymbol{S}^\mathsf{T}\boldsymbol{S})^2(\boldsymbol{I} + \varepsilon\boldsymbol{S}^\mathsf{T}\boldsymbol{S})^{-1}\big\|_F && (210)\\
&\ge \nu_{\min}(t/3; a_0, b_0) - \frac{2b_0}{t}\varepsilon\|\boldsymbol{S}\|_{\rm op}^4\,. && (211)
\end{aligned}$$

Here in (a) we used Wielandt-Hoffman and (b) holds eventually almost surely, by part 5 of Lemma F.1. Recall that $\|\boldsymbol{S}\|_{\rm op} \le 2 + \sqrt{\alpha} \le 3$ with probability $1 - \exp(-cN)$. Hence, taking $\varepsilon = t\nu_{\min}(t/3; a_0, b_0)/(4 \cdot 3^4 b_0)$, we get that, with high probability

$$\frac{1}{N}\sum_{i=1}^{N} f(\lambda_i(\boldsymbol{B}^{\varepsilon})) \ge \frac{1}{2}\nu_{\min}(t/3; a_0, b_0)\,. \tag{212}$$

Since f is $2/t$ Lipschitz, Eq. (205) implies

$$\mathbb{P}\left(\frac{1}{N}\sum_{i=1}^{N} f(\lambda_i(\boldsymbol{B}^{\varepsilon})) \le \frac{1}{2}\nu_{\min}(t/3; a_0, b_0) - u\right) \le 2\,\exp\Big(-\frac{\varepsilon N^2 t^2 u^2}{C(a_0, b_0)}\Big) \tag{213}$$

$$\le 2\,\exp\Big(-\frac{N^2 t^4 \nu_{\min}(t/3; a_0, b_0) u^2}{C'(a_0, b_0)}\Big)\,, \tag{214}$$

and the proof of the claim follows by noting that

$$\mathbb{P}\left(\lambda_{\min}(\boldsymbol{B}^{\varepsilon}) \ge -z_* + t\right) \le \mathbb{P}\left(\frac{1}{N}\sum_{i=1}^{N} f(\lambda_i(\boldsymbol{B}^{\varepsilon})) \le 0\right)\,. \tag{215}$$

□

Lemma F.3. *Assume we let $n, d \to \infty$ with $n/d \to \alpha \in (0, 1]$. Recall the definition of Hamiltonian $H(\boldsymbol{x}) := \|\boldsymbol{F}(\boldsymbol{x})\|_2^2/2$, and that $\mathsf{T}_{\boldsymbol{x}}$ denotes the tangent space to the sphere of radius $\|\boldsymbol{x}\|_2$ at $\boldsymbol{x}$.*

Let $\lambda(\boldsymbol{x}) := \lambda_{\min}(\nabla^2 H(\boldsymbol{x})|_{\mathsf{T},\boldsymbol{x}})$ and define

$$z_0(\boldsymbol{x}) := -z_*(\alpha; \sqrt{2\alpha\xi''(\|\boldsymbol{x}\|^2)H(\boldsymbol{x})/n}, \xi'(\|\boldsymbol{x}\|^2))\,. \tag{216}$$

Then, for any $\varepsilon > 0$ there exists a constant $C(\varepsilon) > 0$ such that for large enough n and d,

$$\mathbb{P}\Big(\forall \boldsymbol{x} \in \mathsf{B}^d(1) : \lambda(\boldsymbol{x}) \le d(z_0(\boldsymbol{x}) + \varepsilon)\Big) \ge 1 - C(\varepsilon)\, e^{-d/C(\varepsilon)}\,. \tag{217}$$

Proof. Lemmas 4.1 and F.2 imply that for any $\varepsilon, C_0 > 0$, there exists a constant $C(C_0, \varepsilon) > 0$ (depending on ξ as well) such that, for large n, d for any $\boldsymbol{x} \in \mathsf{B}^d(1)$

$$\mathbb{P}\big(\lambda(\boldsymbol{x}) \ge d(z_0(\boldsymbol{x}) + \varepsilon); H(\boldsymbol{x}) \le C_0 d\big) \le C(C_0, \varepsilon)\, e^{-2d^2/C(C_0,\varepsilon)}\,. \tag{218}$$

Applying Proposition A.1, Lemma A.6, and Weyl's inequality, we can work on the event

$$\mathcal{G} := \Big\{ \max_{\boldsymbol{x}\in\mathsf{B}^d(1)} H(\boldsymbol{x}) \le C_0 d,\ \max_{\boldsymbol{x}\in\mathsf{B}^d(1)} \|\boldsymbol{DF}|_{\mathsf{T},\boldsymbol{x}}\|_{\rm op} \le C_0\sqrt{d},\ \mathrm{Lip}(H; \mathsf{B}^d(\boldsymbol{0}; 1)) \le C_0\, d,$$
$$\mathrm{Lip}(\boldsymbol{DF}; \mathsf{B}^d(1)) \le C_0\, d,\ \mathrm{Lip}(\lambda; \mathsf{B}^d(1)) \le C_0\, d\Big\}\,,$$

for C_0 a sufficiently large constant (dependent on ξ). Indeed $\mathbb{P}(\mathcal{G}) \ge 1 - C_1 e^{-d/C_1}$ for a certain constant C_1 and large n, d. Further, on the same event $\mathrm{Lip}(z_0; \mathsf{B}^d(1)) \le C_0$ (eventually enlarging C_0).

Let $N^d(\delta_d)$ be a δ_d-net in $\mathsf{B}^d(1)$, with $\delta_d = 1/d$. Then for large n, d,

$$\begin{aligned}\mathbb{P}\big(\exists \boldsymbol{x} \in \mathsf{B}^d(1) :\ \lambda(\boldsymbol{x}) \ge & d(z_0(\boldsymbol{x}) + \varepsilon); \mathcal{G}\big) \le \mathbb{P}\big(\exists \boldsymbol{x} \in N^d(\delta_d) :\ \lambda(\boldsymbol{x}) \ge d(z_0(\boldsymbol{x}) + \varepsilon) - C_0' d\delta_d; \mathcal{G}\big)\\ &\le |N^d(\delta_d)| \max_{\boldsymbol{x}\in N^d(\delta_d)} \mathbb{P}\big(\lambda(\boldsymbol{x}) \ge d(z_0(\boldsymbol{x}) + \varepsilon/2);\ H(\boldsymbol{x}) \le C_0 d\big)\\ &\le \exp\Big(C_1 d\log d - \frac{d^2}{C(C_0, \varepsilon/2)}\Big)\,,\end{aligned}$$

which proves the desired claim. □

F.2 Proof of Theorem 4

Recall the definition of $z_0(\boldsymbol{x})$ in Lemma F.3 and define the event $\mathcal{E}(\varepsilon)$ by

$$\mathcal{E}(\varepsilon) := \Big\{ \forall \boldsymbol{x} \in \mathsf{B}^d(1) : \lambda_{\min}(\nabla^2 H(\boldsymbol{x})|_{\mathsf{T},\boldsymbol{x}}) \le d(z_0(\boldsymbol{x}) + \varepsilon); \\ \mathrm{Lip}(H; \mathsf{B}^d(1)) \le C_0 d;\ \mathrm{Lip}(\nabla^2 H; \mathsf{B}^d(1)) \le C_0 d \Big\}. \tag{219}$$

By Lemma A.6 and Lemma F.3 we can choose C sufficiently large, so that, for any $\varepsilon > 0$, $\mathbb{P}(\mathcal{E}(\varepsilon)) \ge 1 - C(\varepsilon)\exp(-d/C(\varepsilon))$ for large n, d.

Consider the sequence $\boldsymbol{x}^k$ produced by the Hessian descent algorithm. By Taylor's theorem, on the event $\mathrm{Lip}(\nabla^2 H; \mathsf{B}^d(1)) \le C_0 d$, we have

$$\begin{aligned} H(\boldsymbol{x}^{k+1}) &\le H(\boldsymbol{x}^k) - s_k\sqrt{\delta}\,\langle \boldsymbol{v}(\boldsymbol{x}^k), \nabla H(\boldsymbol{x}^k)\rangle + \frac{1}{2}\delta\,\langle \boldsymbol{v}(\boldsymbol{x}^k), \nabla^2 H(\boldsymbol{x}^k)\boldsymbol{v}(\boldsymbol{x}^k)\rangle + C_0 d\,\delta^{3/2} \\ &\le H(\boldsymbol{x}^k) + \frac{1}{2}\delta\lambda_{\min}(\nabla^2 H(\boldsymbol{x}^k)|_{\mathsf{T},\boldsymbol{x}^k}) + 2C_0 d\delta^{3/2}, \end{aligned} \tag{220}$$

where in the second line we used the definition of s_k and $\boldsymbol{v}(\boldsymbol{x}^k)$. We will hereafter work on the event $\mathcal{E}(\varepsilon)$, with $\varepsilon = \delta^{3/2}$. We thus have that

$$H(\boldsymbol{x}^{k+1}) \le H(\boldsymbol{x}^k) - \frac{1}{2}\delta d z_0(\boldsymbol{x}_k) + 3C_0 d\delta^{3/2}. \tag{221}$$

For $t = k\delta \in [0,1]$ with integer k define $U(t) := H(\boldsymbol{x}^{t/\delta})/n$, and on $(k\delta, (k+1)\delta)$ define $U(t)$ by linearly interpolating the two end values. Defining $[t]_\delta := \delta\lfloor t/\delta \rfloor$ and

$$\Phi(u,t) := -\frac{1}{2\alpha} z_*\big(\alpha; \sqrt{2\alpha\xi''(t)u}, \xi'(t)\big), \tag{222}$$

we can rewrite Eq. (220) as (excluding the points $t = k\delta$)

$$\frac{\mathrm{d}U}{\mathrm{d}t}(t) \le \Phi\big(U([t]_\delta), [t]_\delta\big) + C'\delta^{3/2}, \tag{223}$$

$$U(0) \le \frac{1}{2}\xi(0) + \delta. \tag{224}$$

where the bound in the last line holds wth high probability since $H(\mathbf{0})/n$ concentrates around its expectation. Since $u \mapsto \Phi(u,t)$ is a Lipschitz function of u (a fact that follows from Weyl's inequality), for all δ small enough, $U((k+1)\delta)$ is a monotone increasing function of $U(k\delta)$, and hence it follows that $U(t) \le U_*(t)$ where

$$\frac{\mathrm{d}U_*}{\mathrm{d}t}(t) = \Phi\big(U_*([t]_\delta), [t]_\delta\big) + C'\delta^{3/2}, \tag{225}$$

$$U_*(0) = \frac{1}{2}\xi(0) + \delta. \tag{226}$$

Finally, letting $u(t) = u(t;\alpha,\xi)$ denote the solution of the ODE (35), and using the fact that $t \mapsto \Phi(u,t)$ is a Lipschitz function of t, we obtain $|U_*(t) - u(t)| \le C''\sqrt{\delta}$ by standard discretization arguments for ODEs.

G Analysis of the two-phase algorithm

G.1 Proof of Theorem 5

The proof is based on the following state evolution characterization of the AMP iteration of Eqs. (48), (49).

Proposition G.1. *Consider the AMP iteration of Eqs. (48), (49) with initialization given by Eq. (50), and Onsager coefficients in Eqs. (51) to (53). Let $(M_\ell : \ell \ge 0)$, $(H_\ell : \ell \ge 0)$ be two independent centered Gaussian processes with covariances $\boldsymbol{Q}^M = (Q^M_{k,l})_{k,l\ge 0}$, $\boldsymbol{Q}^H = (Q^H_{k,l})_{k,l\ge 0}$ defined recursively via*

$$Q^H_{k+1,l+1} = \xi(Q^M_{k,l})\,, \qquad Q^M_{k+1,l+1} = \gamma^2\xi'(Q^M_{k,l})Q^H_{k,l}\,. \tag{227}$$

with initialization $Q^M_{0,\ell} = Q^H_{0,\ell} = 0$ for all $\ell \ge 0$.

Then, for any L, and any locally Lipschitz function $\psi : \mathbb{R}^{L+1} \to \mathbb{R}$, with $|\psi(\boldsymbol{x})| \le C(1 + \|\boldsymbol{x}\|_2^2)$, we have

$$\operatorname*{p-lim}_{n,d\to\infty} \frac{1}{n}\sum_{i=1}^{n} \psi(\sqrt{n}h_i^0, \dots, \sqrt{n}h_i^L) = \mathbb{E}\big\{\psi(H_0, \dots, H_L)\big\}\,, \tag{228}$$

$$\operatorname*{p-lim}_{n,d\to\infty} \frac{1}{d}\sum_{i=1}^{d} \psi(\sqrt{d}m_i^0, \dots, \sqrt{d}m_i^L) = \mathbb{E}\big\{\psi(M_0, \dots, M_L)\big\}\,. \tag{229}$$

We next use this result to prove the theorem.

Proof of Theorem 5. By construction, for any $k \ge 0$, we have that $Q^M_{\ell,\ell+k} = Q^H_{\ell,\ell+k} = q_\ell$, where the sequence $(q_\ell)_{\ell\ge 0}$ satisfies

$$q_{\ell+1} = \gamma^2\,\xi(q_\ell)\xi'(q_\ell)\,, \qquad q_0 = 0\,. \tag{230}$$

Fixed points of this map are solutions of the equation

$$\frac{1}{\gamma^2} = V(q)\,, \qquad V(q) := \frac{\xi(q)\xi'(q)}{q}\,. \tag{231}$$

Notice that $q \mapsto V(q)$ is a strictly convex function on $(0, q_{\max})$, where $q_{\max} \in [1, \infty]$ is the maximum radius of convergence of ξ (because $\xi(q)$ is real analytic with non-negative coefficients, and hence so is $(\xi(q)\xi'(q) - \xi(0)\xi'(0))/q$). Further, since $\xi(0), \xi'(0) > 0$, we have $V(q) \uparrow \infty$ as $q \downarrow 0$. Therefore $V(q)$ has a unique minimum at $q_{\rm RS}$. Further $\xi(q) \uparrow \infty$ as $q \uparrow \infty$ unless ξ is linear, in which case $V(q)$ is monotone decreasing. Therefore $q_{\rm RS} < \infty$ if and only if ξ is nonlinear. In the nonlinear case, this implies that the fixed point equation (231) has two solutions $q_1(\gamma) < q_{\rm RS} < q_2(\gamma)$ if $\gamma^2 < 1/V_* := 1/\min_{q\in(0,q_{\max})} V(q)$ and no solution for $\gamma^2 > 1/V_*$. In the linear case, we have only one solution $q_1(\gamma) < q_{\rm RS} = \infty$.

Since $q_1(\gamma)$ decreases continuously from $q_{\rm RS}$ to 0 as γ decreases from $1/\sqrt{V_*}$ to 0, for any $q \in (0, q_{\rm RS})$ we can select $\gamma = \gamma_0(q, \xi)$, where

$$\gamma_0(q, \xi) := -\sqrt{\frac{q}{\xi(q)\xi'(q)}} \tag{232}$$

and this yields that the smallest fixed point of (230) coincides with q. Further, $V'(q) < 0$ (again by strict convexity). Letting $f(r) := \gamma^2\,\xi(r)\xi'(r)$, this implies $f'(q) < 1$ strictly. Since f is convex

non-decreasing, we conclude that there exists a constant $C > 0$ (depending on ξ, q) such that, if $\gamma = \gamma_0(q, \xi)$, then

$$q - C\, e^{-\ell/C} \le q_\ell \le q \quad \forall \ell\,. \tag{233}$$

Notice that this implies that, for all $\ell, k \ge 0$ (the constant C will change from line to line):

$$\left|Q^M_{\ell,\ell+k} - q\right| \le Ce^{-\ell/C}\,, \qquad \left|Q^H_{\ell,\ell+k} - \xi(q)\right| \le Ce^{-\ell/C}\,. \tag{234}$$

By Proposition G.1, this implies that, for all $\ell, k \ge 0$

$$\operatorname*{p-lim}_{n,d\to\infty} \left|\|\boldsymbol{m}^\ell\|_2^2 - q\right| \le Ce^{-\ell/C} \qquad \operatorname*{p-lim}_{n,d\to\infty} \left\|\boldsymbol{m}^\ell - \boldsymbol{m}^{\ell+k}\right\|_2^2 \le Ce^{-\ell/C}\,, \tag{235}$$

$$\operatorname*{p-lim}_{n,d\to\infty} \left|\|\boldsymbol{h}^\ell\|_2^2 - \xi(q)\right| \le Ce^{-\ell/C} \qquad \operatorname*{p-lim}_{n,d\to\infty} \left\|\boldsymbol{h}^\ell - \boldsymbol{h}^{\ell+k}\right|_2^2 \le Ce^{-\ell/C}\,, \tag{236}$$

and therefore,

$$\operatorname*{p-lim}_{n,d\to\infty} \left|B_\ell - \frac{1}{\sqrt{\alpha}}\xi'(q)\right| \le Ce^{-\ell/C}\,, \tag{237}$$

$$\operatorname*{p-lim}_{n,d\to\infty} \left|C_\ell - \sqrt{\alpha}\xi'(q)\right| \le Ce^{-\ell/C}\,, \tag{238}$$

$$\operatorname*{p-lim}_{n,d\to\infty} \left|D_\ell - \xi(q)\xi'(q)\right| \le Ce^{-\ell/C}\,. \tag{239}$$

Therefore, using Eq. (48), we get

$$\begin{aligned}
\operatorname*{p-lim}_{n,d\to\infty} \frac{1}{n}\|\boldsymbol{F}(\boldsymbol{m}^\ell)\|^2 &= \operatorname*{p-lim}_{n,d\to\infty} \|\boldsymbol{h}^{\ell+1} + \gamma B_\ell \boldsymbol{h}^{\ell-1}\|^2 \\
&= \operatorname*{p-lim}_{n,d\to\infty} \left(1 + \gamma_0(q,\xi)B_\ell\right)^2 \cdot \|\boldsymbol{h}^\ell\|_2^2 + O(e^{-\ell/C}) \\
&= \left(1 + \gamma_0(q,\xi)\frac{1}{\sqrt{\alpha}}\xi'(q)\right)^2 \xi(q) + O(e^{-\ell/C}) \\
&= \left(\sqrt{\xi(q)} - \sqrt{\frac{1}{\alpha}q\xi'(q)}\right)^2 + O(e^{-\ell/C})\,.
\end{aligned}$$

The claim of the theorem now follows by fixing

$$\gamma = \gamma_*(q, \alpha, \xi) = \gamma_0(q \wedge q_0(\alpha), \xi)\,. \tag{240}$$

While not needed for the proof of the theorem we also note that our analysis implies that

$$\lim_{n,d\to\infty} \frac{1}{n}\left\|\mathsf{P}_{\mathsf{T},\boldsymbol{m}^L} \nabla H(\boldsymbol{m}^L)\right\|_2 \le C\, e^{-L/C}\,. \tag{241}$$

(Recall that the typical scale of $\|\nabla H(\boldsymbol{m})\|_2$ at a non-random point $\boldsymbol{m}$ is $\Theta(n)$.) To prove this claim, recall that $\nabla H(\boldsymbol{m}^\ell) = \boldsymbol{DF}(\boldsymbol{m}^\ell)^{\mathsf{T}}\boldsymbol{F}(\boldsymbol{m}^\ell)$ and therefore, using Eqs. (48) and (49)

$$\begin{aligned}
\frac{1}{n}\nabla H(\boldsymbol{m}^\ell) &= \frac{1}{\sqrt{n}}\boldsymbol{DF}(\boldsymbol{m}^\ell)^{\mathsf{T}}\left(\boldsymbol{h}^{\ell+1} + \gamma B_\ell \boldsymbol{h}^{\ell-1}\right) \\
&= \frac{1}{\sqrt{n}}(1 + \gamma B_\ell)\boldsymbol{DF}(\boldsymbol{m}^\ell)^{\mathsf{T}}\boldsymbol{h}^\ell + \boldsymbol{e}_1^\ell
\end{aligned}$$

$$= \frac{1}{\gamma\sqrt{\alpha}}(1+\gamma B_\ell)\big(\boldsymbol{m}^{\ell+1} + \gamma C_\ell \boldsymbol{m}^{\ell-1} + \gamma^2 D_\ell \boldsymbol{m}^{\ell-1}\big) + \boldsymbol{e}_1^\ell$$
$$= \frac{1}{\gamma\sqrt{\alpha}}(1+\gamma B_\ell)\big(1 + \gamma C_\ell + \gamma^2 D_\ell\big)\boldsymbol{m}^\ell + \boldsymbol{e}_1^\ell + \boldsymbol{e}_2^\ell\,,$$

where, by Eqs. (235), (236), we have

$$\operatorname*{p-lim}_{n,d\to\infty}\Big\{\|\boldsymbol{e}_1^\ell\|_2 + \|\boldsymbol{e}_2^\ell\|_2\Big\} \le C\,\exp(-\ell/C)\,.$$

This immediately implies Eq. (241). □

G.2 Proof of Theorem 6

Theorem Theorem 6 follows from Theorem 5 using the same exactly the same argument as for Theorem 4, with the following adaptations:

1. The initial value of the energy for the second phase is given by $H(\boldsymbol{m}^L)/n = u_{\rm RS}(q,\alpha,\xi)+o_P(1)$.
2. The Hessian descent algorithm takes palce in the $(d-1)$-dimensional space $V_L := \{\boldsymbol{x}\in\mathbb{R}^d : \langle \boldsymbol{x}, \boldsymbol{m}^L\rangle = 0\}$.
3. Within this subspace, we attempt to solve an optimization problem that has the same form as the original one, namely

$$\text{minimize} \quad \|\tilde{\boldsymbol{F}}(\boldsymbol{x};\boldsymbol{m}^L)\|_2^2\,, \qquad \tilde{\boldsymbol{F}}(\boldsymbol{x};\boldsymbol{m}^L) := \boldsymbol{F}(\boldsymbol{m}^L+\boldsymbol{x})\,, \tag{242}$$
$$\text{subj. to} \quad \boldsymbol{x}\in V_L,\ \ \|\boldsymbol{x}\|_2^2 = 1-\|\boldsymbol{m}^L\|_2^2 \tag{243}$$

   Of course, the reduction from dimension d to $d-1$ is immaterial, and the change in the radius constraint does not affect the argument given for Theorem 5.
4. For a non-random $\boldsymbol{v}\in\mathsf{B}^d(1)$, the random function $\tilde{\boldsymbol{F}}(\,\cdot\,;\boldsymbol{v})$ restricted to $\boldsymbol{v}^\perp$ is again a centered Gaussian process with covariance

$$\mathbb{E}\big\{\tilde F_i(\boldsymbol{x}_1)\tilde F_j(\boldsymbol{x}_1;\boldsymbol{v})\big\} = \delta_{ij}\,\xi\big(\|\boldsymbol{v}\|_2^2 + \langle \boldsymbol{x}_1,\boldsymbol{x}_2\rangle\big)\,. \tag{244}$$

   Hence, by Theorem 4, Hessian descent on the subspace $\boldsymbol{v}^\perp$ with respect to $\tilde{\boldsymbol{F}}$ outputs $\boldsymbol{x}^*\in\mathbb{S}^{d-1}$ with energy $\|\boldsymbol{F}(\boldsymbol{x}^*)\|^2/(2n) \le u(t_*;\boldsymbol{v})+C\delta+o_P(1)$, where $t_* := 1-\|\boldsymbol{v}\|_2^2$, where $u(\,\cdot\,;\boldsymbol{v})$ solves

$$\frac{\mathrm{d}u}{\mathrm{d}t}(t;\boldsymbol{v}) = -\frac{1}{2\alpha}z_*\big(\alpha;\sqrt{2\alpha u(t)\xi''(\|\boldsymbol{v}\|^2+t)},\xi'(\|\boldsymbol{v}\|^2+t)\big)\,, \qquad u(0) = \frac{1}{2n}\|F(\boldsymbol{v})\|^2\,. \tag{245}$$

5. The proof of Theorem 4 only uses a uniform upper bound over the minimum eigenvalue of Hessian of the energy function, and therefore the last claim actually applies uniformly over $\boldsymbol{v}\in\mathsf{B}^d(1)$. As a consequence it also applies to the random point $\boldsymbol{m}^L$.

G.3 Proof of Proposition G.1

This statement is a direct consequence of [EAMS21, Theorem 6]. We begin by restating the latter in a form that is adapted to our use here. For each $k\ge 1$, let $\boldsymbol{W}^{(k)}\in(\mathbb{R}^N)^{\otimes k}$ be an independent,

Gaussian symmetric tensor. More precisely, if $(\boldsymbol{G}^{(k)})_{k\geq 0}$ is a collection of tensors with i.i.d. $\mathsf{N}(0,1)$ entries, then

$$\boldsymbol{W}^{(k)} := \frac{1}{k!}\sqrt{\frac{k}{N^{k-1}}}\sum_{\pi\in\mathfrak{S}_k}\boldsymbol{G}^{(k)}_{\pi}\,, \tag{246}$$

where the sum is over permutations of k objects, and $\boldsymbol{G}^{(k)}_{\pi}$ is obtained by permuting the indices of $\boldsymbol{G}^{(k)}$. Let $\nu(t) = \sum_{k\geq 1} c_k^2 t^k$ for some coefficients c_k, $D \geq 1$ be an integer, $\Phi_\ell : \mathbb{R}^D \times \mathbb{R} \to \mathbb{R}^D$ be Lipschitz functions, and denote by $\boldsymbol{D}\Phi_\ell(z;v) \in \mathbb{R}^{D\times D}$ the Jacobian of Φ_ℓ with respect to its first argument (which exists in weak sense).

Define the sequence of random variables V, $(Z_\ell)_{\ell\geq 0}$ by letting $(Z_0, V) \sim p_0$ be an arbitrary random vector taking values in $\mathbb{R}^D \times \mathbb{R}$, and $(Z_\ell)_{\ell\geq 1}$ be a centered Gaussian process, again with Z_ℓ again taking values in $\mathbb{R}^D$ and covariance $\boldsymbol{Q}_{jk} := \mathbb{E}[Z_j Z_k^{\mathsf{T}}]$ determined recursively via

$$\boldsymbol{Q}_{j+1,k+1} = \nu\Big(\mathbb{E}\big\{\Phi_j(Z_j;V)\Phi_k(Z_k;V)^{\mathsf{T}}\big\}\Big)\,. \tag{247}$$

(Throughout this section we will not use boldface for D-dimensional vectors.)

For any dimension $N \in \mathbb{N}$, consider iterates $\boldsymbol{z}^\ell \in \mathbb{R}^{N\times D}$ given by

$$\boldsymbol{z}^{\ell+1} = \sum_{k=1}^{\infty} c_k \boldsymbol{W}^{(k)}\{\Phi_\ell(\boldsymbol{z}^\ell;\boldsymbol{v})\} - \Phi_{\ell-1}(\boldsymbol{z}^{\ell-1};\boldsymbol{v})\mathsf{B}_\ell\,, \tag{248}$$

$$\mathsf{B}_\ell := \nu'\Big(\mathbb{E}[\Phi_\ell(Z_\ell;V)\Phi_{\ell-1}(Z_{\ell-1};V)^{\mathsf{T}}]\Big)\odot\mathbb{E}[\boldsymbol{D}\Phi_\ell(Z_\ell;V)]\,, \tag{249}$$

where $\boldsymbol{v} \in \mathbb{R}^N$ and $\boldsymbol{z}^0 \in \mathbb{R}^{N\times D}$ are non-random, and ν' is applied entrywise, and $\odot$ denotes entrywise multiplication. Further $\Phi_\ell(\boldsymbol{z}^\ell;\boldsymbol{v}) \in \mathbb{R}^{N\times D}$ denotes the matrix whose i-th row is $\Phi_\ell(\boldsymbol{z}^\ell;\boldsymbol{v})_i = \Phi_\ell(\boldsymbol{z}_i^\ell; v_i)$. Further, for $\boldsymbol{x} \in \mathbb{R}^{N\times D}$, $\boldsymbol{x} = (x_1,\dots,x_N)^{\mathsf{T}}$, $x_i \in \mathbb{R}^D$, we denote by $\boldsymbol{W}^{(k)}\{\boldsymbol{x}\} \in \mathbb{R}^{N\times D}$ the matrix whose i-th row is

$$\boldsymbol{W}^{(k)}\{\boldsymbol{x}\}_i = \sum_{j_1,\dots,j_{k-1}=1}^{N} W^{(k)}_{ij_1\dots j_{k-1}}\boldsymbol{x}_{j_1}\odot\cdots\odot\boldsymbol{x}_{j_{k-1}}\,. \tag{250}$$

With these definitions, we have the following.

Proposition G.2 (Theorem 6 in [EAMS21]). *Assume that the empirical distribution of $\boldsymbol{v}, \boldsymbol{z}^0$, namely $\hat{p}_{\boldsymbol{v},\boldsymbol{z}^0} := N^{-1}\sum_{i\leq N}\delta_{v_i,z_i^0}$ converges in W_2 distance to p_0. Then, for any $L \geq 0$ and any function $\psi : \mathbb{R}^{D(L+1)+1} \to \mathbb{R}$ that is locally Lipschitz and at most quadratic growth (i.e. $|\psi(\boldsymbol{x})| \leq C(1+\|\boldsymbol{x}\|_2^2)$), we have*

$$\operatorname*{p-lim}_{N\to\infty}\frac{1}{N}\sum_{i=1}^{N}\psi(\boldsymbol{z}_i^0,\dots,\boldsymbol{z}_i^L;v_i) = \mathbb{E}\Big\{\psi(Z_0,\dots,Z_L;V)\Big\} \tag{251}$$

The same holds if B_ℓ is replaced by its empirical version

$$\hat{\mathsf{B}}_\ell := \nu'\Big(\frac{1}{N}\sum_{i=1}^{N}\Phi_\ell(z_i^\ell;v_i)\Phi_{\ell-1}(z_i^{\ell-1};v_i)^{\mathsf{T}}]\Big)\odot\frac{1}{N}\sum_{i=1}^{N}\boldsymbol{D}\Phi_\ell(z_i^\ell;v_i)\,. \tag{252}$$

Proof. As mentioned, this is an adaptation from [EAMS21, Theorem 6], with three points of difference. However, each of these points can be reduced to the version stated in [EAMS21].

1. In [EAMS21] it is assumed that $D = 1$, but memory across iterations is allowed. Namely, iterates $\boldsymbol{x}^\ell \in \mathbb{R}^N$ are updated via

$$\boldsymbol{x}^{\ell+1} = \sum_{k=1}^{\infty} c_k \boldsymbol{W}^{(k)} \{ f_\ell(\boldsymbol{x}^0, \dots, \boldsymbol{x}^\ell) \} - f_{\ell-1}(\boldsymbol{x}^0, \dots, \boldsymbol{x}^\ell) \tilde{\mathsf{B}}_\ell \,, \,. \tag{253}$$

   It is clear that this case also covers the seemingly more general one with iterates $\boldsymbol{z}^\ell \in \mathbb{R}^{N \times D}$ by grouping the $\boldsymbol{z}^\ell = (\boldsymbol{x}^{\ell D}, \boldsymbol{x}^{\ell D+1}, \cdots, \boldsymbol{x}^{\ell D+D-1})$. Indeed, it covers a more general form of the present statement in which Φ_ℓ depends on $\boldsymbol{z}^0, \dots, \boldsymbol{z}^\ell$ rather than just on $\boldsymbol{z}^\ell$.

2. The statement in [EAMS21] does not allow for dependency on the vector $\boldsymbol{v}$. However, the case with $\boldsymbol{v}$ can be reduced to the one without $\boldsymbol{v}$ simply by encoding $\boldsymbol{v}$ as an additional column of $\boldsymbol{z}^0$, and increasing D to $D + 1$. Since, as mentioned in the previous point, the result of [EAMS21] implies the case in which both D is arbitrary, and Φ_ℓ depends on $\boldsymbol{z}^0$, …, $\boldsymbol{z}^\ell$, dependence on $\boldsymbol{v}$ can be captured.

3. Finally, [EAMS21] uses the deterministic version of coefficient B_ℓ. One can show by induction over ℓ that: (i) $\hat{\mathsf{B}}_\ell = \mathsf{B}_\ell + o_P(1)$; (ii) Denoting by $\hat{\boldsymbol{z}}^\ell$ the iterates that result from using the empirical version, we have $\|bz^\ell - \hat{\boldsymbol{z}}^\ell\|_2 = o_P(1)$.

Normalizations are different but equivalent to the ones of [EAMS21]. □

We next set $N = n + d$ and

$$\boldsymbol{z}^\ell = \begin{pmatrix} \boldsymbol{x}^\ell \\ \boldsymbol{y}^\ell \end{pmatrix}, \quad \boldsymbol{x}^\ell \in \mathbb{R}^{n \times D}, \quad \boldsymbol{y}^\ell \in \mathbb{R}^{d \times D} \,. \tag{254}$$

Further, we set $v_i = 0$ for $i \le n$, $v_i = 1$ for $i > n$, and write $J_\ell(x) = \Phi_\ell(x; 0)$, $H_\ell(x) = \Phi_\ell(x; 1)$, and write $\overline{W}^{(k,q)}_{i;i_1,\dots,i_q,j_1,\dots j_{k-q}} = W^{(k)}_{i,i_1,\dots,i_q,n-1+j_1,\dots n-1+j_{k-q}}$, for $q \le k-1$. In what follows indices denoted by $i, i_1, i_2, \dots$ run over $[n]$, and indices denoted by $j, i_1, i_2, \dots$ run over $[d]$. We can then rewrite the iteration as

$$\begin{aligned} x_i^{\ell+1} = &\sum_{k=1}^{\infty} c_k \sum_{q=0}^{k-1} \binom{k-1}{q} \sum_{i_1 \dots i_q \le n} \sum_{j_{q+1} \dots j_{k-1} \le n} \overline{W}^{(k,q)}_{i;i_1,\dots,i_q,j_1,\dots j_{k-1-q}} \cdot \\ &\cdot J_\ell(x^\ell_{i_1}) \odot \cdots \odot J_\ell(x^\ell_{i_q}) \odot H_\ell(y^\ell_{j_1}) \odot \cdots \odot H_\ell(y^\ell_{j_{k-q}}) - \mathsf{B}_\ell^{\mathsf{T}} F_{\ell-1}(x_i^{\ell-1}) \,, \end{aligned} \tag{255}$$

$$\begin{aligned} y_i^{\ell+1} = &\sum_{k=1}^{\infty} c_k \sum_{q=0}^{k-1} \binom{k-1}{q} \sum_{i_1 \dots i_q \le n} \sum_{j_{q+1} \dots j_{k-1} \le n} \overline{W}^{(k,q)}_{i_1;i_2,\dots,i_q,j,j_1 \dots j_{k-1-q}} \cdot \\ &\cdot J_\ell(x^\ell_{i_1}) \odot \cdots \odot J_\ell(x^\ell_{i_q}) \odot H_\ell(y_{j_1}) \odot \cdots \odot H_\ell(y^\ell_{j_{k-q}}) - \mathsf{B}_\ell^{\mathsf{T}} H_{\ell-1}(y_j^{\ell-1}) \,. \end{aligned} \tag{256}$$

We next set $D = 3$ and

$$J_\ell(x = (x_1, x_2, x_3)) = \Big(0, \;\; \overline{f}_\ell(x_1), \;\; 0 \Big) \,, \tag{257}$$

$$H_\ell(y = (y_1, y_2, y_3)) = \Big(\overline{h}_\ell(y_2 - y_3), \;\; \overline{g}_\ell(y_2 - y_3), \;\; \overline{g}_\ell(y_2 - y_3) \Big) \,, \tag{258}$$

Writing $\hat{\boldsymbol{x}}^\ell$ for the first column of $\boldsymbol{x}^\ell$ and $\hat{\boldsymbol{y}}^\ell$, $\hat{\boldsymbol{y}}^\ell_0$ for the second and third columns of $\boldsymbol{y}^\ell$. In terms of these variables, the AMP iteration reads

$$\hat{x}_i^{\ell+1} = \sum_{k=1}^{\infty} c_k \sum_{j_1,\dots,j_{k-1} > n} \overline{W}^{(k,0)}_{i;i_1 \dots j_{k-1}} \overline{h}_\ell(\hat{y}^\ell_{j_1} - \hat{y}^\ell_{0,j_1}) \cdots \overline{h}_\ell(\hat{y}^\ell_{j_{k-1}} - \hat{y}^\ell_{0,j_{k-1}}) - b_{1,\ell} \overline{f}_{\ell-1}(\hat{x}_i^{\ell-1}) \,, \tag{259}$$

$$\hat{y}_j^{\ell+1} = \sum_{k=1}^{\infty} c_k \sum_{q=0}^{k-1} \binom{k-1}{q} \sum_{i_1 \dots i_q \le n} \sum_{j_{q+1} \dots j_{k-1} \le n} \overline{W}^{(k,q)}_{i_1;i_2,\dots,i_q,j,j_1 \dots j_{k-1-q}} \cdot \overline{f}_\ell(\hat{x}_{i_1}^\ell) \cdots \overline{f}_\ell(\hat{x}_{i_q}^\ell) \overline{g}_\ell(\hat{y}_{j_1}^\ell - \hat{y}_{0,j_1}^\ell) \cdots \overline{g}_\ell(\hat{y}_{j_{k-q}}^\ell - \hat{y}_{0,j_{k-q}}^\ell) - b_{2,\ell} \overline{g}_{\ell-1}(\hat{y}_j^{\ell-1} - \hat{y}_{0,j}^{\ell-1}) - b_{3,\ell} \overline{h}_{\ell-1}(\hat{y}_j^{\ell-1} - \hat{y}_{0,j}^{\ell-1}) \,, \tag{260}$$

$$\hat{y}_{0,j}^{\ell+1} = \sum_{k=1}^{\infty} c_k \sum_{j_1,\dots,j_{k-1} \le d} \overline{W}^{(k,0)}_{j_1;j,j_2,\dots,j_{k-1}} \overline{g}_\ell(\hat{y}_{j_1}^\ell - \hat{y}_{0,j_1}^\ell) \cdots \overline{g}_\ell(\hat{y}_{j_{k-1}}^\ell - \hat{y}_{0,j_{k-1}}^\ell) - b_{4,\ell} \overline{g}_{\ell-1}(\hat{y}_j^{\ell-1} - \hat{y}_{0,j}^{\ell-1}) - b_{5,\ell} \overline{h}_{\ell-1}(\hat{y}_j^{\ell-1} - \hat{y}_{0,j}^{\ell-1}) \,, \tag{261}$$

for suitable sequence of constants $b_{1,\ell}, \dots, b_{5,\ell}$. It is straightforward (albeit tedious) to write expressions for these constants, as well as the state evolution characterization that follows from Proposition G.2.

We next introduce a small parameter ε, and set $\overline{f}_\ell(x) = \varepsilon f_\ell(x/\varepsilon)$, $\overline{g}_\ell(x) = g_\ell(x/\varepsilon)$ $\overline{h}_\ell(x) = h_\ell(x/\varepsilon)$. By defining $x_i^\ell := (\hat{y}_j^\ell - \hat{y}_{0,j}^\ell)/\varepsilon$, and taking the difference of Eqs. (260) and (261), we get

$$\hat{x}_i^{\ell+1} = \sum_{k=1}^{\infty} c_k \sum_{j_1,\dots,j_{k-1} \le d} \overline{W}^{(k,0)}_{i;j_1 \dots j_{k-1}} h_\ell(x_{j_1}^\ell) \cdots h_\ell(x_{j_{k-1}}^\ell) - b_{*,\ell} f_{\ell-1}(\hat{x}_i^{\ell-1}) + O(\varepsilon) \,, \tag{262}$$

$$x_j^{\ell+1} = \sum_{k=1}^{\infty} c_k (k-1) \sum_{i_1 \le n} \sum_{j_1 \dots j_{k-2} \le d} \overline{W}^{(k,0)}_{i_1;j,j_1 \dots,j_{k-2}} f_\ell(\hat{x}_{i_1}^\ell) g_\ell(x_{j_1}^\ell) \cdots g_\ell(x_{j_{k-2}}^\ell) - a_{1,\ell} g_{\ell-1}(x_j^{\ell-1}) - a_{2,\ell} h_{\ell-1}(x_j^{\ell-1}) + O(\varepsilon) \,. \tag{263}$$

The order ε term can be controlled uniformly over n, d on a high probability event (controlling the operator norms of tensors $\boldsymbol{W}^{(k)}$), hence allowing to derive a state evolution characterization of the recursion in which $O(\varepsilon)$ terms are dropped.

We finally define

$$F_i(\boldsymbol{x}) := \sum_{k=1}^{\infty} c_k \sum_{j_1,\dots,j_{k-1} \le d} \overline{W}^{(k,0)}_{i;j_1 \dots j_{k-1}} x_{j_1}^\ell \cdots x_{j_{k-1}}^\ell \,, \tag{264}$$

and note that the above recursion (dropping $O(\varepsilon)$ terms) can be rewritten as

$$\hat{\boldsymbol{x}}^{\ell+1} = \boldsymbol{F}(h_\ell(\boldsymbol{x}^\ell)) - b_{*,\ell} f_{\ell-1}(\hat{\boldsymbol{x}}^{\ell-1}) \,, \tag{265}$$

$$\boldsymbol{x}^{\ell+1} = \boldsymbol{D}\boldsymbol{F}(\boldsymbol{x}^\ell)^\mathsf{T} f_\ell(\hat{\boldsymbol{x}}^\ell) - a_{1,\ell} g_{\ell-1}(\boldsymbol{x}^{\ell-1}) - a_{2,\ell} h_{\ell-1}(\boldsymbol{x}^{\ell-1}) \,. \tag{266}$$

This is exactly the form of the algorithm introduced in the main text, (48), (49), where functions f_ℓ, h_ℓ, g_ℓ are all linear.

The proof of Proposition G.1 follows from keeping track of the coefficients in the Onsager terms (the memory terms in the AMP recursions), as well as of the state evolution recursion, along the chain of reductions just defined.

References

[ABA13] Antonio Auffinger and Gérard. Ben Arous, *Complexity of random smooth functions on the high-dimensional sphere*, Ann. Probab. **41** (2013), no. 6, 4214–4247. MR 3161473

[AC17] Antonio Auffinger and Wei-Kuo Chen, *Parisi formula for the ground state energy in the mixed p-spin model*, The Annals of Probability **45** (2017), no. 6b, 4617–4631.

[ADH+19] Sanjeev Arora, Simon Du, Wei Hu, Zhiyuan Li, and Ruosong Wang, *Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks*, International Conference on Machine Learning, PMLR, 2019, pp. 322–332.

[AGZ09] Greg W. Anderson, Alice Guionnet, and Ofer Zeitouni, *An introduction to random matrices*, Cambridge University Press, 2009.

[Auf13] Auffinger, Antonio and Ben Arous, Gérard and Cernỳ, Jirí, *Random matrices and complexity of spin glasses*, Communications on Pure and Applied Mathematics **66** (2013), no. 2, 165–201.

[AW09] Jean-Marc Azaïs and Mario Wschebor, *Level sets and extrema of random processes and fields*, John Wiley & Sons, 2009. MR 2478201

[AZLL19] Zeyuan Allen-Zhu, Yuanzhi Li, and Yingyu Liang, *Learning and generalization in overparameterized neural networks, going beyond two layers*, Advances in Neural Information Processing Systems **32** (2019), 6158–6169.

[BHMM19] Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal, *Reconciling modern machine-learning practice and the classical bias–variance trade-off*, Proceedings of the National Academy of Sciences **116** (2019), no. 32, 15849–15854.

[BLLT20] Peter L Bartlett, Philip M Long, Gábor Lugosi, and Alexander Tsigler, *Benign overfitting in linear regression*, Proceedings of the National Academy of Sciences (2020).

[BMR21] Peter L. Bartlett, Andrea Montanari, and Alexander Rakhlin, *Deep learning: a statistical viewpoint*, Acta Numerica **30** (2021), 87–201.

[CDMFF11] Mireille Capitaine, Catherine Donati-Martin, Delphine Féral, and Maxime Février, *Free convolution with a semicircular distribution and eigenvalues of spiked deformations of wigner matrices*, Electron. J. Probab **16** (2011), no. 64, 1750–1792.

[COB19] Lenaic Chizat, Edouard Oyallon, and Francis Bach, *On lazy training in differentiable programming*, Advances in Neural Information Processing Systems, 2019, pp. 2937–2947.

[CS92] Andrea Crisanti and H-J Sommers, *The sphericalp-spin interaction spin glass model: the statics*, Zeitschrift für Physik B Condensed Matter **87** (1992), no. 3, 341–354.

[CS95] Andrea Crisanti and Hans-J. Sommers, *Thouless-Anderson-Palmer Approach to the Spherical p-Spin Spin Glass Model*, J. Phys. I France **5** (1995), no. 7, 805–813.

[DZPS18] Simon S Du, Xiyu Zhai, Barnabas Poczos, and Aarti Singh, *Gradient descent provably optimizes over-parameterized neural networks*, International Conference on Learning Representations, 2018.

[EAMS21] Ahmed El Alaoui, Andrea Montanari, and Mark Sellke, *Optimization of mean-field spin glasses*, The Annals of Probability **49** (2021), no. 6, 2922–2960.

[Gor85] Yehoram Gordon, *Some inequalities for gaussian processes and applications*, Israel Journal of Mathematics **50** (1985), no. 4, 265–289.

[JT17] Aukosh Jagannath and Ian Tobasco, *Low temperature asymptotics of spherical mean field spin glasses*, Communications in Mathematical Physics **352** (2017), no. 3, 979–1017.

[Kac43] Mark Kac, *On the average number of real roots of a random algebraic equation*, Bull. Amer. Math. Soc. **49** (1943), 314–320. MR 7812

[KU23] Persia Jana Kamali and Pierfrancesco Urbani, *Dynamical mean field theory for models of confluent tissues and beyond*, arXiv:2306.06420 (2023).

[MS17] James A Mingo and Roland Speicher, *Free probability and random matrices*, vol. 35, Springer, 2017.

[MS23a] Andrea Montanari and Eliran Subag, *Solving overparametrized systems of random equations II: On Smale's 17th problem over the reals*, 2023, In preparation.

[MS23b] ———, *Solving overparametrized systems of random equations III: Lipschitz hardness*, 2023, In preparation.

[OS20] Samet Oymak and Mahdi Soltanolkotabi, *Towards moderate overparameterization: global convergence guarantees for training shallow neural networks*, IEEE Journal on Selected Areas in Information Theory (2020).

[Ric45] Stephen O. Rice, *Mathematical analysis of random noise*, Bell System Tech. J. **24** (1945), 46–156. MR 11918

[Saa11] Yousef Saad, *Numerical methods for large eigenvalue problems: revised edition*, SIAM, 2011.

[Sch42] Isaac J. Schoenberg, *Positive definite functions on spheres*, Duke Math. J. **9** (1942), 96–108. MR 0005922 (3,232c)

[Sub17a] Eliran Subag, *The complexity of spherical p-spin models—a second moment approach*, Ann. Probab. **45** (2017), no. 5, 3385–3450. MR 3706746

[Sub17b] ———, *The geometry of the Gibbs measure of pure spherical spin glasses*, Invent. Math. **210** (2017), no. 1, 135–209. MR 3698341

[Sub21] ———, *Following the ground states of full-rsb spherical spin glasses*, Communications on Pure and Applied Mathematics **74** (2021), no. 5, 1021–1044.

[Sub23] ———, *Concentration for the zero set of random polynomial systems*, arXiv preprint arXiv:2303.11924 (2023).

[SZ21] Eliran Subag and Ofer Zeitouni, *Concentration of the complexity of spherical pure p-spin models at arbitrary energies*, J. Math. Phys. **62** (2021), no. 12, Paper No. 123301, 15. MR 4346481

[Urb23] Pierfrancesco Urbani, *A continuous constraint satisfaction problem for the rigidity transition in confluent tissues*, Journal of Physics A: Mathematical and Theoretical **56** (2023), no. 11, 115003.

[ZBH+21] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals, *Understanding deep learning (still) requires rethinking generalization*, Communications of the ACM **64** (2021), no. 3, 107–115.