Escaping mediocrity: how two-layer networks learn hard generalized linear models

Luca Arnaboldi LUCA.ARNABOLDI@EPFL.CH

Florent Krzakala FLORENT.KRZAKALA@EPFL.CH
Bruno Loureiro BRUNO.LOUREIRO@DI.ENS.FR
Ludovic Stephan LUDOVIC.STEPHAN@EPFL.CH

Abstract

This study explores the sample complexity for two-layer neural networks to learn a generalized linear target function under Stochastic Gradient Descent (SGD), focusing on the challenging regime where many flat directions are present at initialization. It is well-established that in this scenario $n = O(d \log d)$ samples are typically needed. However, we provide precise results regarding the pre-factors in high-dimensional contexts and for varying widths. Notably, our findings suggest that overparametrization can only enhance convergence by a constant factor within this problem class. These insights are grounded in the reduction of SGD dynamics to a stochastic process in lower dimensions, where escaping mediocrity equates to calculating an exit time. Yet, we demonstrate that a deterministic approximation of this process adequately represents the escape time, implying that the role of stochasticity may be minimal in this scenario.

1. Introduction

In this manuscript we are interested in the supervised task of learning the following target function:

$$y = \sigma_\star\left(w_\star^\top x\right) + \sqrt{\Delta} z, \tag{1}$$

where $x \sim \mathcal{N}(0, {}^1\!/_d I_d), z \sim \mathcal{N}(0,1)$. This target belongs to a general class of models known as *single-index* or *generalised linear* models, where the labels depend on the covariates $x \in \mathbb{R}^d$ only through its projection on a fixed direction $w_\star \in \mathbb{R}^d$, followed by a non-linear real-valued function $\sigma_\star : \mathbb{R} \to \mathbb{R}$. The popularity of this model is that different separation results can be shown in the high-dimensional limit where $d \gg 1$:

• In the well-specified setting where we fit an isotropic single-index model with the same hypothesis class (i.e. $f_\theta(x) = \sigma_\star(w^\top x)$), it has been shown that the sample complexity of one-pass SGD[1] is determined by the first non-zero Hermite coefficient of the target $\sigma_\star$, also known as the *information exponent* [7]. Problems with non-zero first Hermite have information exponent $k = 1$, and $w_\star$ can be learned at linear sample complexity $n = O(d)$. Instead, problems with zero first and non-zero second Hermite coefficient have information exponent $k = 2$, requiring instead $n = O(d \log d)$ samples [7, 42].
• For fully-connected two-layer neural networks $f_\theta(x) = a^\top \sigma(Wx)$, several results are known under different assumptions. For fixed first layer weights $W \in \mathbb{R}^{p \times d}$ (a.k.a. random features model) and large enough width p, learning the k-th order Hermite coefficient $\sigma_\star$ requires $n = O(d^k)$ samples [28], implying a sample complexity of $n = O(d^2)$ for a quadratic problem, e.g. $\sigma_\star(x) = x^2$. Recently,

1. Which we recall the reader is equivalent to the convergence rate.

it was shown that wide networks ($p \to \infty$) can achieve the well-specified sample complexity of $n = O(d)$ under one-pass SGD, *provided that all Hermite coefficients of both $\sigma_\star, \sigma$ are non-zero* [9]. In particular, this assumption cover only problems with information exponent $k = 1$, excluding hard cases such as quadratic problems. Finally, for $\sigma(x) = \sigma_\star(x) = x^2$, [38] has shown that for p large enough, full-batch gradient flow achieves sample complexity $n = 2d$, although at a running time of $t = O(\log d)$.

With the exception of [9], the works mentioned above cover the scaling of the sample complexity in the high-dimensional limit. Our goal is, instead, to derive sharp results for the sample complexity of learning (1) with a fully-connected two-layer neural network in the challenging case where $\sigma_\star$ has a vanishing first Hermite coefficient. As discussed above, this case violates the "standard learning scenario" of [9], and can be seen as a proxy for hard learning problems for descent-based algorithms. For concreteness, in the following we focus on the purely quadratic case:

$$y = \left(w_\star^\top x\right)^2 + \sqrt{\Delta} z, \qquad\qquad w_\star \in \mathbb{S}^{d-1}(\sqrt{d}) \tag{2}$$

Learning the target (2) consists of learning the non-linearity $\sigma_\star(x) = x^2$ and the direction $w_\star$. In this work, we focus our attention in the second part, considering the following architecture with squared-activation:

$$f_\Theta(x) = \frac{1}{p}\sum_{i=1}^{p} a_i (w_i^\top x)^2. \tag{3}$$

where $\Theta = (a, W)$ is the set of trainable weights, which are trained with one-pass stochastic gradient descent (SGD):

$$\Theta^{\nu+1} = \Theta^\nu - \gamma \nabla_\Theta \ell(y^\nu, f_{\Theta^\nu}(x^\nu)) \tag{4}$$

with square loss $\ell(y, x) = 1/2(y - x)^2$ and initial condition $\Theta^0 = (a^0, W^0)$. Note that at each step ν, the gradient is evaluated at a fresh pair of data $(x^\nu, y^\nu) \in \mathbb{R}^{d+1}$ drawn from the model (2). In particular, this implies that after $\nu \in [n]$ steps, the algorithm has seen n data points.

Learning in this problem is hard, and can be compared to finding a needle in a haystack. Indeed, with the exception of one direction that points towards $\pm w_\star$, the population risk at (random) initialization is mostly flat. This slows down the dynamics, which takes a long time to establish a significant correlation with the signal - a scenario we refer to as *escaping mediocrity*.

At first, the particular case of purely quadratic activation might appear too specific. Indeed, as we will see later the population risk for this task has a a global maximum at initialization and a degenerated set of global minima. The choice of more general $\sigma_\star$ and σ with zero first Hermite but not necessarily zero higher-order coefficients might give rise to other critical points such as saddle-points, giving rise to a more complex SGD dynamics. However, since the focus of this work is on escaping mediocrity, our conclusions will hold, up to constants, to more general activations with information exponent equal to 2 .

Summary of results — Our main contributions in this manuscript are:

• We derive a deterministic set of ODEs providing an exact and analytically tractable description of the one-pass SGD dynamics in the high-dimensional limit $d \to \infty$, and characterize the leading order stochastic corrections to this limit.

• We provide an analytical formula for the number of samples required for one-pass SGD to learn the phase retrieval target in high-dimensions at arbitrary network width. We show that overparametrization can only improve convergence by a constant factor for phase retrieval.
• Finally, we compute the leading order stochastic corrections to the exit time, and show that stochasticity does not help escaping the flat directions at initialization. This suggests that the deterministic descriptions is enough to fully capture the phenomenology of the dynamics in this problem.

All the codes used for numerical experiments are provided in this anonymous repository. Further related work is discussed in App. A.

We introduce our key theoretical tool, which consists in low-dimensional reduction of the projected SGD dynamics (4); in Appendix B we derive the high-dimensional limit $d \to \infty$ of interest.

Sufficient statistics — The key observation is to notice that the population risk only depends on the hidden-layer weights $W \in \mathbb{R}^{p\times d}$ through the the second layer weights $a \in \mathbb{R}^p$ and the weights correlation matrices $\Omega \in \mathbb{R}^{(p+1)\times(p+1)}$

$$\Omega := \begin{pmatrix} Q & m \\ m^\top & 1 \end{pmatrix} = \begin{pmatrix} {}^1\!/_d \mathrm{W}\mathrm{W}^\top & {}^1\!/_d \mathrm{W} w_\star \\ {}^1\!/_d \left(\mathrm{W} w_\star\right)^\top & 1 \end{pmatrix} \tag{5}$$

The explicit expression of the population risk is:

$$\mathcal{R}(\Theta) = \mathbb{E}\left[\ell(y, f_\Theta(x))\right] = \frac{\Delta+3}{2} - \frac{1}{p}\sum_{j=1}^{p} a_j \left(Q_{jj} + 2m_j^2\right) + \frac{1}{2p^2}\sum_{j,l=1}^{p} a_j a_l (Q_{jj}Q_{ll} + 2Q_{jl}^2) \tag{6}$$

Notice that the matrices M, Q are precisely the second moments of the pre-activations $(\lambda_\star, \lambda) = (w_\star^\top x, Wx) \in \mathbb{R}^{p+1}$. Therefore, to characterize the evolution of the risk throughout SGD, it is sufficient to track the evolution of the first layer weights a_i and the correlation matrices m, Q, which consists of $p(p+1)$ parameters.

2. Escaping mediocrity in the well-specified scenario

As a starting point, we consider the well-specified case of $p = 1$. In this section, we show that in the high-dimensional limit, the sample complexity constant for one-pass SGD can be well estimated from the deterministic reduction (12). In particular, we show that the stochastic corrections from a finer analysis of the process (12) can be neglected.

Exit time from deterministic limit— We now move to the description of the one-pass SGD dynamics. Our key goal in this section is to determine how much data / how long SGD takes in order to find the signal in the high-dimensional limit $d \to \infty$. As we discuss in App. B, in this limit the sufficient statistics concentrate, with its evolution being described by the following deterministic ODE:

$$\frac{\mathrm{d}\bar{m}(t)}{\mathrm{d}t} = \bar{m}(t)\left[4(1-6\gamma)(1-\bar{m}^2(t)) - 2\gamma\Delta\right] \quad \text{with} \quad \bar{m}(t) \in [-1, 1] \tag{7}$$

with initial condition $\bar{m}(0) = {}^1\!/_d\, w_\star^\top w^0$. See App. D for an explicit derivation. Figure 1 (left) compares the evolution of the risk predicted from solving the high-dimensional ODEs (7) with different finite size ($d = 3000$) simulated instances of the problem, showing a a good agreement between the theory and the averaged population risk over the different runs. Given the spherical constraint, the population risk is now simply given by $\mathcal{R}(m) = 2\left(1 - m^2\right) + {}^\Delta\!/_2$. From this expression, it is clear that $m = \pm 1$ are global minima and $m = 0$ is a global maximum. Therefore, the information theoretically minimum achievable risk is $\min \mathcal{R}(m) = \mathcal{R}(\pm 1) = {}^\Delta\!/_2$.

We start with two immediate observations that can be drawn from (7). First, we have a necessary upper bound on the learning rate for learning to occur: $\gamma < {}^1\!/_6$. Moreover, from fixed-point stability analysis we can get the value where $\bar{m}$ converges for large times, and, consequently, the asymptotic excess population risk achievable in this setting is:

$$\lim_{t\to\infty} \mathcal{R}(\bar{m}(t)) - {}^\Delta\!/_2 = \frac{\gamma\Delta}{1 - 6\gamma}. \tag{8}$$

We now move to our main problem: estimating the time SGD takes to escape mediocrity at initialization. Let $T \in [0, 1]$ be the relative difference with respect to the initial value of the risk, and let t_{ext} be the time when the risk exits the region above the threshold T, see Fig. 1 (right) for an illustration. By construction, t_{ext} can be found by solving the following equation:

$$(1 - T)\left(\mathcal{R}\big(\bar{m}(0)\big) - \frac{\Delta}{2}\right) = \left(\mathcal{R}\left(\bar{m}\left(t_{\text{ext}}\right)\right) - \frac{\Delta}{2}\right). \tag{9}$$

The above can be exactly solved by numerically integrating (7) and then finding the root of (9). However, an analytical expression for the ODE exit time can be found from the following two observations:

- From the discussion around equation (19), initializing at random in high-dimensions imply that $\bar{m}(0) = \varepsilon \ll 1$, so we can consider the linearization of equation (7) in ε and solve it analytically. For small enough T, this will lead us to an accurate result;
- Even if the ODE trajectories are deterministic, the exit time t_{ext} is a random variable of the random initialization.

Note there these lead to two natural notions of average exit time over the initial conditions. The first one is obtained by taking the expected value over initial conditions before solving the cross-threshold equation, while the second is to take the expected value exit time obtained from solving (7) over a fixed initial condition

$$t_{\text{ext}}^{(\text{anl})} = \frac{\log\left[Td + (1 - T)\right]}{8(1 - 6\gamma) - 4\gamma\Delta} \qquad t_{\text{ext}}^{(\text{qnc})} = \mathbb{E}_{\mu_0\sim\chi^2(1)}\left[\frac{\log\left[\frac{Td}{\mu_0} + (1 - T)\right]}{8(1 - 6\gamma) - 4\gamma\Delta)}\right]. \tag{10}$$

Borrowing the jargon from statistical physics we refer to $t_{\text{ext}}^{(\text{anl})}$ as the *annealed exit time*, while $t_{\text{ext}}^{(\text{qnc})}$ as the *quenched exit time*. Some comments on this result are in order:

- By concavity of the logarithm function, we have $t_{\text{ext}}^{(\text{qnc})} \geq t_{\text{ext}}^{(\text{anl})}$.
- For both notions, we have $t_{\text{ext}} = O(\log d)$ implying $n = O(d \log d)$ samples are required to escape mediocrity, consistent with the rates in the literature [7, 14, 42].
- Both exit times are monotonically increasing in both $\gamma \in [0, {}^1\!/_6]$ and $\Delta \geq 0$. Recalling that $\delta t = {}^\gamma\!/_d$, this implies the existence of an optimal learning rate $\gamma_{\text{opt}} = {}^1\!/_{(12\,+\,\Delta)}$ that minimizes the number of samples required to escape mediocrity.

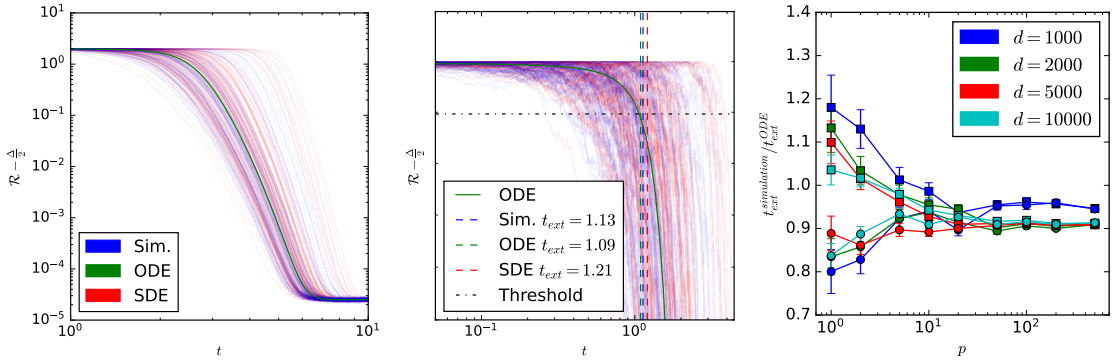

Figure 1: multiple run of the simulated SGD and the numerically integrated SDE, always starting from the same initial condition, with $d = 3000$. All the t_{ext} presented are obtained by solving numerically (9). The SDE captures the variance that the ODE doesn't exhibit, but the t_{ext} do not change considerably.

Does stochasticity matters? — Note that the initial correlation parameter at random initialization (18) is given by $m^0 = O(1/\sqrt{d})$. Therefore, in the high-dimensional limit $d \to \infty$ in which the ODE description (7) is exact, we have $\bar{m}(0) = 0$. This is a fixed point (7), which suggests that that strictly in the high-dimensional limit SGD is trapped forever at mediocrity. However, in practice we always have $d < \infty$, meaning that at initialization we always have a non-zero correlation with the signal $m^0 = \varepsilon \ll 1$. Moreover, at high but finite dimensions, (7) is just an approximation to the actual stochastic dynamics (12). Indeed, this is precisely what we used in order to estimate the exit time from the deterministic ODE (7). While the stochastic corrections to the high-dimensional limit does not radically change the convergence rate scaling [42] (and hence the mediocrity picture), it is important to ask whether it leads to important corrections on the precise exit time.

Stochastic corrections to the deterministic high-dimensional limit of one-pass SGD have been recently discussed in a broad setting by [8]. In particular, this work has shown that close to a fixed point the the process for the sufficient statistics (12) can be well approximated in the high-dimensional limit by a diffusion process with drift potential given by the corresponding deterministic ODEs. We follow a similar strategy, and consider the following process specialized to the case $p = 1$:

$$\mathrm{d}m_1 = \Psi_1(\Omega)\,\mathrm{d}t + \sqrt{\frac{\gamma}{d}}\,\sigma_m(\Omega)\cdot \mathrm{d}B_t, \qquad \mathrm{d}Q_{11} = \Phi_{11}(\Omega)\,\mathrm{d}t + \sqrt{\frac{\gamma}{d}}\,\sigma_Q(\Omega)\cdot \mathrm{d}B_t \tag{11}$$

where $\mathrm{d}B_t$ is a 2-dimensional Wiener process, and σ_M and σ_Q are defined as the standard deviation vector of the sufficient statistics; details in App. G . Notice that the stochastic correction is proportional to $\sqrt{\gamma/d}$, consistent to a first order correction to the deterministic limit. Similarly to the discussion in Sec. 2, the spherical constraint can be imposed by projecting the process in the sphere. This is discussed in detail in App. D. Figure 1 compares different instances of finite size simulations with instances of the spherical SDE with the same initial condition. Although the stochastic correction offers a better description of the process at large but finite dimensions, we find that quite surprisingly they have a small impact in the exit time. Hence, the formulas (10) derived in the Sec. 2 for random initialization provide a good approximation to the exit time. In App. F we discuss how to derive an exit time formulae with the stochastic corrections, both annealed and quenched one. As just showed, the new formulas do not offer any improvements compared to the deterministic ones, nevertheless the stochastic process can describe the dynamic even when the initialization is exactly $m = 0$, that is a fixed point of the ODE.

To summarize, in this section we have shown that the deterministic ODEs provides a good approximation for the precise number of samples required for escaping mediocrity in high-dimensions. In other words, stochasticity *does not help* in navigating the flat directions at initialization and in correlating with the signal.

3. The role of width

Thus far our discussion has focused on the well-specified case. We now discuss the role of width in escaping mediocrity. Our starting point are the deterministic ODEs (17) for the sufficient statistics derived in Sec. B. As in our previous analysis, we focus on the spherical setting where $w_i \in \mathbb{S}^{d-1}(\sqrt{d})$, implying $Q_{jj} = 1$, see App. D for a detailed derivation. First, we derive analytical expressions for the exit time for arbitrary width $p \geq 1$ in the particular case where the second layer is fixed at initialization $a_j^0 = 1, \forall j \in [n]$. The role played by the second layer is then discussed in App. I . Differently from the $p = 1$ case, the process cannot be described be a single sufficient statistics, and instead we have to track $p(p-1)/2$ non-diagonal entries of Q (it is a symmetric matrix), and p components of the vector m. Note that equation (9) remains valid to define t_{ext}, and can be solved numerically. An analytical expression for the exit time can be derived under similar assumptions to the ones discussed in Section 2, although the derivation is significantly more cumbersome. Full details and the explicit expression of t_{ext} can be found in App. E.

Notice that t_{ext} is a monotonically decreasing function of the width. Nevertheless, for any $p \geq 1$, the leading order dependence in the dimension is $t_{\text{ext}} = \log d$. Hence, despite helping escaping mediocrity, increasing the width cannot mitigate it. This can be contrasted to other aspects in which overparametrization can significantly help optimization, for instance with global convergence [5]. Interestingly, the minimal escaping time $t_{\text{ext}}^{(\text{anl})} = 1/4 \log\big(T(p+1)d + (p+1)(1-T)/2p\big)$, obtained by choosing the learning rate that minimizes the sample complexity for escaping, has the same pre-factor for any width $p \geq 1$, with the only differences being the dependence in p inside the logarithm and in the time scaling $t = \nu\gamma/pd$. At infinite width $p \to \infty$, this simply amounts to a factor $\frac{12+\Delta}{2+\Delta}$ with respect to $p = 1$. Details of this computation can be found in App. E.4.

Figure 1(right) compares our analytical formulas (47) & (46) with real one-pass SGD simulations. The simulation are averaged over many different instance of the initial conditions, and the ratio γ/p is kept constant when varying p, for not having discrepancies due to the different learning rate scaling. It's interesting to notice how the two different formulas gives the same outcome for large width $p \gg 1$. Moreover, for narrow networks they essentially differ from by a d independent constant. Figure 1(right) also suggests that, as for $p = 1$, the stochasticity can be neglected in the estimation of the exit time. In App. G we provide further evidence of that.

References

[1] Emmanuel Abbe, Enric Boix Adsera, and Theodor Misiakiewicz. The merged-staircase property: a necessary and nearly sufficient condition for sgd learning of sparse functions on two-layer neural networks. In *Conference on Learning Theory*, pages 4782–4887. PMLR, 2022.

[2] Emmanuel Abbe, Enric Boix-Adsera, and Theodor Misiakiewicz. Sgd learning on neural networks: leap complexity and saddle-to-saddle dynamics, 2023.

[3] Luca Arnaboldi, Ludovic Stephan, Florent Krzakala, and Bruno Loureiro. From high-dimensional & mean-field dynamics to dimensionless odes: A unifying approach to sgd in two-layers networks, 2023.

[4] Benjamin Aubin, Bruno Loureiro, Antoine Baker, Florent Krzakala, and Lenka Zdeborová. Exact asymptotics for phase retrieval and compressed sensing with random generative priors. In Jianfeng Lu and Rachel Ward, editors, *Proceedings of The First Mathematical and Scientific Machine Learning Conference*, volume 107 of *Proceedings of Machine Learning Research*, pages 55–73. PMLR, 20–24 Jul 2020.

[5] Francis Bach and Lénaic Chizat. Gradient descent on infinitely wide neural networks: Global convergence and generalization. In *International Congress of Mathematicians*, 2022.

[6] Jean Barbier, Florent Krzakala, Nicolas Macris, Léo Miolane, and Lenka Zdeborová. Optimal errors and phase transitions in high-dimensional generalized linear models. *Proceedings of the National Academy of Sciences*, 116(12):5451–5460, 2019. doi: 10.1073/pnas.1802705116.

[7] Gerard Ben Arous, Reza Gheissari, and Aukosh Jagannath. Online stochastic gradient descent on non-convex losses from high-dimensional inference. *Journal of Machine Learning Research*, 22(106):1–51, 2021. URL http://jmlr.org/papers/v22/20-1288.html.

[8] Gerard Ben Arous, Reza Gheissari, and Aukosh Jagannath. High-dimensional limit theorems for sgd: Effective dynamics and critical scaling. *Advances in Neural Information Processing Systems*, 35:25349–25362, 2022.

[9] Raphaël Berthier, Andrea Montanari, and Kangjie Zhou. Learning time-scales in two-layers neural networks, 2023.

[10] Etienne Boursier, Loucas Pillaud-Vivien, and Nicolas Flammarion. Gradient flow dynamics of shallow relu networks for square loss and orthogonal inputs. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 20105–20118. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/7eeb9af3eb1f48e29c05e8dd3342b286-Paper-Conference.pdf.

[11] Emmanuel J. Candès, Yonina C. Eldar, Thomas Strohmer, and Vladislav Voroninski. Phase retrieval via matrix completion. *SIAM Journal on Imaging Sciences*, 6(1):199–225, 2013. doi: 10.1137/110848074.

[12] Emmanuel J. Candès, Thomas Strohmer, and Vladislav Voroninski. Phaselift: Exact and stable signal recovery from magnitude measurements via convex programming. *Communications on Pure and Applied Mathematics*, 66(8):1241–1274, 2013. doi: https://doi.org/10.1002/cpa.21432.

[13] Emmanuel J. Candès, Xiaodong Li, and Mahdi Soltanolkotabi. Phase retrieval via wirtinger flow: Theory and algorithms. *IEEE Transactions on Information Theory*, 61(4):1985–2007, 2015. doi: 10.1109/TIT.2015.2399924.

[14] Yuxin Chen, Yuejie Chi, Jianqing Fan, and Cong Ma. Gradient descent with random initialization: fast global convergence for nonconvex phase retrieval. *Mathematical Programming*, 176 (1):5–37, Jul 2019. ISSN 1436-4646. doi: 10.1007/s10107-019-01363-6.

[15] Lenaic Chizat and Francis Bach. On the global convergence of gradient descent for over-parameterized models using optimal transport. *Advances in neural information processing systems*, 31, 2018.

[16] M Copelli and N Caticha. On-line learning in the committee machine. *Journal of Physics A: Mathematical and General*, 28(6):1615–1625, mar 1995. doi: 10.1088/0305-4470/28/6/016.

[17] Damek Davis, Dmitriy Drusvyatskiy, and Courtney Paquette. The nonsmooth landscape of phase retrieval. *IMA Journal of Numerical Analysis*, 40(4):2652–2695, 01 2020. ISSN 0272-4979. doi: 10.1093/imanum/drz031.

[18] Jonathan Dong, Lorenzo Valzania, Antoine Maillard, Thanh-an Pham, Sylvain Gigan, and Michael Unser. Phase retrieval: From computational imaging to machine learning: A tutorial. *IEEE Signal Processing Magazine*, 40(1):45–57, 2023. doi: 10.1109/MSP.2022.3219240.

[19] Behrooz Ghorbani, Song Mei, Theodor Misiakiewicz, and Andrea Montanari. Limitations of lazy training of two-layers neural network. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[20] Sebastian Goldt, Madhu Advani, Andrew M Saxe, Florent Krzakala, and Lenka Zdeborová. Dynamics of stochastic gradient descent for two-layer neural networks in the teacher-student setup. *Advances in neural information processing systems*, 32, 2019.

[21] Sebastian Goldt, Marc Mézard, Florent Krzakala, and Lenka Zdeborová. Modeling the influence of data structure on learning in neural networks: The hidden manifold model. *Phys. Rev. X*, 10: 041044, Dec 2020. doi: 10.1103/PhysRevX.10.041044.

[22] Sebastian Goldt, Bruno Loureiro, Galen Reeves, Florent Krzakala, Marc Mézard, and Lenka Zdeborová. The gaussian equivalence of generative models for learning with two-layer neural networks. In *Proceedings of Machine Learning Research*, volume 145, pages 1–46. 2nd Annual Conference on Mathematical and Scientific Machine Learning, 2021.

[23] Karl Hajjar and Lénaïc Chizat. On the symmetries in the dynamics of wide two-layer neural networks. *Electronic Research Archive*, 31(4):2175–2212, 2023. ISSN 2688-1594. doi: 10.3934/era.2023112. URL https://www.aimspress.com/article/doi/10.3934/era.2023112.

[24] Arthur Jacot, François Ged, Berfin Şimşek, Clément Hongler, and Franck Gabriel. Saddle-to-saddle dynamics in deep linear networks: Small initialization training, symmetry, and sparsity, 2022.

[25] Kishore Jaganathan, Yonina C. Eldar, and Babak Hassibi. Phase Retrieval: An Overview of Recent Developments. CRC Press, 2016. ISBN 9781315371474.

[26] Antoine Maillard, Bruno Loureiro, Florent Krzakala, and Lenka Zdeborová. Phase retrieval in high dimensions: Statistical and computational phase transitions. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 11071–11082. Curran Associates, Inc., 2020.

[27] Song Mei, Andrea Montanari, and Phan-Minh Nguyen. A mean field view of the landscape of two-layer neural networks. *Proceedings of the National Academy of Sciences*, 115(33): E7665–E7671, 2018.

[28] Song Mei, Theodor Misiakiewicz, and Andrea Montanari. Generalization error of random feature and kernel methods: Hypercontractivity and kernel matrix concentration. *Applied and Computational Harmonic Analysis*, 59:3–84, 2022. ISSN 1063-5203. doi: https://doi.org/10.1016/j.acha.2021.12.003. Special Issue on Harmonic Analysis and Machine Learning.

[29] Francesca Mignacco, Pierfrancesco Urbani, and Lenka Zdeborová. Stochasticity helps to navigate rough landscapes: comparing gradient-descent-based algorithms in the phase retrieval problem. *Machine Learning: Science and Technology*, 2(3):035029, jul 2021. doi: 10.1088/2632-2153/ac0615.

[30] Marco Mondelli and Andrea Montanari. Fundamental limits of weak recovery with applications to phase retrieval. In Sébastien Bubeck, Vianney Perchet, and Philippe Rigollet, editors, *Proceedings of the 31st Conference On Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pages 1445–1450. PMLR, 06–09 Jul 2018.

[31] G. Reents and R. Urbanczik. Self-averaging and on-line learning. *Phys. Rev. Lett.*, 80:5445–5448, Jun 1998. doi: 10.1103/PhysRevLett.80.5445.

[32] P Riegler and M Biehl. On-line backpropagation in two-layered neural networks. *Journal of Physics A: Mathematical and General*, 28(20):L507–L513, oct 1995. doi: 10.1088/0305-4470/28/20/002.

[33] Grant Rotskoff and Eric Vanden-Eijnden. Trainability and accuracy of artificial neural networks: An interacting particle system approach. *Communications on Pure and Applied Mathematics*, 75(9):1889–1935, 2022. doi: https://doi.org/10.1002/cpa.22074. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/cpa.22074.

[34] David Saad and Sara Solla. Dynamics of on-line gradient descent learning for multilayer neural networks. In D. Touretzky, M. C. Mozer, and M. Hasselmo, editors, *Advances in Neural Information Processing Systems*, volume 8. MIT Press, 1996.

[35] David Saad and Sara A. Solla. On-line learning in soft committee machines. *Phys. Rev. E*, 52: 4225–4243, Oct 1995. doi: 10.1103/PhysRevE.52.4225.

[36] David Saad and Sara A. Solla. Exact solution for on-line learning in multilayer neural networks. *Phys. Rev. Lett.*, 74:4337–4340, May 1995. doi: 10.1103/PhysRevLett.74.4337.

[37] Stefano Sarao Mannelli, Giulio Biroli, Chiara Cammarota, Florent Krzakala, Pierfrancesco Urbani, and Lenka Zdeborová. Complex dynamics in simple neural networks: Understanding gradient flow in phase retrieval. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan,

and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 3265–3274. Curran Associates, Inc., 2020.

[38] Stefano Sarao Mannelli, Eric Vanden-Eijnden, and Lenka Zdeborová. Optimization and generalization of shallow neural networks with quadratic activation functions. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 13445–13455. Curran Associates, Inc., 2020.

[39] Justin Sirignano and Konstantinos Spiliopoulos. Mean field analysis of neural networks: A central limit theorem. *Stochastic Processes and their Applications*, 130(3):1820–1852, 2020.

[40] Ju Sun, Qing Qu, and John Wright. A geometric analysis of phase retrieval. In *2016 IEEE International Symposium on Information Theory (ISIT)*, pages 2379–2383, 2016. doi: 10.1109/ISIT.2016.7541725.

[41] Yan Shuo Tan and Roman Vershynin. Phase retrieval via randomized Kaczmarz: theoretical guarantees. *Information and Inference: A Journal of the IMA*, 8(1):97–123, 04 2018. ISSN 2049-8772. doi: 10.1093/imaiai/iay005.

[42] Yan Shuo Tan and Roman Vershynin. Online stochastic gradient descent with arbitrary initialization solves non-smooth, non-convex phase retrieval. *Journal of Machine Learning Research*, 24(58):1–47, 2023. URL http://jmlr.org/papers/v24/20-902.html.

[43] Rodrigo Veiga, Ludovic Stephan, Bruno Loureiro, Florent Krzakala, and Lenka Zdeborova. Phase diagram of stochastic gradient descent in high-dimensional two-layer neural networks. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022.

[44] Caibin Zeng. Mean exit time and escape probability for the ornstein–uhlenbeck process. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 30(9):093127, 2020.

Appendix A. Further related work

The investigation of a deterministic high-dimensional limit of one-pass SGD for two-layer neural networks draws back to the seminal works of [34–36], and was followed by a stream of works that span decades of research [3, 16, 20–22, 31, 32, 43]. More recently, the stochastic corrections around fixed points of the dynamics have been investigated by [8]. In a complementary research line, [15, 19, 27, 33, 39]) have shown that an alternative deterministic description of SGD can be obtained in the infinite width-limit, a.k.a. mean-field regime. High-dimensional reductions of the mean-field equations have been studied by [1, 2, 9, 23]. Recently, [3, 43] has shown that these apparently different limits of one-pass SGD can be unified in a single description.

There has been a recent surge of interest in studying how increasing degrees of complexity in the target function are incrementally learned by SGD [1, 2, 9, 10, 24], with an emerging staircase picture where complexity is sequentially learned in different scenarios. This picture, however, is bound to classes of targets where SGD develops strong correlations with the target directions at initialization, a notion which was mathematically formalized by the so-called information exponent (IE) by [7]. Instead, targets for which the landscape at initialization is mostly flat (IE ≥ 2) are hard for SGD at high-dimensions, translating to very slow dynamics. This is precisely the case for the phase retrieval problem (IE $= 2$), a classic inverse problem arising in many scientific areas, from X-ray crystallography to astronomical imaging [18, 25]. Phase retrieval has been widely studied in the literature as a prototypical example of a hard inverse problem [4, 6, 11–13, 26, 30], providing a simple yet challenging example of a non-convex optimization problem which is hard for descent-based algorithms [14, 17, 29, 37, 38, 40–42].

Appendix B. High-dimensional limit of SGD

As shown in Appendix C, starting from Eq. (4) we can derive a set of self-consistent stochastic processes describing the evolution of (a, m, Q):

$$a_j^{\nu+1} - a_j^{\nu} = \frac{\gamma}{pd}\mathcal{E}^{\nu}\lambda_j^2 \tag{12}$$

$$m_j^{\nu+1} - m_j^{\nu} =: \mathcal{M}_j(a, \lambda_\star, \lambda) = 2\frac{\gamma}{pd}\mathcal{E}^{\nu} a_j \lambda_j \lambda_\star \tag{13}$$

$$Q_{jl}^{\nu+1} - Q_{jl}^{\nu} =: \mathcal{Q}_{jl}(a, \lambda_\star, \lambda) = 2\frac{\gamma}{pd}\mathcal{E}^{\nu}\left(a_j + a_l\right)\lambda_j\lambda_l \tag{14}$$

$$+ 4\frac{\gamma^2}{p^2 d}\mathcal{E}^{\nu 2}||x^{\nu}||^2 a_j a_l \lambda_j \lambda_l \tag{15}$$

where we have defined the displacement vector

$$\mathcal{E}^{\nu} := \frac{1}{p}\sum_{j=1}^{p} a_j(\lambda_j^{\nu})^2 - (\lambda^{\star\nu})^2 + \sqrt{\Delta}z^{\nu}, \tag{16}$$

and we used γ/d as the learning rate of the second layer, in order to have the same high-dimensional scaling.

High-dimensional limit — As of now we have not made any assumptions on the dimension of the problem; the stochastic processes defined in (12) are exact, with the right-hand side depending

implicitly on (m, Q) through the moments of $(\lambda_\star, \lambda)$. However, our goal is to study this process in the high-dimensional limit $d \to \infty$ where learning is hard and simulating (4) can be computationally demanding. Defining a step-size $\delta t = \gamma/pd$ and a continuous extension of (a^ν, m^ν, Q^ν) to continuous time $(a(\nu\delta t), m(\nu\delta t), Q(\nu\delta t))$ by linear interpolation, it can be shown that in the high-dimensional limit $d \to \infty$ the sufficient statistics $(a(t), m(t), Q(t))$ concentrate in their expectation $(\bar{a}(t), \bar{m}(t), \bar{Q}(t))$, which satisfies the following system of ordinary differential equations (ODEs):

$$\begin{aligned}
\frac{\mathrm{d}\bar{a}_j}{\mathrm{d}t} &= \mathbb{E}_{(\lambda,\lambda_\star)\sim\mathcal{N}(0_{p+1},\Omega)}\left[\mathcal{E}\lambda_j^2\right] \\
\frac{\mathrm{d}\bar{m}_j}{\mathrm{d}t} &= \mathbb{E}_{(\lambda,\lambda_\star)\sim\mathcal{N}(0_{p+1},\Omega)}\left[\mathcal{M}_j(a,\lambda_\star,\lambda)\right] =: \Psi_j\left(\Omega\right) \\
\frac{\mathrm{d}\bar{Q}_{jl}}{\mathrm{d}t} &= \mathbb{E}_{(\lambda,\lambda_\star)\sim\mathcal{N}(0_{p+1},\Omega)}\left[\mathcal{Q}_{jl}(a,\lambda_\star,\lambda)\right]
\end{aligned} \tag{17}$$

with initial conditions given by $(\bar{a}(0), \bar{m}(0), \bar{Q}(0)) = (a^0, {}^1\!/\!_d W^0 w_\star, {}^1\!/\!_d W^0 {W^0}^\top)$. The explicit expression of these expected values can be found in Appendix C. As discussed in the related works, the high-dimensional limit of one-pass SGD for two-layer neural networks have been studied under different settings in the literature [3, 8, 9, 20, 31, 34, 41, 43]. However, to our best knowledge our work is the first to derive and study these equations for the squared activation in the high-dimensional limit.

Initialization and mediocrity — In the noiseless case $\Delta = 0$, it is easy to check that $a_j = 1$ and $w_j = \pm w_\star$ ($m_j = \pm 1$ and $Q_{jl} = 1$) is indeed a stationary point of (17) that corresponds to two degenerated global minima of the population risk (6). Adding a noise $\Delta > 0$ only shift these values. Similarly, it is easy to check that $m_i = 0$ and $Q_{ij} = 0$ for $i \neq j$ are also stationary points. These correspond to taking $w_j \perp w_l \perp w_\star$ for all $j \neq l$ in (17), and is a global maximum of (6). This stationary point plays an important role in the dynamics. Indeed, in the absence of knowledge on the process that generated the data (2), it is customary to initialize the weights randomly:

$$w_j^0 \sim \mathcal{N}(0, I_d), \qquad\qquad j = 1, \cdots, p. \tag{18}$$

When $d \to \infty$, the weights are be orthogonal with high probability. In terms of the sufficient statistics:

$$Q_{jj} \sim \text{Dirac}(1), \qquad j \neq l: \ \sqrt{d}\, Q_{jl}^0 \xrightarrow{d\to+\infty} \mathcal{N}(0,1) \quad \text{and} \quad \sqrt{d}\, m_j^0 \xrightarrow{d\to+\infty} \mathcal{N}(0,1). \tag{19}$$

Therefore, since the variance of (m^0, Q^0) decays as ${}^1\!/\!_d$, the higher the dimension, the closer a random initialization is to a stationary point of the dynamics. Moreover, of all the d directions, there exists $d - p - 1$ directions orthogonal to $w_\star$ and $\{w_j^0\}_{j\in[p]}$ along which the population risk (6) remains constant. The proliferation of flat directions close to initialization severely slows down the SGD dynamics at high-dimensions, which typically requires $n = O(d \log d)$ steps to develop a significant correlation with the signal in order to escape this region. This scenario, which we refer to as *escaping mediocrity*, is common to many hard learning problems [7]. In the following, we leverage the exact description (17) derived in this section to estimate precisely how much data is required for SGD to escape mediocrity in the prototypical phase retrieval problem (1).

Spherical constraint — A phenomenon that is observed when starting from the initial conditions above is a change in the norms of the weights w_i without effectively correlating with $w_\star$. In this phase, sometimes referred as *norm learning*, $m \approx Q_{jl} \approx 0$ for $j \neq l$, while Q_{jj} changes considerably, resulting in a slightly decrease in the population risk towards a plateau that reflects mediocrity. Since the focus of this study is precisely on escaping mediocrity (i.e. developing non-zero correlation with the signal), in the following we will fix the norm of the weights $w_i^\nu \in \mathbb{S}^{d-1}(\sqrt{d})$ at initialization and throughout the dynamics $\nu \in [n]$. This assumption, which was also the focus of [7], amounts to imposing a spherical constraint at every step of SGD, also known as *projected SGD*:

$$w_j^{\nu+1} = \frac{w_j^\nu - \gamma \nabla_{w_j} \ell(y^\nu, f_{\Theta^\nu}(x^\nu))}{\left\| w_j^\nu - \gamma \nabla_{w_j} \ell(y^\nu, f_{\Theta^\nu}(x^\nu)) \right\|} \sqrt{d}. \tag{20}$$

The high-dimensional limit of these equations lead to the following ODEs for the evolution of the sufficient statistics (M, Q):

$$\frac{\mathrm{d}\bar{m}_j}{\mathrm{d}t} = \Psi_j(\Omega) - \frac{\bar{m}_j}{2}\Phi_{jj}(\Omega), \qquad \frac{\mathrm{d}\bar{Q}_{jl}}{\mathrm{d}t} = \Phi_{jl}(\Omega) - \frac{\bar{Q}_{jl}}{2}\left(\Phi_{jj}(\Omega) + \Phi_{ll}(\Omega)\right). \tag{21}$$

Note that $Q_{jj} = 1$ is consistently fixed.

Appendix C. Explicit ODEs derivation

C.1. Derivation of the process

Let's start by reminding the definition of *displacement* at step ν

$$\mathcal{E}^\nu := \frac{1}{p}\sum_{j=1}^{p} a_j (\lambda_j^\nu)^2 - (\lambda_\star^\nu)^2 + \sqrt{\Delta} z^\nu, \tag{22}$$

from which it's easy to write the loss function

$$\ell(y^\nu, f_{\Theta^\nu}(x^\nu)) = \frac{1}{2}\left(\mathcal{E}^\nu\right)^2. \tag{23}$$

The gradient respect to the parameters is given

$$\begin{aligned} \partial_{a_j} \ell(y^\nu, f_{\Theta^\nu}(x^\nu)) &= \frac{1}{p}\mathcal{E}^\nu (\lambda_j^\nu)^2 \\ \nabla_{w_j} \ell(y^\nu, f_{\Theta^\nu}(x^\nu)) &= \frac{1}{p}\mathcal{E}^\nu 2 a_j^\nu \lambda_j^\nu x^\nu \end{aligned} \tag{24}$$

Using γ as learning rate for the weights w_j and γ/p for the second layer, we have the following update equations

$$\begin{aligned} a_j^{\nu+1} &= a_j^\nu - \frac{\gamma}{pd}\mathcal{E}^\nu (\lambda_j^\nu)^2 \\ w_j^{\nu+1} &= w_j^\nu - \frac{\gamma}{p}\mathcal{E}^\nu 2 a_j^\nu \lambda_j^\nu x^\nu \end{aligned} \tag{25}$$

Applying the definition of the sufficient statistics, $m_j = {}^{w_j w_\star}/_d$ and $Q_{jl} = {}^{w_j w_l}/_d$, we can recover Eqs. (12).

C.2. Explicit ODE

To get the explicit form of our ODEs we need to compute some expected value over the preactivations $(\lambda_\star, \lambda)$. These are gaussian variables, whose correlation matrix is given by Ω. Similarly, the population risk is defined as

$$\mathcal{R} = \mathbb{E}_{(\lambda,\lambda^\star)\sim\mathcal{N}(0_{p+1},\Omega)}\left[\frac{1}{2}\mathcal{E}^2\right].$$

Let's look close to the random variable we need for this expected values, expressing them just as function of local fields. To be more concise, from now on with $\mathbb{E}$ we always mean the expected value over $(\lambda, \lambda^\star) \sim \mathcal{N}(0_{p+1}, \Omega)$. For the risk we just need

$$\mathbb{E}\left[\mathcal{E}^2\right] = \lambda_\star^4 + \frac{1}{p^2}\sum_{j,l=1}^{p} a_j a_l \mathbb{E}\left[\lambda_j^2\lambda_l^2\right] - \frac{2}{p}\sum_{j=1}^{p} a_j \mathbb{E}\left[\lambda_j^2\lambda_\star^2\right],$$

while for the ODE of $\bar{a}$

$$\mathbb{E}\left[\mathcal{E}\lambda_j^2\right] = \frac{1}{p}\sum_{l=1}^{p} a_l \mathbb{E}\left[\lambda_l^2\lambda_j^2\right] - \mathbb{E}\left[\lambda_\star^2\lambda_j^2\right],$$

where we omitted the noise part since it averages out with the expectation. The equation for $\bar{m}$ requires

$$2a_j\mathbb{E}\left[\mathcal{E}\lambda_j\lambda_\star\right] = \frac{2}{p}\sum_{l=1}^{p} a_j a_l \mathbb{E}\left[\lambda_l^2\lambda_j\lambda_\star\right] - 2\mathbb{E}\left[\lambda_\star^2\lambda_j\lambda_\star\right],$$

while we need two different expectations for $\bar{Q}$

$$2(a_j + a_l)\mathbb{E}\left[\mathcal{E}\lambda_j\lambda_\star\right] = 2(a_j + a_l)\left[\frac{1}{p}\sum_{s=1}^{p} a_s \mathbb{E}\left[\lambda_s^2\lambda_j\lambda_l\right] - \mathbb{E}\left[\lambda_\star^2\lambda_j\lambda_l\right]\right] \quad \text{and}$$

$$4\frac{\gamma}{p}\mathcal{E}^2 a_j a_l \lambda_j\lambda_l = 16\frac{\gamma}{p}a_j a_l\left[\mathbb{E}\left[\lambda_j\lambda_l\lambda_\star^4\right] - \frac{2}{p}\sum_{s=1}^{p} a_s\mathbb{E}\left[\lambda_j\lambda_l\lambda_\star^2\lambda_s^2\right] + \frac{1}{p^2}\sum_{s,t=1}^{p}\left(a_s a_t\mathbb{E}\left[\lambda_j\lambda_l\lambda_s^2\lambda_t^2\right] + \Delta\mathbb{E}\left[\lambda_j\lambda_l\right]\right)\right].$$

We are left to compute some distribution moments of a multivariate Gaussian of second, fourth and sixh order. In the anonymous repository can be found a Mathematica script to address this task; alternatively Isserlis' Theorem can be applied. We introduce a shorthand in the notation

$$\omega_{\alpha\beta} \coloneqq [\Omega]_{\alpha\beta},$$

where the indices α and β can discriminate between local fields λ (if $\alpha, \beta \in [1, \dots, p]$, and $\lambda_\star$ (if $\alpha, \beta = p + 1$). The final result are given by

$$\begin{aligned}
\mathbb{E}\left[\lambda_\alpha\lambda_\beta\right] &= \omega_{\alpha\beta} \\
\mathbb{E}\left[\lambda_\alpha^2\lambda_\beta^2\right] &= \omega_{\alpha\alpha}\omega_{\beta\beta} + 2\omega_{\alpha\beta}^2 \\
\mathbb{E}\left[\lambda_\alpha\lambda_\beta\lambda_\gamma^2\right] &= \omega_{\alpha\beta}\omega_{\gamma\gamma} + 2\omega_{\alpha\gamma}\omega_{\beta\gamma} \\
\mathbb{E}\left[\lambda_\alpha\lambda_\beta\lambda_\gamma^2\lambda_\delta^2\right] &= \omega_{\alpha\beta}\omega_{\gamma\gamma}\omega_{\delta\delta} + 2\omega_{\alpha\beta}\omega_{\gamma\delta}^2 + 2\omega_{\alpha\gamma}\omega_{\beta\gamma}\omega_{\delta\delta} + \\
&\qquad 4\omega_{\alpha\gamma}\omega_{\beta\delta}\omega_{\gamma\delta} + 4\omega_{\alpha\delta}\omega_{\beta\gamma}\omega_{\gamma\delta} + 2\omega_{\alpha\delta}\omega_{\beta\delta}\omega_{\gamma\gamma}
\end{aligned}$$

By retracing all steps backward and making the necessary substitutions, we can arrive at an explicit form of the ODEs and population risk. While the full risk expression can be found in Eq. (6), we report here just the case $a_j = 1$ for the ODEs since they have a compact matrix form

$$\frac{\mathrm{d}m}{\mathrm{d}t} = 2\left(\rho - \frac{\operatorname{Tr}[Q]}{p}\right)m + 4\left(\rho m - \frac{Qm}{p}\right) \tag{26a}$$

$$\begin{aligned}\frac{\mathrm{d}Q}{\mathrm{d}t} &= 4\left(\rho - \frac{\operatorname{Tr}[Q]}{p}\right)Q + 8\left(\frac{mm^\top}{k} - \frac{Q^2}{p}\right) \\ &+ \frac{4\gamma}{p}\Bigg\{\left[3\rho^2 Q + 12\rho mm^\top\right] + \frac{1}{p^2}\left[\left(\operatorname{Tr}[Q]^2 + 2\operatorname{Tr}\left[Q^2\right]\right)Q + 4\operatorname{Tr}[Q]Q^2 + 8Q^3\right] \\ &- \frac{2}{p}\Bigg[\left(\rho\operatorname{Tr}[Q] + 2\operatorname{Tr}\left[mm^\top\right]\right)Q + 2\operatorname{Tr}[Q]mm^\top \\ &+ 2\rho Q^2 + 4\left(mm^\top Q + Qmm^\top\right)\Bigg] + \Delta Q\Bigg\},\end{aligned} \tag{26b}$$

where $\rho \coloneqq {w_\star^2}/{d}$. For completeness, this is Eq. (6) for the case $a_j = 1$

$$\mathcal{R}(\Omega) = \frac{3+\Delta}{2} - \frac{\rho\operatorname{Tr}[Q] + 2\operatorname{Tr}\left[mm^\top\right]}{p} + \frac{\operatorname{Tr}[Q]^2 + 2\operatorname{Tr}\left[Q^2\right]}{2p^2}. \tag{27}$$

Appendix D. Spherically constrained ODE and SDE

D.1. Spherical constraint for ODE

Let's recall the update rule for the weights

$$w_j^{\nu+1} = \frac{w_j^\nu - \gamma \nabla_{w_j} \ell(y^\nu, f_{\Theta^\nu}(x^\nu))}{\left\| w_j^\nu - \gamma \nabla_{w_j} \ell(y^\nu, f_{\Theta^\nu}(x^\nu)) \right\|} \sqrt{d}, \tag{28}$$

we will find the leading order approximation of it and then apply the argument as the unconstrained case for deriving the ODEs. To shorten the notation we will use ℓ^ν for indicating $\ell(y^\nu, f_{\Theta^\nu}(x^\nu))$.

Let's start by computing the normalization factor

$$\begin{aligned}
\frac{1}{\left\| w_j^\nu - \gamma \nabla_{w_j} \ell^\nu \right\|} &= \left[\left(w_j^\nu - \gamma \nabla_{w_j} \ell^\nu \right) \cdot \left(w_j^\nu - \gamma \nabla_{w_j} \ell^\nu \right) \right]^{-\frac{1}{2}} \\
&= \left[\left\| w_j^\nu \right\|^2 - 2\gamma w_j^\nu \cdot \nabla_{w_j} \ell^\nu + \gamma^2 \left\| \nabla_{w_j} \ell^\nu \right\|^2 \right]^{-\frac{1}{2}} \\
&= \frac{1}{\sqrt{d}} \left[1 - \frac{1}{d} \left(2\gamma w_j^\nu \cdot \nabla_{w_j} \ell^\nu - \gamma^2 \left\| \nabla_{w_j} \ell^\nu \right\|^2 \right) \right]^{-\frac{1}{2}} \\
&= \frac{1}{\sqrt{d}} \left[1 + \frac{1}{2d} \left(2\gamma w_j^\nu \cdot \nabla_{w_j} \ell^\nu - \gamma^2 \left\| \nabla_{w_j} \ell^\nu \right\|^2 \right) + o\left(d^{-1} \right) \right].
\end{aligned}$$

Note that we kept both two terms in the expansion because we can show that both the norm and the scalar product with a weight vector are order 1

$$\begin{aligned}
\left\| \nabla_{w_j} \ell^\nu \right\|^2 &\sim \frac{2\mathcal{E}^2 \lambda_j^\nu}{p^2} \frac{\sum_{i=1}^d \left(x_i^\nu \right)^2}{d} \sim \frac{2\mathcal{E}^2 \lambda_j^\nu}{p^2} \frac{\chi_d^2}{d} = \mathcal{O}(1), \\
w_j \cdot \nabla_{w_l} \ell^\nu &\sim \frac{2\mathcal{E} \lambda_l^\nu}{p} \sum_{i=1}^d \frac{w_{j,i}}{\sqrt{d}} \mathcal{N}(0,1) \sim \frac{2\mathcal{E} \lambda_l^\nu}{p^2} \mathcal{N}\left(0, \sum_{i=1}^d \frac{w_{j,i}^2}{d} \right) \sim \frac{2\mathcal{E} \lambda_l^\nu}{p^2} \mathcal{N}(0,1) = \mathcal{O}(1).
\end{aligned}$$

We can now plug the expansion back into the original update rule

$$w_j^{\nu+1} = \left(w_j^\nu - \gamma \nabla_{w_j} \ell^\nu \right) \left[1 + \frac{1}{2d} \left(2\gamma w_j^\nu \cdot \nabla_{w_j} \ell^\nu - \gamma^2 \left\| \nabla_{w_j} \ell^\nu \right\|^2 \right) + o\left(d^{-1} \right) \right] \sqrt{d}.$$

We are ready to go over the steps that take us from the update rule on vector weights to those on order parameters. We report by way of example the steps performed for m; the accounts for Q are

similar, just a bit more tedious.

$$\begin{aligned} m_j^{\nu+1} &= \frac{w_j^{\nu+1}\cdot w^*}{d} \\ &= \left(m_j^\nu - \frac{\gamma w^*\cdot\nabla_{w_l}\ell^\nu}{d}\right)\left[1+\frac{1}{2d}\left(2\gamma w_j^\nu\cdot\nabla_{w_j}\ell^\nu - \gamma^2\left\|\nabla_{w_j}\ell^\nu\right\|^2\right)+o\left(d^{-1}\right)\right] \\ &= m_j^\nu - \frac{\gamma w^*\cdot\nabla_{w_l}\ell^\nu}{d} + \frac{m_{jr}^\nu}{2d}\left(2\gamma w_j^\nu\cdot\nabla_{w_j}\ell^\nu - \gamma^2\left\|\nabla_{w_j}\ell^\nu\right\|^2\right)+o\left(d^{-1}\right) \\ &= m_j^\nu + \frac{1}{d}\left[\frac{\gamma}{p}\lambda_\star\mathcal{E}^\nu\lambda_j^\nu - \frac{m_j^\nu}{2}\left(2\frac{\gamma}{p}\mathcal{E}^\nu\lambda_j^\nu\lambda_j^\nu + \frac{\gamma^2}{p^2}\mathcal{E}^\nu\lambda_j^{\nu 2}\right)\right]+o\left(d^{-1}\right). \end{aligned} \tag{29}$$

We can now take the limit $d\to+\infty$, claiming that the theorem [20, 43] proving ODE convergence is still valid. Indeed, the error term $o(d^{-1})$ in Eq. (29) has an average order of $O(d^{-2})$, which can be absorbed in the term Γ^ν of Theorem A.1 in [43]. The rest of the proof proceeds the same way, noting that the square function can be assumed to be Lipschitz since the dynamics take place on the sphere.

The differential equation that describes the evolution of m is

$$\frac{\mathrm{d}\bar{m}_j(t)}{\mathrm{d}t} = \mathbb{E}_{\lambda,\lambda^*\sim\mathcal{N}(0,\Omega(t))}\left[2\mathcal{E}\lambda_j\lambda_\star - \frac{\bar{m}_j(t)}{2}\left(4\mathcal{E}\lambda_j\lambda_j + 4\frac{\gamma}{p}\mathcal{E}^2\lambda_j^2\right)\right];$$

Using the definitions introduced in Equations (21) we can write the equation in a nicer form

$$\frac{\mathrm{d}\bar{m}_j(t)}{\mathrm{d}t} = \Psi_j(\Omega) - \frac{\bar{m}_j(t)}{2}\Phi_{jj}(\Omega). \tag{30}$$

Essentially, the spherical constraint can be imposed by using a term proportional to the unconstrained Q update.

Without reporting all the calculations, we can write an analogous differential equation for Q evolution

$$\frac{\mathrm{d}\bar{Q}_{jl}(t)}{\mathrm{d}t} = \Phi_{jl}(\Omega) - \frac{\bar{Q}_{jl}(t)}{2}\left(\Phi_{jj}(\Omega)+\Phi_{ll}(\Omega)\right). \tag{31}$$

Note that $\dfrac{\mathrm{d}\bar{Q}_{jj}(t)}{\mathrm{d}t} = 0$ if $Q_{jj}(t)=1$, as it should be since the norm of spherical vectors must not change.

In Figure 2 we show two examples of integration of ODEs, for different values of p. Simulating for large but finite d does not kills the stochasticity in the SGD runs, but we can clearly see how the ODE well describe the dynamics on average.

D.2. Spherical constraint for SDE

This subsection we derive the spherical constraint for the SDE, with $p=1$. We assume that the stochastic process is given

$$\begin{aligned} \mathrm{d}m &= \Psi_1(\Omega)\,\mathrm{d}t + \sqrt{\frac{\gamma}{d}}\,\sigma_m(\Omega)\cdot\mathrm{d}B_t \\ \mathrm{d}Q &= \Phi_{11}(\Omega)\,\mathrm{d}t + \sqrt{\frac{\gamma}{d}}\,\sigma_Q(\Omega)\cdot\mathrm{d}B_t, \end{aligned} \tag{32}$$

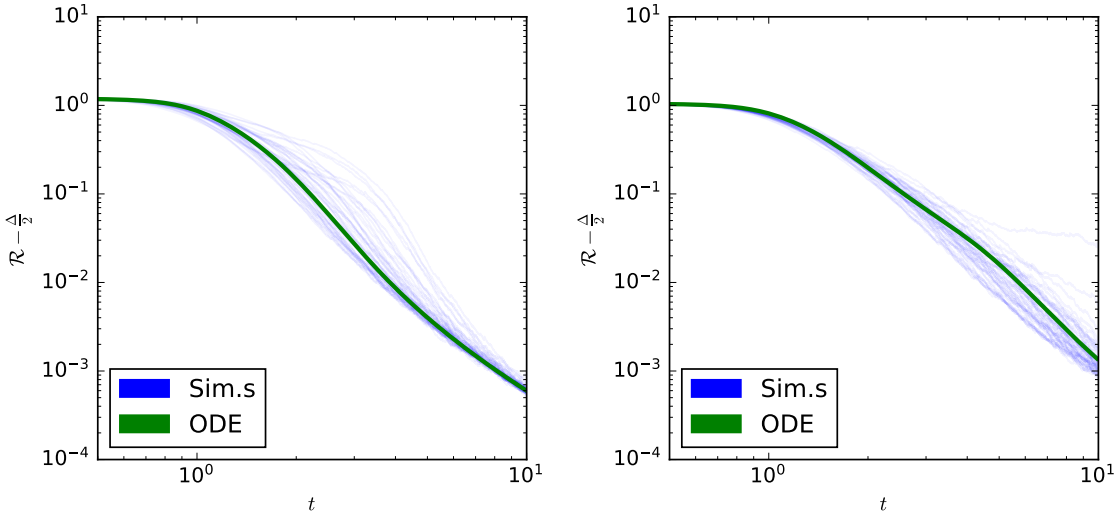

Figure 2: comparison of ODE integration and many SGD runs for $p = 5$ (left) and $p = 20$ (right). Both the experiments have $d = 5000$.

without providing explicit expressions for $\sigma_m(\Omega)$ and $\sigma_Q(\Omega)$; see Appendix G for that.

The derivation is basically following the steps of the previous Section. Starting from the unconstrainted update rule for the weights

$$w_j^{\nu+1} = w_j^\nu - \gamma \nabla_{w_j} \ell^\nu,$$

we can find an expression for the two unconstrained differentials

$$\begin{aligned} \mathrm{d}q &= \frac{-2\gamma w \cdot \nabla \ell^\nu + \gamma^2 \left\|\nabla \ell^\nu\right\|^2}{d} \\ \mathrm{d}m &= \frac{-\gamma w_\star \cdot \nabla \ell^\nu}{d}. \end{aligned} \tag{33}$$

Since we are forcing the weight on the sphere, the update rule that actually has to be used is

$$w_j^{\nu+1} = \frac{w_j^\nu - \gamma \nabla_{w_j} \ell^\nu}{w_j^\nu - \gamma \nabla_{w_j} \ell^\nu} \sqrt{d};$$

multiplying both sides by $w_\star$ and subtracting m we get

$$\mathrm{d}m_S = \frac{m + \mathrm{d}m}{\|w - \gamma \nabla \ell^\nu\|} \sqrt{d} - m,$$

where we introduce m_S to differentiate the constrained variable from m. Let's estimate the normalization factor

$$\begin{aligned} \|w - \gamma \nabla \ell^\nu\| =& \sqrt{(w - \gamma \nabla \ell^\nu)^2} = \sqrt{w^2 - 2w \cdot \gamma \nabla \ell^\nu + \gamma^2 \left\|\nabla \ell^\nu\right\|^2} \\ =& \sqrt{d} \sqrt{\frac{w^2}{d} + \frac{-2w \cdot \gamma \nabla \ell^\nu + \gamma^2 \left\|\nabla \ell^\nu\right\|^2}{d}} = \sqrt{d}\sqrt{q + \mathrm{d}q} \\ =& \sqrt{d}\sqrt{1 + \mathrm{d}q}, \end{aligned}$$

where in the last step we used the constraint $q = 1$. We can now plug it back in $\mathrm{d}m_S$

$$\mathrm{d}m_S = \frac{m + \mathrm{d}m}{\sqrt{1 + \mathrm{d}q}} - m,$$

and expanding up to leading orders we get

$$\begin{aligned}\mathrm{d}m_S =&(m + \mathrm{d}m)(1 + \mathrm{d}q)^{-\frac{1}{2}} - m = (m + \mathrm{d}m)\left(1 - \frac{1}{2}\,\mathrm{d}q + \frac{3}{8}\,\mathrm{d}q^2\right) - m \\ =&\,\mathrm{d}m - \frac{m}{2}\,\mathrm{d}q - \frac{1}{2}\,\mathrm{d}m\,\mathrm{d}q + \frac{3}{8}m\,\mathrm{d}q^2\end{aligned}$$

In principle, we can now use the Itô Lemma on differentials Equations (32), obtaining

$$\mathrm{d}m^2 = \frac{\gamma}{d}\sigma_m^2(\Omega)\,\mathrm{d}t, \quad \mathrm{d}q^2 = \frac{\gamma}{d}\sigma_m^2(\Omega)\,\mathrm{d}t, \quad \text{and} \quad \mathrm{d}m\,\mathrm{d}q = \frac{\gamma}{d}\sigma_m(\Omega)\cdot\sigma_Q(\Omega)\,\mathrm{d}t.$$

It's interesting to note that these lead to a drift correction (and not just stochastic), but it's of second order. As expected, in numerical simulations we can't see the effect for these corrections, hence we neglet them in what follows. Finally, we write the explicit Brownian motion for the constrained dynamic

$$\mathrm{d}m_S = \left(\Psi_1(\Omega) - \frac{m_S}{2}\Phi_{11}(\Omega)\right)\mathrm{d}t + \sqrt{\frac{\gamma}{d}}\left(\sigma_m(\Omega) - \frac{m_S}{2}\sigma_Q(\Omega)\right)\cdot\mathrm{d}B_t. \tag{34}$$

Of course, all functions depending on Ω should be evaluated at $m = m_S, q = 1$.

Appendix E. Derivation of the expected exit time formulas

E.1. Linearization of the equations

The linear approximation of Ψ around $m_j \approx 0$ is given by

$$\Psi_j = 4\left(m_j - \frac{m_j}{p}\right) = 4\left(1-\frac{1}{p}\right)m_j, \tag{35}$$

while for Φ we distinguish the cases $j = l$ or not

$$\begin{aligned} j \neq l \quad \Phi_{jl} &= 4\left[2\left(-2\frac{Q_{jl}}{p}\right)\right] + \\ &+ \frac{4\gamma}{p}\left\{3Q_{jl} - \frac{2}{p}\left[pQ_{jl} + 4Q_{jl}\right] + \frac{1}{p^2}\left[\left(p^2+2p\right)Q_{jl} + 8pQ_{jl} + 24Q_{jl}\right] + \Delta Q_{jl}\right\} \\ &= -\frac{16}{p}Q_{jl} + \frac{4\gamma}{p}\left\{3 - 2 - \frac{8}{p} + 1 + \frac{2}{p} + \frac{8}{p} + \frac{24}{p^2} + \Delta\right\}Q_{jl} \\ &= -\frac{16}{p}Q_{jl} + \frac{4\gamma}{p}\left\{2 + \frac{2}{p} + \frac{24}{p^2} + \Delta\right\}Q_{jl} \end{aligned} \tag{36}$$

$$\begin{aligned} j = l \quad \Phi_{jj} &= 4\left[2\left(-\frac{1}{p}\right)\right] + \frac{4\gamma}{p}\left\{3 - \frac{2}{p}\left[p+2\right] + \frac{1}{p^2}\left[\left(p^2+2p\right) + 4p + 8\right] + \Delta\right\} \\ &= -\frac{8}{p} + \frac{4\gamma}{p}\left\{3 - 2 - \frac{4}{p} + 1 + \frac{2}{p} + \frac{4}{p} + \frac{8}{p^2} + \Delta\right\} \\ &= -\frac{8}{p} + \frac{4\gamma}{p}\left\{2 + \frac{2}{p} + \frac{8}{p^2} + \Delta\right\} \end{aligned} \tag{37}$$

Given these linear approximations, we are ready to write down the equations valid as long as the risk stays in the first plateau

$$\begin{aligned} \frac{\mathrm{d}\left[m(t)\right]_j}{\mathrm{d}t} &= \left[4\left(1-\frac{1}{p}\right) + \frac{4}{p} - \frac{2\gamma}{p}\left(2 + \frac{2}{p} + \frac{8}{p^2} + \Delta\right)\right]\left[m(t)\right]_j \\ &= \left[4 - \frac{2\gamma}{p}\left(2 + \frac{2}{p} + \frac{8}{p^2} + \Delta\right)\right]\left[m(t)\right]_j \\ &= 4\left[1 - \frac{\gamma}{p}\left(1 + \frac{1}{p} + \frac{4}{p^2} + \frac{\Delta}{2}\right)\right]\left[m(t)\right]_j, \end{aligned} \tag{38}$$

$$\begin{aligned} \frac{\mathrm{d}\left[Q(t)\right]_{jl}}{\mathrm{d}t} &= \left[-\frac{16}{p} + \frac{4\gamma}{p}\left(2 + \frac{2}{p} + \frac{24}{p^2} + \Delta\right) + \frac{8}{p} - \frac{4\gamma}{p}\left(2 + \frac{2}{p} + \frac{8}{p^2} + \Delta\right)\right]\left[Q(t)\right]_{jl} \\ &= \left[-\frac{8}{p} + \frac{4\gamma}{p}\left(\frac{16}{p^2}\right)\right]\left[Q(t)\right]_{jl} \\ &= -\frac{8}{p}\left[1 - \frac{8\gamma}{p^2}\right]\left[Q(t)\right]_{jl}. \end{aligned} \tag{39}$$

We observe that the evolution of the sufficient statistics is uncoupled in the starting saddle. We can shorthand the notation by introducing ω_Q and ω_M

$$\begin{aligned} \frac{\mathrm{d}m_j}{\mathrm{d}t} &= \omega_M m_j \\ \frac{\mathrm{d}Q_{jl}}{\mathrm{d}t} &= -\omega_Q Q_{jl} \quad \text{when } j \neq l. \end{aligned} \tag{40}$$

These equations admit a simple solution given by

$$\begin{aligned} m_j(t) &= m_j(0)\exp[\omega_M t] \\ Q_{jl}(t) &= Q_{jl}(0)\exp[-\omega_Q t]. \end{aligned} \tag{41}$$

E.2. Solving the approximated risk equation

From Eq. (27) when the weights are on the sphere, it follows that the risk is given by

$$\mathcal{R}(Q,m) - \frac{\Delta}{2} = 1 + \frac{1}{p} + \frac{1}{p^2}\sum_{j,l=1;j\neq l}^{p} Q_{jl}^2 - \frac{2}{p}\sum_{j=1}^{p} m_j^2 \tag{42}$$

Eqs. (41) can be used to obtain a deterministic time evolution of the risk. The only source of randomness left is from the initial conditions, we can define two random variables as

$$\frac{\mu_0}{d} \coloneqq \sum_{j=1}^{p}\left[m_j(0)\right]^2 \quad \text{and} \quad \frac{\tau_0}{d} \coloneqq \sum_{j,l=1;j\neq l}^{p}\left[Q_{jl}(0)\right]^2, \tag{43}$$

and get an expression for the risk in function of time

$$\mathcal{R}(t) - \frac{\Delta}{2} = 1 + \frac{1}{p} + \frac{d\tau_0}{p^2}\exp[-2\omega_Q t] - \frac{2d\mu_0}{p}\exp[2\omega_M t] \tag{44}$$

This equation is not polished enough to be solved analytically yet. First of all we need to assume that the risk is decreasing by forcing $\omega_m > 0$. Secondly, we want the exponential proportional to τ_0 to be negligible when $t > 0$: this follow from $\omega_Q > 0$. In principle, this last condition is not needed for the process to converge like the first one, but without it is not possible to find an analytical solution for the cross -threshold equation. Moreover, when $p > 6$: $\omega_m > 0 \implies \omega_Q > 0$, so we can see that the request is not unreasonable.

Wrapping all this consideration together, Eq. (9) is

$$(1-T)\left(1 + \frac{1}{p} + \frac{\tau_0}{dp^2} - \frac{2\mu_0}{dp}\right) = 1 + \frac{1}{p} - \frac{2\mu_0}{dp}\exp[2\omega_M t_{\text{ext}}], \tag{45}$$

from where we can compute the exit time

$$t_{\text{ext}} = \frac{\log\left[\frac{Tp(p+1)d + (2\mu_0 p - \tau_0)(1-T)}{2\mu_0 p}\right]}{2\omega_m}.$$

E.3. Averaging on initial conditions

As of now t_{ext} is still depending on the initial conditions through the random variables μ_0 and τ_0. Following from the choosen initial conditions, and taking into account the dependence between the two random variables

$$\mu_0, \tau_0 \sim \mathcal{P}_p^d \quad \text{where } \mathcal{P}_p^d \equiv \left(d\sum_{j=1}^{p}(u_j \cdot v)^2, 2d\sum_{j=1}^{p}\sum_{l=j+1}^{p}(u_j \cdot u_l)^2 \right) \text{ with } v, u_j \sim \mathbb{S}^{d-1}(1).$$

We have now two possibility to get the expectation of t_{ext}. The first one is known in statistical physics literature as *quenched formula* leaves us with an unexpressed expected value

$$t_{\text{ext}}^{(\text{qnc})} = \mathbb{E}_{\mu_0,\tau_0\sim\mathcal{P}_p^d}\left[\frac{\log\left[\frac{Tp(p+1)d+(2\mu_0 p-\tau_0)(1-T)}{2\mu_0 p}\right]}{2\omega_m}\right]. \tag{46}$$

The second one, often referred as the *annealed formula*, is obtain by simply replacing the random variables with their expected values

$$t_{\text{ext}}^{(\text{anl})} = \frac{\log\left[\frac{T(p+1)d+(p+1)(1-T)}{2p}\right]}{2\omega_m}. \tag{47}$$

Case $p = 1$ For completeness, let's reduce these formulas to the simplest case $p = 1$. In this case τ_0 does not appear, and from Eq. (19) we find that $\mu_0 \sim \chi^2(1)$, so the quenched formula reduces to

$$t_{\text{ext}}^{(\text{qnc})} = \mathbb{E}_{\mu_0\sim\chi^2(1)}\left[\frac{\log\left[\frac{Td}{\mu_0}+(1-T)\right]}{8(1-6\gamma)-4\gamma\Delta)}\right]. \tag{48}$$

E.4. Overparameterization Gain

Let introduce the number of gradient step needed to escape the threshold. Remembering $t = \nu\gamma/pd$, we define

$$s_{\text{ext}}(p, d, \gamma, \Delta, T) \coloneqq \frac{pd}{\gamma} t_{\text{ext}}. \tag{49}$$

In the domain of our interest, namely $s_{\text{ext}} > 0$, s_{ext} is and convex function in γ. Therefore, it exist a unique minimum that correspond to the minimum number of steps required to cross the threshold when p, d are fixed

$$0 = \left.\frac{\partial s_{\text{ext}}(p, d, \gamma, \Delta, T)}{\partial\gamma}\right|_{\gamma=\gamma_{\text{opt}}(p,d,\Delta)} \quad \Longrightarrow \quad \gamma_{\text{opt}}(p, d, \Delta) = \frac{p^3}{8 + 2p + (2 + \Delta)p^2}$$

Note that this learning rate correspond to an exit time whose log's prefactor is constant to $\frac{1}{4}$, as reported in the main. We can also compute the optimal number of steps; we choose to stick with the *annealed formula*: for large p both the estimation lead to the same result, while for small p we are underestimating the exit time of a small factor. Hence, the annealed minimum number of steps is

$$s_{\text{ext}}^{\text{min}}(p, d, \Delta, T) \coloneqq s_{\text{ext}}^{\text{anl}}(p, d, \gamma_{\text{opt}}(p, d, \Delta), \Delta, T) = \frac{d\left[8 + 2p + (2 + \Delta)p^2\right]\log\left[\frac{T(p+1)d+(p+1)(1-T)}{2p}\right]}{4p^2}.$$

This formula can be used to estimate the overparametrization gain. Noting that $s_{\text{ext}}^{\min}$ is monotonically decrising in p, we can define the gain as

$$\lim_{d\to+\infty} \frac{s_{\text{ext}}^{\min}(p=1,d,\Delta,T)}{\lim_{p\to+\infty} s_{\text{ext}}^{\min}(p,d,\Delta,T)} = \frac{12+\Delta}{2+\Delta}.$$

Appendix F. Stochastic correction to exit time formula

In this section we derive a new exit time formula for the case $p = 1$, that takes into account the stochastic corrections of the dynamics.

At first, we require the further assumption $w^0 \perp w_\star^0$. Obviously, this is not realistic, since to achieve this initialization we should know $w_\star$ exactly. Still, this case is could arise interest, since it corresponds to $m^0 = 0$, that is a fixed point of Equation (7). Thus, there is no dynamics in the ODE description, while the SDE are able to jump out from the fixed point and reach the point of convergence of the SGD. Lastly, the following analysis can be generalized to be used with the usual initial conditions.

The key observation is approximating both the noise and the drift of Equation (34) to the first non-zero order, when evaluated at $m = 0$. As we pointed out above, the drift term is null at initialization, so the leading order is the first; this corresponds to the linearization of Equation (7), and the linear factor is

$$\mu := \left[4(1 - 6\gamma) - 2\gamma\Delta\right]. \tag{50}$$

Instead, the noise term is not vanishing at $m = 0$. Of course, $\sigma_Q(\Omega)$ does not contribute because is proportional to m_S. Moreover, looking at the expression for Σ in Appendix G, we can observe that the covariance matrix is diagonal since the non-diagonal term is proportional to m_S as well. All considered, the only term is contributing is the variance of m: the two-dimensional Brownian motion can be reduced to one in dimension 1, with variance

$$\sigma^2 := \frac{\gamma}{pd}(48 + 4\Delta).$$

Summarizing, the SDE near the initialization can be described as

$$\mathrm{d}m = \mu m \,\mathrm{d}t + \sigma \,\mathrm{d}b_t, \tag{51}$$

where b_t is a one-dimensional Wiener process with unit variance. The approximated evolution of m is a *expansive Ornstein–Uhlenbeck* process, since $\mu > 0$. We have already shown in the main that there is a direct map between m and the population risk; we can deal just with m for measuring the exit time. Given an expansive OU-process staring at $m = 0$, the mean first exit time from the interval $(-\sqrt{T}, \sqrt{T})$ is given by [44]

$$t_{\text{ext}}^{(SDE)} = \frac{T}{\sigma^2}\,{}_2F_2\left(1, 1; 3/2, 2; -\frac{\mu}{\sigma^2}T\right) \tag{52}$$

where ${}_2F_2$ is a generalized hypergeometric function; see [44] for reference. The formula does not have a closed form like the one derived earlier, and it works only when T is small enough that the exit point is not too far from initial conditions, otherwise the approximation we did do not hold anymore.

As reported in [44], the formula can be generalized also to the initial condition we used in the rest of the manuscript, where $m \sim \mathcal{N}\left(0, 1/\sqrt{d}\right)$. Again, we have to average over the initial conditions: (52) coincides with the *annealed* version, while we do not present here the *quenched* one to be concise.

Appendix G. SDEs for arbitrary width

In this appendix we generalize the discussion of the spherical constrained SDE to network of arbitrary width. As we already did at beginning of Section 3, we fix the second layer to $a_j = 1$ for everything follow.

G.1. Unconstrained SDE with $p > 1$

The updates of overlapping matrixes elements can be written as

$$\begin{aligned} \mathrm{d}Q_{jl} &= \frac{-\gamma(w_j \cdot \nabla_{w_l}\ell + w_l \cdot \nabla_{w_j}\ell) + \gamma^2 \nabla_{w_j}\ell^\nu \cdot \nabla_{w_l}\ell^\nu}{d} = \frac{\gamma}{pd}\mathcal{Q}_{jl}, \\ \mathrm{d}m_j &= \frac{-\gamma w_\star \cdot \nabla_{w_j}\ell}{d} = \frac{\gamma}{pd}\mathcal{M}_j. \end{aligned}$$

where we recalled the definition of the random variables $\mathcal{M}_j$ and $\mathcal{Q}_{jl}$; the factor $1/p$ is missing, but it will come out from the gradients ∇_{w_j}. As we already seen, the usual argument provides setting $\mathrm{d}t = \gamma/pd$ and say that the remaining factor is concentrating to its expected value

$$\begin{aligned} \mathrm{d}m_j &= \Psi_j \,\mathrm{d}t \\ \mathrm{d}Q_{jl} &= \Phi_{jl}\,\mathrm{d}t, \end{aligned}$$

where $\Phi_{jl} = \mathbb{E}\left[\mathcal{Q}_{jl}\right]$ and $\Psi_j = \mathbb{E}\left[\mathcal{M}_j\right]$. We can now go beyond this and add corrections to concentration, namely adding a Brownian motion, following [8]:

$$\begin{aligned} \mathrm{d}m_j &= \Psi_j \,\mathrm{d}t + \sqrt{\frac{\gamma}{pd}}\sigma_j^m \cdot \mathrm{d}B_t \\ \mathrm{d}Q_{jl} &= \Phi_{jl}\,\mathrm{d}t + \sqrt{\frac{\gamma}{pd}}\sigma_{jl}^Q \cdot \mathrm{d}B_t. \end{aligned} \tag{53}$$

The σ_j^m and σ_{jl}^Q are the rows of the matrix obtained by taking the square root of the covariance matrix of all the $p + p^2$ random variable $\mathcal{M}_j$ and $\mathcal{Q}_{jl}$:

$$\begin{pmatrix} \sigma_1^m \\ \vdots \\ \sigma_p^m \\ \sigma_{11}^Q \\ \vdots \\ \sigma_{pp}^Q \end{pmatrix} := \sqrt{\begin{pmatrix} \mathrm{Var}\left[\mathcal{M}_1\right] & \cdots & \mathrm{Cov}\left[\mathcal{M}_1, \mathcal{M}_p\right] & \mathrm{Cov}\left[\mathcal{M}_1, \mathcal{Q}_{11}\right] & \cdots & \mathrm{Cov}\left[\mathcal{M}_1, \mathcal{Q}_{pp}\right] \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \mathrm{Cov}\left[\mathcal{M}_p, \mathcal{M}_1\right] & \cdots & \mathrm{Var}\left[\mathcal{M}_p\right] & \mathrm{Cov}\left[\mathcal{M}_p, \mathcal{Q}_{11}\right] & \cdots & \mathrm{Cov}\left[\mathcal{M}_p, \mathcal{Q}_{pp}\right] \\ \mathrm{Cov}\left[\mathcal{Q}_{11}, \mathcal{M}_1\right] & \cdots & \mathrm{Cov}\left[\mathcal{Q}_{11}, \mathcal{M}_p\right] & \mathrm{Var}\left[\mathcal{Q}_{11}\right] & \cdots & \mathrm{Cov}\left[\mathcal{Q}_{11}, \mathcal{Q}_{pp}\right] \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \mathrm{Cov}\left[\mathcal{Q}_{pp}, \mathcal{M}_1\right] & \cdots & \mathrm{Cov}\left[\mathcal{Q}_{pp}, \mathcal{M}_p\right] & \mathrm{Cov}\left[\mathcal{Q}_{pp}, \mathcal{Q}_{11}\right] & \cdots & \mathrm{Var}\left[\mathcal{Q}_{pp}\right] \end{pmatrix}}, \tag{54}$$

and $\mathrm{d}B_t$ is a $p(1+p)$-dimensional Wiener process.

G.2. Spherical Constrained SDE

Let's move now to the actual spherical derivation. The steps used are exaclty the same we did in D, Since we are forcing the weight on the sphere, the update rule that has to be used is

$$w_j^{\nu+1} = \frac{w_j^\nu - \gamma \nabla_{w_j}\ell^\nu}{\left\| w_j^\nu - \gamma \nabla_{w_j}\ell^\nu \right\|}\sqrt{d};$$

We introduce now $\mathfrak{M}_j$ and $\mathfrak{Q}_{jl}$ to discriminate between the spherical variable and the unconstrained ones. Multiplying both sides of the update rule by $w_\star$ and subtracting $\mathfrak{M}_j$ we get

$$\mathrm{d}\mathfrak{M}_j = \frac{\mathfrak{M}_j + \mathrm{d}m_{jr}}{\left\|w_j - \gamma\nabla_{w_j}\ell^\nu\right\|}\sqrt{d} - \mathfrak{M}_j.$$

Similarly, the product of two update rules brings us to

$$\begin{aligned}\mathrm{d}\mathfrak{Q}_{jl} &= \frac{w_j - \gamma\nabla_{w_j}\ell^\nu}{\left\|w_j - \gamma\nabla_{w_j}\ell^\nu\right\|}\frac{w_l - \gamma\nabla_{w_l}\ell^\nu}{\left\|w_l - \gamma\nabla_{w_l}\ell^\nu\right\|} - \mathfrak{Q}_{jl}\\ &= \frac{\mathfrak{Q}_{jl} + \mathrm{d}Q_{jl}}{\left\|w_j - \gamma\nabla_{w_j}\ell^\nu\right\|\left\|w_l - \gamma\nabla_{w_l}\ell^\nu\right\|}d - \mathfrak{Q}_{jl}\end{aligned}$$

Let's estimate the normalization factor

$$\begin{aligned}\left\|w_j - \gamma\nabla_{w_j}\ell^\nu\right\| =& \sqrt{(w_j - \gamma\nabla_{w_j}\ell^\nu)^2} = \sqrt{w_j^2 - 2w_j\cdot\gamma\nabla_{w_j}\ell^\nu + \gamma^2\left\|\nabla_{w_j}\ell^\nu\right\|^2}\\ =& \sqrt{d}\sqrt{\frac{w_j^2}{d} + \frac{-2w_j\cdot\gamma\nabla_{w_j}\ell^\nu + \gamma^2\left\|\nabla_{w_j}\ell^\nu\right\|^2}{d}} = \sqrt{d}\sqrt{Q_{jj} + \mathrm{d}Q_{jj}}\\ =& \sqrt{d}\sqrt{1 + \mathrm{d}Q_{jj}},\end{aligned}$$

Expanding up to leading orders we get

$$\begin{aligned}\mathrm{d}\mathfrak{M}_j =& (\mathfrak{M}_j + \mathrm{d}m_{jr})(1 + \mathrm{d}Q_{jj})^{-\frac{1}{2}} - \mathfrak{M}_j\\ =& (\mathfrak{M}_j + \mathrm{d}m_{jr})\left(1 - \frac{1}{2}\mathrm{d}Q_{jj} + \frac{3}{8}\mathrm{d}q_{jj}^2\right) - \mathfrak{M}_j\\ =& \mathrm{d}m_{jr} - \frac{\mathfrak{M}_j}{2}\mathrm{d}Q_{jj} - \frac{1}{2}\mathrm{d}m_{jr}\,\mathrm{d}Q_{jj} + \frac{3}{8}\mathfrak{M}_j\,\mathrm{d}Q_{jj}^2\end{aligned}$$

$$\begin{aligned}\mathrm{d}\mathfrak{Q}_{jl} =& (\mathfrak{Q}_{jl} + \mathrm{d}Q_{jl})(1 + \mathrm{d}Q_{jj})^{-\frac{1}{2}}(1 + \mathrm{d}Q_{ll})^{-\frac{1}{2}} - \mathfrak{Q}_{jl}\\ =& (\mathfrak{Q}_{jl} + \mathrm{d}Q_{jl})\left(1 - \frac{1}{2}\mathrm{d}Q_{jj} + \frac{3}{8}\mathrm{d}q_{jj}^2\right)\left(1 - \frac{1}{2}\mathrm{d}Q_{ll} + \frac{3}{8}\mathrm{d}q_{ll}^2\right) - \mathfrak{Q}_{jl}\\ =& \mathrm{d}Q_{jl} - \frac{\mathfrak{Q}_{jl}}{2}(\mathrm{d}Q_{jj} + \mathrm{d}Q_{ll}) + \frac{\mathfrak{Q}_{jl}}{8}(3\,\mathrm{d}Q_{jj}^2 + 3\,\mathrm{d}Q_{ll}^2 + 2\,\mathrm{d}Q_{jj}\,\mathrm{d}Q_{ll}) - \frac{1}{2}(\mathrm{d}Q_{jl}\,\mathrm{d}Q_{jj} + \mathrm{d}Q_{jl}\,\mathrm{d}Q_{ll})\end{aligned}$$

We can now use the Itô Lemma on differentials Equations (53), obtaining

$$\mathrm{d}x_a\,\mathrm{d}x_b = \frac{\gamma}{pd}\sigma_{x_a}\cdot\sigma_{x_b}\,\mathrm{d}t,$$

so we have some extra drift terms at an higher order.

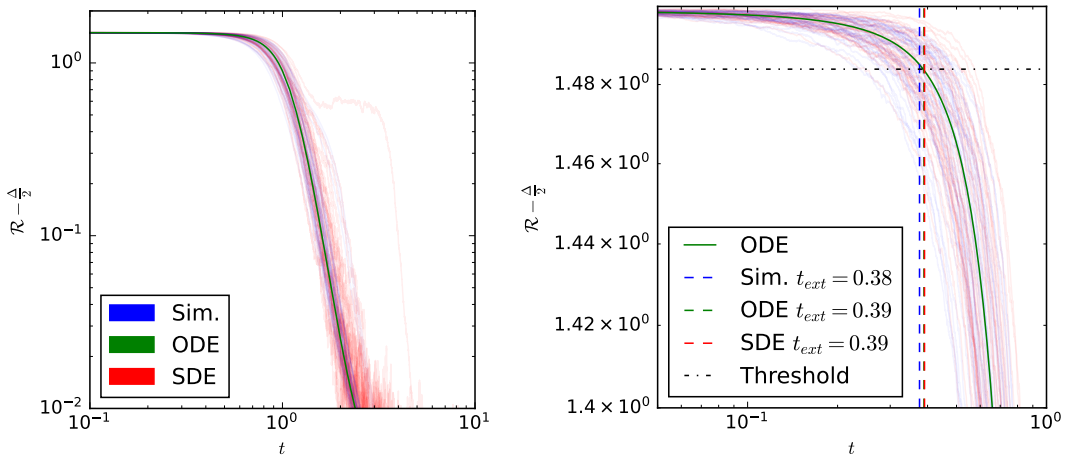

Figure 3: multiple run of the simulated SGD and the numerically integrated SDE, always starting from the same initial condition, with $d = 3000$. All the t_{ext} presented are obtained by solving numerically (9). The SDE captures the variance that the ODE doesn't exhibit, but the t_{ext} do not change considerably.

G.3. Special cases

Computing explicitly the variance in Equation (54) is not conceptually different from what exposed in Appendix C: expanding all the expression we are left expectations of polynomials of the preactivations. Even if not complex, it is required to compute up to twelfth moments of a multivariate normal distributions, that could lead to very long results. We developed a Mathematica script for addressing this computation, available in the GitHub repository; it can be used for the covariance of arbitrary network width. In the same repository, we provide the code for numerically integrating the SDE, in the cases $p = 1, 2$. These two special case are briefly discussed here.

Explicit expression for $p = 1$ We report here the expression for the variances and covariance introduced in (32)

$$\begin{aligned}
\text{Var}\left[\mathcal{Q}_{11}\right] =& 576\gamma^2\Delta m^4 - 2496\gamma^2\Delta m^2 - 11520\gamma^2 m^6 + 54144\gamma^2 m^4 - 73728\gamma^2 m^2 \\
&+ 544\gamma\Delta m^2 - 11136\gamma m^4 + 22272\gamma m^2 + 320 m^4 - 1600 m^2 + 32\gamma^2\Delta^2 \\
&+ 1920\gamma^2\Delta + 31104\gamma^2 - 544\gamma\Delta - 11136\gamma + 48\Delta + 1280 \\
\text{Cov}\left[\mathcal{M}_1, \mathcal{Q}_{11}\right] =& 72\gamma\Delta m^3 - 72\gamma\Delta m - 1440\gamma m^5 + 2880\gamma m^3 \\
&- 1440\gamma m + 24\Delta m - 480 m^3 + 480 m \\
\text{Var}\left[\mathcal{M}_1\right] =& 8\Delta m^2 - 192 m^4 + 144 m^2 + 4\Delta + 48
\end{aligned} \tag{55}$$

Numerical experiments for $p = 2$ In Figure 3 we show the same numerical experiment we presented in Section 2, repeated in the case $p = 2$. Again, we show there is no benefit including the stochasticity in the analysis. All the evidence indicate that even for larger p there is no effect in including the correction and the time estimation based on the ODE is accurate.

Appendix H. Landscape geometry

In this appendix we discuss the landscape of the population risk for the phase retrieval problem $p = 1$. We start by discussing the Euclidean case first, moving to the sphere next.

We recall the reader that the loss function is given by given by:

$$\ell(w) := \ell(y, f_\Theta(x)) = \frac{1}{2}\left((w_\star^\top x)^2 + \sqrt{\Delta}z - (w^\top x)^2\right)^2 \tag{56}$$

As before, we define the pre-activations (or local fields):

$$\lambda_\star = w_\star^\top x, \qquad\qquad \lambda = w^\top x \tag{57}$$

and the displacement vector:

$$\begin{aligned}\mathcal{E}(w) &:= (w_\star^\top x)^2 + \sqrt{\Delta}z - (w^\top x)^2 \\ &= \lambda_\star^2 + \sqrt{\Delta}z - \lambda^2\end{aligned} \tag{58}$$

Note that:

$$\nabla_w \mathcal{E}(w) = -2\lambda x \tag{59}$$

Therefore, the Euclidean gradient of the loss is given by:

$$\nabla_w \ell(w) = -2\mathcal{E}(w)\lambda x \tag{60}$$

And the Euclidean Hessian is:

$$\begin{aligned}\nabla^2 \ell(w) &= 2\left(2\lambda^2 - \mathcal{E}(w)\right) xx^\top \\ &= 2(3\lambda^2 - \lambda_\star^2 + \sqrt{\Delta}z)xx^\top\end{aligned} \tag{61}$$

Averaged geometry — We now compute the expected geometry by taking population averages of the above. It will be useful to define the correlation variables:

$$\rho = \frac{||w_\star||^2}{d}, \qquad\qquad m = \frac{w_\star^\top w}{d}, \qquad\qquad q = \frac{||w||^2}{d} \tag{62}$$

which we recall are the second moments of the pre-activations $(\lambda_\star, \lambda)$. With this notation, the population risk (the expected value of (56)) reads:

$$\mathcal{R}(m, q) = \Delta/2 + 3\rho^2 + 3q^2 - 4m^2 - 2\rho q \tag{63}$$

In order to compute the expected gradient, we need the following moments:

$$\mathbb{E}[\lambda_\star^2 \lambda x] = \rho\theta + 2mw_\star, \qquad\qquad \mathbb{E}[\lambda_\star^3 x] = 3qw \tag{64}$$

Which gives:

$$\nabla_w \mathcal{R}(w) = -2\mathbb{E}\left[\mathcal{E}(w)\lambda x\right] = -2((\rho - 3q)w + 2mw_\star) \tag{65}$$

Finally, let's compute the expected Hessian. For that, we will need the following moments:

$$\mathbb{E}[\lambda_\star^2 x x^\top] = \rho I_d + w_\star w_\star^\top, \qquad \mathbb{E}[\lambda^2 x x^\top] = q I_d + w w^\top \tag{66}$$

Therefore:

$$\nabla_w^2 \mathcal{R}(w) = 2\mathbb{E}\left[(3\lambda^2 - \lambda_\star^2 + \sqrt{\Delta}z) x x^\top\right] = 2\left(-(\rho - 3q) I_d + 3 w w^\top - w_\star w_\star^\top\right) \tag{67}$$

We are now ready to compute the critical points of the Euclidean landscape and evaluate their nature. By definition, the critical points are defined as solutions of $\nabla_w \mathcal{R}(w) = 0$. These are:

- $w = 0$ ($(m, q) = (0, 0)$): The Hessian of this critical point is given by:

$$\nabla_w^2 \mathcal{R}(0) = -2(\rho I_d + w_\star w_\star^\top) \prec 0 \tag{68}$$

  This is a negative-definite matrix with $d - 1$ negative eigenvalues -2ρ and one negative eigenvalue $-2(\rho + 1)$ with eigenvector $\theta_\star$. Therefore, this is a local maximum. The risk associated is given by:

$$\mathcal{R}(0) = \frac{\Delta}{2} + 3\rho^2 \tag{69}$$

- $(m, q) = (0, \rho/3)$: This defines a line of critical points $\{w \in \mathbb{R}^d : w \perp w_\star \text{ and } ||w|| = 1/\sqrt{3}||w_\star||\}$. The Hessian is given by:

$$\nabla_w^2 \mathcal{R} = 2(3 w w^\top - w_\star w_\star^\top) \tag{70}$$

  Note this is a rank-two matrix with $d - 2$ zero eigenvalues (associated to flat directions), one negative eigenvalue with eigenvector $w_\star$ and a positive eigenvalue with eigenvector perpendicular to the minima. This is a saddle-point, and have population risk:

$$\mathcal{R}(0, \rho/3) = \frac{\Delta}{2} + \frac{10}{3}\rho^2 \tag{71}$$

- $w = \pm w_\star$ ($(m, q) = (\pm\rho, \rho)$): From the definition of our problem, this is the global minima. The expected Hessian is given by:

$$\nabla_w^2 \mathcal{R}(\pm w_\star) = 4(\rho I_d + w_\star w_\star^\top) \succ 0 \tag{72}$$

  which is indeed a positive definite matrix. This defines the minimum achievable population risk:

$$\min_{w \in \mathbb{R}^d} \mathcal{R}(w) = \mathcal{R}(\pm\rho, \rho) = \frac{\Delta}{2} \tag{73}$$

This is consistent with the discussion in [14], where the critical points and their nature were reported, but explicit expressions for the expected gradient and Hessian were not given. From this geometry, a neat picture for the one-pass SGD dynamics in the unconstrained problem can be drawn. Consider $w_\star \in \mathbb{S}^{d-1}$ with a random initialization at high-dimensions:

$$w^0 \sim \mathcal{N}(0, I_d). \tag{74}$$

Note that with high-probability the random initial weights is almost orthogonal to the signal, and we have $(m, q) \approx (0, 1)$. Note this is not a critical point, and the initial gradient $\nabla_w \mathcal{R}(w^0) = 4w^0$ is orthogonal to $w_\star$. Indeed, for $p = 1$ and $\Delta = 0$ the unconstrained ODEs (17) are given by:

$$\dot{m}(t) = 6\, m(t)(\rho - q(t)) \tag{75}$$

$$\dot{q}(t) = 4\left(q(t)(\rho - 3q(t)) + 2m(t)^2\right) + 12\gamma\left(q(t)(\rho^2 + 5q(t)^2 - 2\rho q(t)) + 4m^2(\rho - 2q(t))\right) \tag{76}$$

where $\dot{} := {}^{\mathrm{d}}/_{\mathrm{d}t}$ and we dropped the bars for clarity. Therefore, in the initial stage of the dynamics $\dot{m} \approx 0$ remains almost constant, while q decreases, and the dynamics flow in the direction of the saddle-point $(0, {}^{\rho}/_{3})$. As we have seen, the saddle is mostly flat, with a single negative curvature direction pointing towards $w_\star$. This is precisely the mediocrity stage, where the dynamics slows down and SGD gets stuck for a long time before being able to develop significant correlation with $w_\star$ and escape.

H.1. Landscape in the sphere

Recall that the orthogonal projector on a vector $u \in \mathbb{S}^{d-1}$ is given by:

$$\mathrm{Proj}_u = I_d - uu^\top \tag{77}$$

Therefore, the gradient on the sphere is given by:

$$\begin{aligned}\mathrm{grad}_{\mathbb{S}^{d-1}}\ell(w) &= \mathrm{Proj}_{\mathbb{S}^{d-1}}(\nabla_w \ell(w)) = (I_d - ww^\top)\nabla_w \ell(w) \\ &= -2\mathcal{E}(w)\lambda(x - \lambda w)\end{aligned} \tag{78}$$

Similarly, the Hessian on the sphere can be written as:

$$\begin{aligned}\mathrm{Hess}_{\mathcal{S}^{d-1}}\ell(w) &= \mathrm{Proj}_{\mathbb{S}^{d-1}}\left(\nabla_w^2 \ell(w)\right) - \langle w, \nabla_w \ell(w)\rangle I_d \\ &= (3\lambda^2 - \lambda_\star^2)(xx^\top - \lambda\theta x^\top) + \mathcal{E}(w)\lambda^2 I_d\end{aligned} \tag{79}$$

Averaged geometry — Recall that in the sphere we have $\rho = q = 1$. Therefore, the population risk now only depends on the correlation $m = \langle w_\star, w\rangle$, and reads:

$$\mathcal{R}(w) = 2(1 - m^2) + \frac{\Delta}{2} \tag{80}$$

Luckily, half of the moments we need have been already computed above. To get the gradient on the sphere, we just need to compute:

$$\mathbb{E}[\lambda_\star^2 \lambda^2] = 1 + 2m^2, \qquad \mathbb{E}[\lambda^4] = 3 \tag{81}$$

Therefore, the averaged spherical gradient is given by:

$$\mathrm{grad}_{\mathbb{S}^{d-1}}\mathcal{R}(w) = 4m\,(mw - w_\star) \tag{82}$$

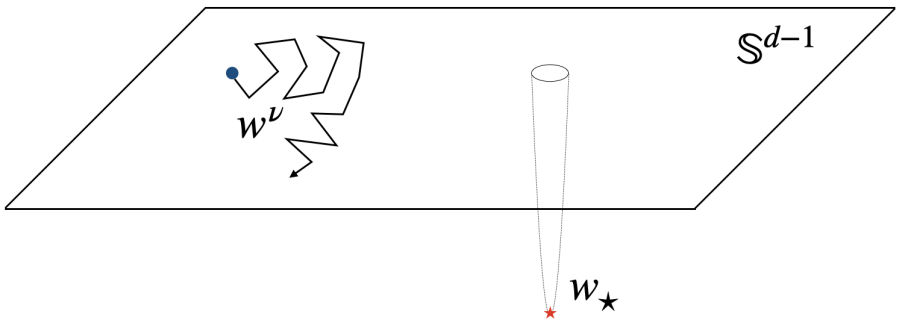

Figure 4: Low-dimensional illustration of mediocrity at initialization. As discussed in Sec. H.1, the expected Hessian at initialization is a strict saddle-point with $d-1$ flat directions and a single negative direction pointing towards the global minimum. This scenario is particularly hard for descent based algorithms such as SGD, that require $n = (d \log d)$ samples / steps to develop significant correlation with the signal.

We also have all moments required to compute the expected spherical Hessian:

$$\mathrm{Hess}_{\mathcal{S}^{d-1}} \mathcal{R}(w) = 4\left[m^2 I_d + m w w_\star^\top - w_\star w_\star^\top\right] \tag{83}$$

As before, we can now analyze the critical points and their nature.

- $w \perp w_\star$ ($m = 0$): The Hessian is given by:

$$\mathrm{Hess}_{\mathcal{S}^{d-1}} \mathcal{R}(w) = -4 w_\star w_\star^\top \tag{84}$$

  Which is a rank one matrix with $d-1$ eigenvalues 0 (flat directions) and a single negative eigenvalue -4 with negative curvature pointing towards the signal $w_\star$. Therefore, this is a strict saddle-point. However, since the risk is a decreasing function of $m^2 \in [0,1]$, this is also the global maximum of the risk.

- $w = \pm w_\star$ ($m = \pm 1$): As before, these are the global minima, and define the minimal achievable risk $\mathcal{R}(\pm w_\star) = \Delta/2$. Indeed, the Hessian is given by:

$$\mathrm{Hess}_{\mathcal{S}^{d-1}} \mathcal{R}(\pm w_\star) = 4 I_d \succ 0 \tag{85}$$

  which is positive-define.

Therefore, the landscape now resembles a golf course: completely flat with a single whole corresponding with the global minimum $w_\star$, see Fig. 4 for an illustration. This is the prototypical image of mediocrity. In particular, differently from the unconstrained case, random initialization

$$w^0 \sim \mathrm{Unif}(\mathbb{S}^{d-1}(\sqrt{d})) \tag{86}$$

now corresponds to initializing close to the saddle-point point $m^0 = 0$: with high-probability the initial weights are orthogonal to the signal at high-dimensions:

$$m^0 = 1/d \langle w^0, w_\star \rangle \approx 1/\sqrt{d} \ll 1. \tag{87}$$

For the reader convenience, we recall that the ODEs (7) describing the evolution of the correlation m is given by:

$$\dot{m}(t) = m(t)\left[4(1-6\gamma)(1-m^2(t)) - 2\gamma\Delta\right] \tag{88}$$

Close to initialization we now have $\dot{m}(0) \approx 0$, slowing down the dynamics close to initialization.

Appendix I. Training the second layer

In the previous section, we have derived analytical expressions for the exit time in the particular case of fixed first layer weights $a_j^0 = 1$. Here, we provide numerical evidence that training the first layer does not significantly change our conclusions.

The key challenge is that by training the first layer we can't measure t_{ext} as the time needed to escape the risk at initialization. Indeed, from equation (9) it can be seen that in the very first steps of the learning the vector a changes slightly to adapt to the initial conditions, thereby fitting the noise. In this scenario, instead of looking directly at the risk, we can instead use the largest component of the correlation vector m as a measure on how much the network has learned. At random initialization, this is of order $1/\sqrt{d}$, grows to 1 as the neural network correlates with the target weights. A natural choice for initializing the second layer weights is $a_j^0 = 1, \forall j \in [p]$. In principle, this initial condition guarantees that the risk at initialization is exactly equal to the case where a_j is fixed. On the other hand, as we already point out, the initial plateau where the dynamics gets stuck depends on the particular first layer initial condition. Even for other choices of initialization, e.g. $a_j \sim \text{Bernoulli}(1/2)$, the dynamics quickly goes the a plateau, so it does not really matter which a_j^0 is used. Therefore, for simplicity we choose an homogeneous initialization $a_j^0 = 1$. Figure 5 compares the evolution of the maximum correlation when learning the second layer or not, for different values of p.

It is important to stress that we are not claiming that the time needed to reach the minimum of the population risk is the same when training or not the second layer, as can be seen in Fig. 5. Instead, our result highlights that the time needed to escape the flat directions at initialization are close. In fact, after the two layer neural network has escaped mediocrity, the dynamics can be very different whether the second layer weights are trained or not. For instance, a could become sparse with just a few neurons contributes to the output, or it could remain close to homogeneous $a_j = 1$, and with all neurons correlating with the target. Although studying the dynamics after escaping mediocrity is surely an interesting endeavor, it's out of the scope of this manuscript.

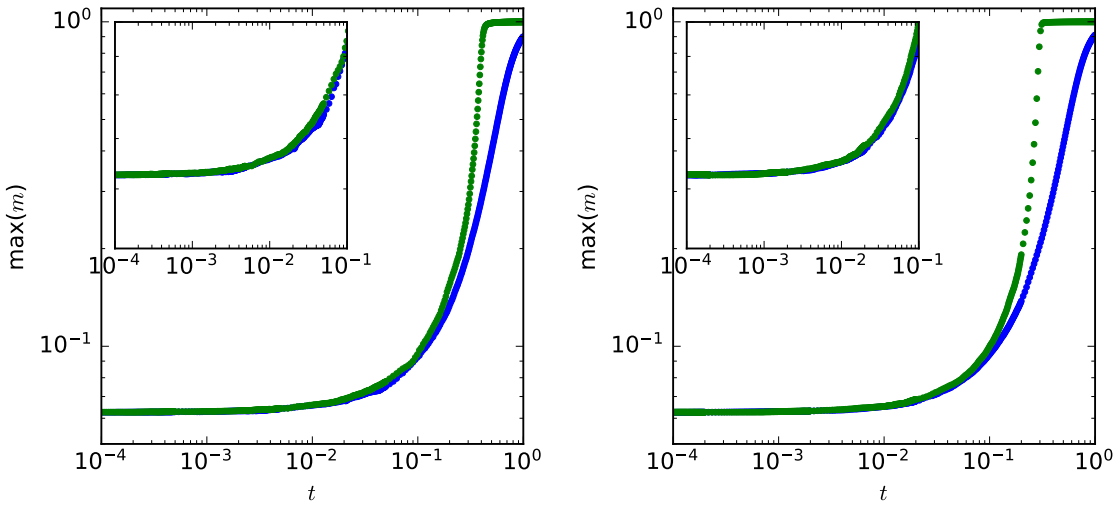

Figure 5: $p = 20$ (left), $p = 50$ (right), $d = 1000$. Comparison between the growth of $\max m$ throughout the learning process, when the second layer is fixed (blue) and trained (green). The dynamics is obviously different far from the starting point, but when we zoom close to the exit point, the two processes have the same behavior, t_{ext} included.