Fundamental limits of weak learnability in high-dimensional multi-index models

author names withheld

Under Review for the Workshop on High-dimensional Learning Dynamics, 2024

Abstract

Multi-index models — functions which only depend on the covariates through a non-linear transformation of their projection on a subspace — are a useful benchmark for investigating feature learning with neural networks. This paper examines the theoretical boundaries of learnability in this hypothesis class, focusing particularly on the minimum sample complexity required for weakly recovering their low-dimensional structure with first-order iterative algorithms, in the high-dimensional regime where the number of samples is $n = \alpha d$ is proportional to the covariate dimension d. Our findings unfold in three parts: (i) first, we identify under which conditions a *trivial subspace* can be learned with a single step of a first-order algorithm for any $\alpha > 0$; (ii) second, in the case where the trivial subspace is empty, we provide necessary and sufficient conditions for the existence of an *e*asy subspace consisting of directions that can be learned only above a certain sample complexity $\alpha > \alpha_c$. The critical threshold α_c marks the presence of a computational phase transition, in the sense that no efficient iterative algorithm can succeed for $\alpha < \alpha_c$. In a limited but interesting set of really hard directions —akin to the parity problem— α_c is found to diverge. Finally, (iii) we demonstrate that interactions between different directions can result in an intricate hierarchical learning phenomenon, where some directions can be learned sequentially when coupled to easier ones. Our analytical approach is built on the optimality of approximate message-passing algorithms among first-order iterative methods, delineating the fundamental learnability limit across a broad spectrum of algorithms, including neural networks trained with gradient descent.

1. Introduction

A fundamental property of neural networks is their ability to learn features from data and adapt to relevant structures in high-dimensional noisy data. However, our mathematical understanding of this mechanism remains limited. A popular model for studying this question is *multi-index models*. Multi-index functions are a class of statistical models encoding the inductive bias that the relevant directions for prediction depend only on a low-dimensional subspace of the covariates. They define a rich class of hypotheses [41], containing many widely studied functions in the statistical learning and theoretical computer science literature. Training a multi-index model typically translates to a non-convex optimization problem. Several authors have thus used multi-index models as a test-bed for understanding the behavior of neural nets and gradient-descent in non-convex, high-dimensional contexts, e.g. [1–4, 8, 14, 15, 21, 24, 47, 50–52]. We refer to Appendix A for an additional discussion.

To serve as a useful benchmark, it is necessary to have an idea of the fundamental limit of learnability in such models. This translates to the question of how many observations from the model are required to obtain a better-than-random prediction within a class of algorithms, also known as *weak learnability*. This can be studied both *statistically* (within the class of all, including exponential, algorithms) or *computationally* (restricted to a particular computational class, such as first-order algorithms). In the single-index case ($p = 1$), weak learnability has been heavily studied under probabilistic assumptions for the weights and data distribution (e.g. i.i.d. Gaussian or uniformly in the sphere). The statistical threshold for learnability has been characterised by [12] when the covariate dimension d is large. Optimal computational thresholds for the class of first-order iterative algorithms were derived in [18, 42, 43, 46] in the same regime. Similar results were also proven for other computational models: [23] provided a lower bound $n \geq d^{\max(1,\ell/2)}$ under the Correlational Statistical Query (CSQ) model, comprising algorithms that take queries of the type $\mathbb{E}[y\varphi(\boldsymbol{x})]$. [26] provided results for the Statistical Query (SQ) model, which allows for more flexible queries of the type $\mathbb{E}[\varphi(\boldsymbol{x}, y)]$, hence a lower sample complexity $n \geq d^{\max(1,\kappa/2)}$ with $\kappa \leq \ell$ defining the *generative exponent*. Aside from the particular case of committee machines [5, 20, 29, 35], results for general multi-index models $p > 1$ are scarce, as they crucially depend on the way different directions are coupled by the link function. The goal of the present work is precisely to close this gap for the class of Gaussian multi-index models in the proportional, high-dimensional regime, and to provide a classification of how hard it is to learn feature directions from data in multi-index models.

2. Settings and definitions

As motivated in the introduction, our main focus in this work will be to study subspace identifiability in the class of *Gaussian multi-index models*.

Definition 1 (Gaussian multi-index models) *Given a covariate $\boldsymbol{x} \sim \mathcal{N}(0, {}^1\!/\!_d\boldsymbol{I}_d)$, we define the class of Gaussian multi-index models as likelihoods of the type $\mathbb{E}[y|\boldsymbol{x}] = g(\boldsymbol{W}^\star\boldsymbol{x})$ where $g : \mathbb{R}^p \to \mathbb{R}$ is the link function and $\boldsymbol{W}^\star \in \mathbb{R}^{p\times d}$ is a weight matrix with i.i.d. rows $\boldsymbol{w}_k^\star \sim \mathcal{N}(0, \boldsymbol{I}_d)$. Note that this uniquely defines a joint distribution $p(\boldsymbol{x}, y)$ over $\mathbb{R}^{d+1}$.*

Given n i.i.d. samples $(\boldsymbol{x}_i, y_i)_{i\in[n]}$ drawn as per Definition 1, we are interested in investigating the computational bottlenecks of estimating $\boldsymbol{W}^\star$ from the samples $(\boldsymbol{x}_i, y_i)_{i\in[n]}$. Note that reconstructing $\boldsymbol{W}^\star$ or a permutation of its rows is equivalent from the perspective of the likelihood theorem 1. Therefore, in this work, we will be interested in *weak subspace learnability*, which corresponds to obtaining an estimation of the subspace spanned by $\boldsymbol{W}^\star$ better than a random estimator. This can be defined in an invariant way as follows

Definition 2 (Weak subspace recovery) *Let $V^\star \subset \mathbb{R}^p$ denote a subspace spanned by vectors representing components along $\boldsymbol{W}^\star$ such that each $\mathbf{v} \in V^\star$ maps to a vector $\mathbf{v}_d$ in $\mathbb{R}^d$ through the map $\mathbf{v}_d = (\boldsymbol{W}^\star)^\top\mathbf{v}$. Given an estimator $\hat{\boldsymbol{W}} \in \mathbb{R}^{p\times d}$ of $\boldsymbol{W}^\star$, we have weak recovery of a $V^\star$ if:*

$$\inf_{\boldsymbol{v}\in V^\star, \|\boldsymbol{v}\|=1} \left\| \frac{\hat{\boldsymbol{W}}(\boldsymbol{W}^\star)^\top\boldsymbol{v}}{d} \right\| = \Theta_d(1). \tag{1}$$

with high probability as $d \to \infty$.

Our main tool for characterising the computational bottlenecks in the Gaussian multi-index problem is an *Approximate Message Passing* (AMP) Algorithm 1 tailored to our problem, which we describe in Appendix H. The key property of AMP is that for a well-chosen $\boldsymbol{g}_{\rm out}$, it is provably optimal within the class of first-order methods [18]. This means that establishing learnability for the optimal AMP algorithm 1 implies a computational lower bound in the class of first-order methods, which includes popular machine learning algorithms such as gradient descent (SG). For the Gaussian multi-index estimation problem 1, the optimal $\boldsymbol{g}_{\rm out}$ is simply given by the optimal denoiser of an effective p-dimensional problem $Y = g(\boldsymbol{Z})$ with $\boldsymbol{Z} \sim \mathcal{N}(\boldsymbol{\omega}, \boldsymbol{V}_p)$:

$$\boldsymbol{g}_{\text{out}}(y, \boldsymbol{\omega}, \boldsymbol{V}) = \mathbb{E}[\boldsymbol{Z}|Y=y] = \frac{\int_{\mathbb{R}^p} \frac{\mathrm{d}\boldsymbol{z}}{\sqrt{\det 2\pi \boldsymbol{V}}} e^{-\frac{1}{2}(\boldsymbol{z}-\boldsymbol{\omega})^\top \boldsymbol{V}^{-1}(\boldsymbol{z}-\boldsymbol{\omega})} P(y|\boldsymbol{z}) V^{-1}(\boldsymbol{z}-\boldsymbol{\omega})}{\int_{\mathbb{R}^p} \frac{\mathrm{d}\boldsymbol{z}}{\sqrt{\det 2\pi \boldsymbol{V}}} e^{-\frac{1}{2}(\boldsymbol{z}-\boldsymbol{\omega})^\top \boldsymbol{V}^{-1}(\boldsymbol{z}-\boldsymbol{\omega})} P(y|\boldsymbol{z})}, \tag{2}$$

where the conditional expectation is defined through the output channel $Y = g(\boldsymbol{Z})$ with $\boldsymbol{Z} \sim \mathcal{N}(\boldsymbol{\omega}, \boldsymbol{V}_p)$. Here $\boldsymbol{\omega}$ can be interpreted as an estimate of the pre-activations.

Our analysis is based on two remarkable properties that make AMP a particularly useful tool for studying high-dimensional estimation. The first property is that for any $t < O(\log d)$, in the high-dimensional limit $d \to \infty$ the performance of AMP can be tracked without actually running the algorithm. This result, known as *state evolution*, makes AMP mathematically tractable in high-dimensions [13]. The second property is its optimality with respect to Bayesian estimation. We discuss this further in Appendix C. For multi-index models, the state evolution equations were derived by [7] and rigorously proven by [34]: in the large asymptotic limit $1/d \hat{\boldsymbol{W}}^t \boldsymbol{W}^{\star\top}$ converge in probability to $\boldsymbol{M}^t$, which evolves as $\boldsymbol{M}^{t+1} = F(\boldsymbol{M}^t)$. The specific form of F is detailed in Appendix B. This maps the problem of characterising the computational bottlenecks of first-order methods for high-dimensional Gaussian multi-index models to the study of the deterministic, p-dimensional dynamical system.

3. The trivial subspace ($\alpha_c = 0$)

We denote by $\mathcal{S}_p^+$, the cone of positive-semi-definite matrices in $\mathbb{R}^p$ and by $\succ$ the associated partial ordering. A starting point is identifying its fixed points and their basins of attraction. In the absence of any prior information on $\boldsymbol{W}^\star$ aside from its distribution, one cannot do better than taking $\hat{\boldsymbol{w}}_k^{t=0} \sim \mathcal{N}(0, \boldsymbol{I}_d)$ with $k \in [p]$ independently at random from the prior. With high-probability, at initialisation, the elements of the overlap matrix $1/d \hat{\boldsymbol{W}}^0 \boldsymbol{W}^{\star\top}$ are $\Theta_d(d^{-1/2})$ numbers. The asymptotic overlap for an uninformed initial condition is thus identically zero $\boldsymbol{M}^0 = 0$, a null-rank matrix. If $\boldsymbol{M}^0 = 0$ is not a fixed point, then $\boldsymbol{M}^1 \succ 0$, implying the weak recovery of a subspace of dimension $k = \mathrm{r}ank(\boldsymbol{M}^1) > 0$ with just a single step of AMP.

Lemma 3 (Existence of uninformed fixed point) *$\boldsymbol{M} = 0 \in \mathbb{R}^{p\times p}$ is a fixed point of eq. (7) if and only if the following condition holds almost surely over Y:*

$$\boldsymbol{g}_{\text{out}}(y, \boldsymbol{\omega} = 0, \boldsymbol{V} = \boldsymbol{I}_p) = \mathbb{E}[\boldsymbol{Z}|Y=y] = \boldsymbol{0}, \tag{3}$$

This implies that as long as the conditional expectation above is not zero almost surely, AMP weakly learns a non-empty subspace immediately in the first iteration for any number of samples $n = \Theta(d)$. For this reason, we refer to this subspace as a "trivial subspace" $T^\star$ composed of all directions $\mathbf{v} \in \mathbb{R}^p$ such that $\mathbb{E}[\mathbf{v}^\top \boldsymbol{Z}|Y=y] \neq \boldsymbol{0}$. Note that for single-index models ($p=1$), the condition in eq. (3) reduces to the one derived by [12, 44, 46]. Interestingly, this is *exactly* the same condition appearing in [26] for weak learnability of single-index models in the SQ model; see eq. (3) therein. We elaborate on this in Appendix E. To make Lemma 3 concrete, let us look at a few examples. A detailed derivation of these examples is discussed in the Appendix G.

(a) For single-index models ($p = 1$), T^* is one dimensional if and only if g is non-even, e.g. $g(z) = \mathrm{He}_3(z)$. This follows from requiring that $g_{out}(y, 0, 1) \neq 0$ for at least one value of y. In particular, on any open interval where g_{out} is invertible we have $g_{out} = g^{-1}$.

(b) For a committee $g(\boldsymbol{z}) = \sum_{i=1}^p sign(z_i)$, the trivial subspace $T^\star$ is again 1-d, spanned by $\mathbf{1} \in \mathbb{R}^p$.

(c) For monomials $g(\boldsymbol{z}) = z_1 \dots z_p$, the trivial subspace $T^\star$ is non-empty if and only if $p = 1$.

(d) For staircase functions [2] $g(\boldsymbol{z}) = z_1 + z_1 z_2 + z_1 z_2 z_3 + \cdots + z_1 \ldots z_p$, the trivial subspace is $T^\star = \mathbb{R}^p$ and is spanned by the canonical basis. In other words, AMP learns all the directions with a single step for any $\alpha > 0$.

This implies that when $T^\star$ is empty, running one-pass SGD after applying any transformation $\mathcal{T}$ fails to obtain weak recovery with $O(d)$ samples. The situation can be very different when reusing batches: [27] showed that full batch GD implicitly applies a transformation $\mathcal{T}$ to the labels, in a mechanism similar to the one discussed above. However, this transformation may not be optimal. For instance, while AMP learns the trivial subspaces of both $g(z) = \mathrm{He}_3(z)$ and staircase functions ((d)) in a single-step, GD requires 2-steps for the first and p steps for the latter [27].

4. Computational phase transitions ($\alpha_c > 0$)

What happens when the trivial subspace $T^\star$ from is empty? In this section, we discuss precisely this case, showing that for some link functions g there exists a critical sample complexity threshold $\alpha_c > 0$ above which some directions become learnable by iterating AMP. To contrast with the trivial subspace, we refer to these as the *easy directions*. While the trivial subspace is characterised by the existence of the fixed point $\boldsymbol{M} = 0$, the easy subspace will be determined by its stability, which crucially depends on the sample complexity $\alpha = n/d$. The stability of $\boldsymbol{M} = \boldsymbol{0}$ can be obtained by linearising the state evolution (7) around $\delta \boldsymbol{M}$ with $\|\delta \boldsymbol{M}\| \approx 0$: $F(\boldsymbol{M}) \approx \alpha \mathcal{F}(\delta \boldsymbol{M}) + \mathcal{O}(\|\delta \boldsymbol{M}\|^2)$, where $\mathcal{F}(\delta \boldsymbol{M})$ is a linear operator on the cone $\mathcal{S}_p^+$ of PSD matrices of dimension p.

Lemma 4 *(Stability of the uninformed fixed point). If $\boldsymbol{M} = \boldsymbol{0} \in \mathbb{R}^{p \times p}$ is a fixed point, then it is an* **unstable** *fixed point of eq. (7) if and only if $n > \alpha_c d$, where the critical sample complexity α_c is:*

$$\frac{1}{\alpha_c} = \sup_{\boldsymbol{M} \in \mathcal{S}_p^+} \|\mathcal{F}(\boldsymbol{M})\|_F, \qquad \mathcal{F}(\boldsymbol{M}) := \mathbb{E}_Y \left[\partial_{\boldsymbol{\omega}} \boldsymbol{g}_{\text{out}}(Y, 0, \boldsymbol{I}_p)\, \boldsymbol{M} \partial_{\boldsymbol{\omega}} \boldsymbol{g}_{\text{out}}(Y, 0, \boldsymbol{I}_p)^\top \right], \tag{4}$$

with $\| \cdot \|_F$ the Frobenius norm, and the extremum being achieved at a unique extremizer $\boldsymbol{M}^\star \in \mathcal{S}_p^+$. Moreover, if $\|\mathcal{F}(\boldsymbol{M})\|_F = 0$, then $\boldsymbol{M} = \boldsymbol{0}$ is stable a fixed point for any $n = \Theta(d)$.

Lemma 4 implies that for $\alpha > \alpha_c$ AMP algorithm 1 initialised from small but non-zero $\boldsymbol{M}^0$ will learn some directions in the easy subspace $E^\star$, which can be characterised as the orthogonal to all the hard directions $\mathbf{v} \in \mathbb{R}^p$ such that $\mathbf{v}^\top \partial_{\boldsymbol{\omega}} \boldsymbol{g}_{out}(Y, \boldsymbol{\omega} = 0, \boldsymbol{V} = \boldsymbol{I}_p) \mathbf{v} = 0$.

Surprisingly, we can show that for $\alpha > \alpha_c$ AMP algorithm 1 spans the full easy subspace $E^\star$ starting from an arbitrarily small but extensive initial correlation (see Appendix F.5) As for Lemma 3, the expression for the weak recovery threshold eq. (4) generalises the single-index expression from [12, 43, 46]. Indeed, for $p = 1$, we have $\partial_\omega g_{\text{out}}(y, 0, 1) = \mathbb{E}[\mathrm{He}_2(Z)|Y = y]$, and therefore ${}^1\!/\!{}_{\alpha_c} = \mathbb{E}_Y \left[\mathbb{E}[\mathrm{He}_2(Z)|Y]^2 \right]$, which is exactly eq. (11) in [12]. In other words, as long as the link function g has a non-zero second-order coefficient in the Hermite basis, it is learnable by algorithm 1 with $n > \alpha_c d$ samples. Optimality of AMP then implies that no first-order method can achieve non-vanishing correlation with $\boldsymbol{w}^\star$ when $n < \alpha_c d$ as $d \to \infty$. We now illustrate Theorem 4 in a few examples of interest.

(a) The monomial $g(\boldsymbol{z}) = z_1 \ldots z_p$ with $p > 1$ can always be learned with $\alpha > \alpha_c(p)$ large enough [19]. For instance, we have $\alpha_c(2) \approx 0.5937$, $\alpha_c(3) \approx 3.725$ and $\alpha_c(4) \approx 4.912$. In Appendix G we derive an analytical formula for $\alpha_c(p)$ for arbitrary p.

(b) The embedding of the sparse parity functions: $g(\boldsymbol{z}) = \mathrm{s}ign(z_1 z_2 ... z_p)$. As this is invariant under permutations of the indices, Lemma 4 implies the existence of a computational phase transition.In

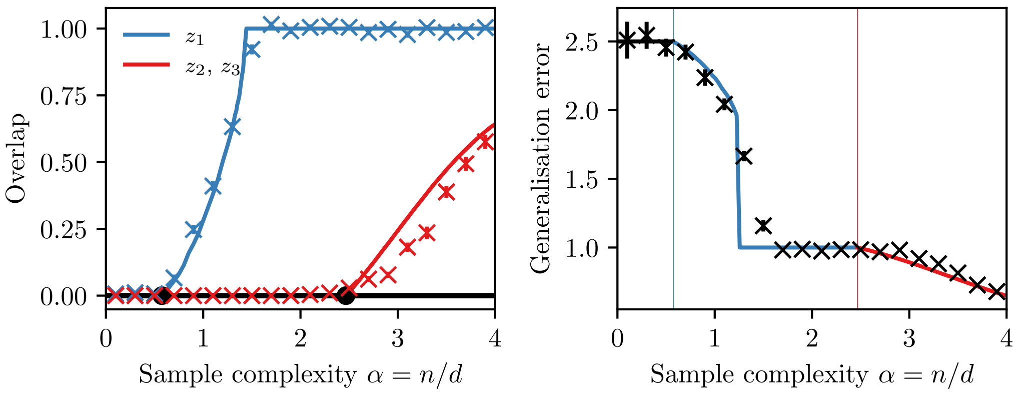

Figure 1: Hierarchical weak learnability for the staircase function $g(z_1, z_2, z_3) = z_1^2 + sign(z_1 z_2 z_3)$. (**Left**): Overlaps with the first direction $|M_{11}|$ (blue), and with the second and third one $1/2(M_{22} + M_{33})$ (red) as a function of the sample complexity $\alpha = n/d$, with solid lines denoting state evolution curves eq. (7), and crosses/dots finite-size runs of AMP algorithm 1 with $d = 500$ and averaged over 72 seeds. All other overlaps are zero (black). The two black dots indicate the critical thresholds at $\alpha_1 \approx 0.575$ and $\alpha_2 = \pi^2/4$. (**Right**) Corresponding generalization error as a function of the sample complexity. Details on the numerical implementation are discussed in Appendix H.

Appendix G.3 we compute analytically the critical value $\alpha_c(p)$. For $p = 1$ this problem is equivalent to phase retrieval, for which we have $\alpha_c(1) = 1/2$. For $p = 2$ we show in the appendix that $\alpha_c(2) = \pi^2/4$, while $\alpha_c(p) = +\infty$ for $p \geq 3$. Figure 2 illustrates this discussion for the 2-sparse parity problem, comparing theoretical prediction eq. (7) with finite-size runs of algorithm 1.

5. Hierarchical Iterative Denoising

Suppose that the estimator $\hat{\boldsymbol{W}}^t$ has developed an overlap along a subspace belonging to the span of $\boldsymbol{W}^\star$, resulting in a non-zero overlap $\boldsymbol{M}^t \succ 0$. How does this affect learning along the orthogonal complement? The central difference with respect to initialization is that the variable $\boldsymbol{\omega}$ in the linear operator $\mathcal{F}$ defined in 4 becomes non-zero (since it is distributed as $\boldsymbol{\omega} = \sqrt{\boldsymbol{M}}\boldsymbol{\xi}$). Crucially, this changes the span of $\mathcal{F}(\boldsymbol{M})$ and hence the stability condition in Lemma 4. In particular, learning some directions might facilitate learning larger subspaces. We call this *Hierarchical Iterative Denoising*. This phenomenon is reminiscent of the specialisation transition in committee machines [7, 50] and of the staircase phenomenon for one-pass SGD introduced in [1]. We formalise this in Appendix D. A concrete example of a function which displays the hierarchical iterative denoising staircase is a linear combination between *hard* parity function and an *easy* polynomial: $g(z_1, z_2, z_3) = z_1^2 + sign(z_1 z_2 z_3)$ The sign part is a sparse parity with $p = 3$, which cannot be learned with $n = \Theta(d)$ samples, but the quadratic polynomial z_1^2 component in the function allows weak-recovery of the first component i.e. $U = (1, 0, 0)$ as long as $\alpha > 1/2$. Hence, conditionally on U the effective multi-index model becomes $sign(z_2 z_3)$, which as discussed in Section 4 is an *easy* function. Figure 1 illustrates this: first, z_1 is learned at $\alpha_1 \approx 0.575$. Then, for *larger* value when $\alpha > \alpha_2$, all directions are learned (see App. (G.4)). Knowing z_1 makes the *hard* 3-parity an *easy* 2-parity problem. It is easy to construct a multi-index model where AMP will iterate over any number of such plateaus. For instance, consider the model: $g(\boldsymbol{z}) = z_1^2 + sign(z_1 z_2 z_3) + sign(z_3 z_4 z_5) + sign(z_5 z_6 z_7)$ After the first plateau to learn z_1, there will be one for z_2, then for z_3, for z_4 and so on.

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 Po-Ling Loh and Maxim Raginsky, editors, *Proceedings of Thirty Fifth Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pages 4782–4887. PMLR, 02–05 Jul 2022. URL https://proceedings.mlr.press/v178/abbe22a.html.

[2] Emmanuel Abbe, Enric Boix Adserà, and Theodor Misiakiewicz. Sgd learning on neural networks: leap complexity and saddle-to-saddle dynamics. In Gergely Neu and Lorenzo Rosasco, editors, *Proceedings of Thirty Sixth Conference on Learning Theory*, volume 195 of *Proceedings of Machine Learning Research*, pages 2552–2623. PMLR, 12–15 Jul 2023. URL https://proceedings.mlr.press/v195/abbe23a.html.

[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. In Gergely Neu and Lorenzo Rosasco, editors, *Proceedings of Thirty Sixth Conference on Learning Theory*, volume 195 of *Proceedings of Machine Learning Research*, pages 1199–1227. PMLR, 12–15 Jul 2023. URL https://proceedings.mlr.press/v195/arnaboldi23a.html.

[4] 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.

[5] Benjamin Aubin, Antoine Maillard, Jean Barbier, Florent Krzakala, Nicolas Macris, and Lenka Zdeborová. The committee machine: computational to statistical gaps in learning a two-layers neural network*. *Journal of Statistical Mechanics: Theory and Experiment*, 2019(12):124023, dec 2019. doi: 10.1088/1742-5468/ab43d2. URL https://dx.doi.org/10.1088/1742-5468/ab43d2.

[6] Benjamin Aubin, Florent Krzakala, Yue Lu, and Lenka Zdeborová. Generalization error in high-dimensional perceptrons: Approaching bayes error with convex optimization. *Advances in Neural Information Processing Systems*, 33:12199–12210, 2020.

[7] Benjamin Aubin, Bruno Loureiro, Antoine Baker, Florent Krzakala, and Lenka Zdeborová. Exact asymptotics for phase retrieval and compressed sensing with random generative priors. In *Mathematical and Scientific Machine Learning*, pages 55–73. PMLR, 2020.

[8] Jimmy Ba, Murat A Erdogdu, Taiji Suzuki, Zhichao Wang, Denny Wu, and Greg Yang. High-dimensional asymptotics of feature learning: How one gradient step improves the representation. *Advances in Neural Information Processing Systems*, 35:37932–37946, 2022.

[9] Dmitry Babichev and Francis Bach. Slice inverse regression with score functions. *Electronic Journal of Statistics*, 12(1):1507 – 1543, 2018. doi: 10.1214/18-EJS1428. URL https://doi.org/10.1214/18-EJS1428.

[10] Afonso S Bandeira, Amelia Perry, and Alexander S Wein. Notes on computational-to-statistical gaps: predictions using statistical physics. *Portugaliae mathematica*, 75(2):159–186, 2018.

[11] Afonso S Bandeira, Ahmed El Alaoui, Samuel Hopkins, Tselil Schramm, Alexander S Wein, and Ilias Zadik. The franz-parisi criterion and computational trade-offs in high dimensional statistics. *Advances in Neural Information Processing Systems*, 35:33831–33844, 2022.

[12] 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. URL https://www.pnas.org/doi/abs/10.1073/pnas.1802705116.

[13] Mohsen Bayati and Andrea Montanari. The dynamics of message passing on dense graphs, with applications to compressed sensing. *IEEE Transactions on Information Theory*, 57(2): 764–785, 2011. doi: 10.1109/TIT.2010.2094817.

[14] Raphaël Berthier, Andrea Montanari, and Kangjie Zhou. Learning time-scales in two-layers neural networks, 2024.

[15] Alberto Bietti, Joan Bruna, and Loucas Pillaud-Vivien. On learning gaussian multi-index models with gradient flow, 2023.

[16] Avrim Blum, Adam Kalai, and Hal Wasserman. Noise-tolerant learning, the parity problem, and the statistical query model. *J. ACM*, 50(4):506–519, jul 2003. ISSN 0004-5411. doi: 10.1145/792538.792543. URL https://doi.org/10.1145/792538.792543.

[17] David R. Brillinger. *A Generalized Linear Model With "Gaussian" Regressor Variables*, pages 589–606. Springer New York, New York, NY, 1982. ISBN 978-1-4614-1344-8. doi: 10.1007/978-1-4614-1344-8_34. URL https://doi.org/10.1007/978-1-4614-1344-8_34.

[18] Michael Celentano, Andrea Montanari, and Yuchen Wu. The estimation error of general first order methods. In Jacob Abernethy and Shivani Agarwal, editors, *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pages 1078–1141. PMLR, 09–12 Jul 2020. URL https://proceedings.mlr.press/v125/celentano20a.html.

[19] Sitan Chen and Raghu Meka. Learning polynomials in few relevant dimensions. In Jacob Abernethy and Shivani Agarwal, editors, *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pages 1161–1227. PMLR, 09–12 Jul 2020. URL https://proceedings.mlr.press/v125/chen20a.html.

[20] Sitan Chen, Aravind Gollakota, Adam Klivans, and Raghu Meka. Hardness of noise-free learning for two-hidden-layer neural networks. *Advances in Neural Information Processing Systems*, 35:10709–10724, 2022.

[21] Elizabeth Collins-Woodfin, Courtney Paquette, Elliot Paquette, and Inbar Seroussi. Hitting the high-dimensional notes: An ode for sgd learning dynamics on glms and multi-index models, 2023.

[22] Arnak S. Dalalyan, Anatoly Juditsky, and Vladimir Spokoiny. A new algorithm for estimating the effective dimension-reduction subspace. *Journal of Machine Learning Research*, 9(53): 1647–1678, 2008. URL http://jmlr.org/papers/v9/dalalyan08a.html.

[23] Alex Damian, Jason D. Lee, and Mahdi Soltanolkotabi. Neural networks can learn representations with gradient descent, 2022.

[24] Alex Damian, Eshaan Nichani, Rong Ge, and Jason D Lee. Smoothing the landscape boosts the signal for sgd: Optimal sample complexity for learning single index models. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 752–784. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/02763667a5761ff92bb15d8751bcd223-Paper-Conference.pdf.

[25] Alex Damian, Eshaan Nichani, Rong Ge, and Jason D Lee. Smoothing the landscape boosts the signal for sgd: Optimal sample complexity for learning single index models. *Advances in Neural Information Processing Systems*, 36, 2024.

[26] Alex Damian, Loucas Pillaud-Vivien, Jason D. Lee, and Joan Bruna. Computational-statistical gaps in gaussian single-index models, 2024.

[27] Yatin Dandi, Emanuele Troiani, Luca Arnaboldi, Luca Pesce, Lenka Zdeborová, and Florent Krzakala. The benefits of reusing batches for gradient descent in two-layer networks: Breaking the curse of information and leap exponents, 2024.

[28] Yash Deshpande and Andrea Montanari. Finding hidden cliques of size n/e n/e in nearly linear time. *Foundations of Computational Mathematics*, 15:1069–1128, 2015.

[29] Ilias Diakonikolas, Daniel M. Kane, Vasilis Kontonis, and Nikos Zarifis. Algorithms and sq lower bounds for pac learning one-hidden-layer relu networks. In Jacob Abernethy and Shivani Agarwal, editors, *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pages 1514–1539. PMLR, 09–12 Jul 2020. URL https://proceedings.mlr.press/v125/diakonikolas20d.html.

[30] David L. Donoho, Arian Maleki, and Andrea Montanari. Message-passing algorithms for compressed sensing. *Proceedings of the National Academy of Sciences*, 106(45):18914–18919, 2009. doi: 10.1073/pnas.0909892106. URL https://www.pnas.org/doi/abs/10.1073/pnas.0909892106.

[31] Massimo Fornasier, Karin Schnass, and Jan Vybiral. Learning functions of few arbitrary linear parameters in high dimensions. *Foundations of Computational Mathematics*, 12(2): 229–262, Apr 2012. ISSN 1615-3383. doi: 10.1007/s10208-012-9115-y. URL https://doi.org/10.1007/s10208-012-9115-y.

[32] Jerome H. Friedman and Werner Stuetzle. Projection pursuit regression. *Journal of the American Statistical Association*, 76(376):817–823, 1981. ISSN 01621459. URL http://www.jstor.org/stable/2287576.

[33] J.H. Friedman and J.W. Tukey. A projection pursuit algorithm for exploratory data analysis. *IEEE Transactions on Computers*, C-23(9):881–890, 1974. doi: 10.1109/T-C.1974.224051.

[34] Cédric Gerbelot and Raphaël Berthier. Graph-based approximate message passing iterations. *Information and Inference: A Journal of the IMA*, 12(4):2562–2628, 09 2023. ISSN 2049-8772. doi: 10.1093/imaiai/iaad020. URL https://doi.org/10.1093/imaiai/iaad020.

[35] Surbhi Goel, Aravind Gollakota, Zhihan Jin, Sushrut Karmalkar, and Adam Klivans. Superpolynomial lower bounds for learning one-layer neural networks using gradient descent. In *International Conference on Machine Learning*, pages 3587–3596. PMLR, 2020.

[36] Yoshiyuki Kabashima. Inference from correlated patterns: a unified theory for perceptron learning and linear vector channels. In *Journal of Physics: Conference Series*, volume 95, page 012001. IOP Publishing, 2008.

[37] Sham M Kakade, Varun Kanade, Ohad Shamir, and Adam Kalai. Efficient learning of generalized linear and single index models with isotonic regression. In J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc., 2011. URL https://proceedings.neurips.cc/paper_files/paper/2011/file/30bb3825e8f631cc6075c0f87bb4978c-Paper.pdf.

[38] Adam Tauman Kalai and Ravi Sastry. The isotron algorithm: High-dimensional isotonic regression. In *COLT 2009 - The 22nd Conference on Learning Theory, Montreal, Quebec, Canada, June 18-21, 2009*, 2009. URL http://www.cs.mcgill.ca/%7Ecolt2009/papers/001.pdf#page=1.

[39] Michael Kearns. Efficient noise-tolerant learning from statistical queries. *J. ACM*, 45(6): 983–1006, nov 1998. ISSN 0004-5411. doi: 10.1145/293347.293351. URL https://doi.org/10.1145/293347.293351.

[40] Mark Grigor'evich Krein and Moisei Aronovich Rutman. Linear operators leaving invariant a cone in a banach space. *Uspekhi Matematicheskikh Nauk*, 3(1):3–95, 1948.

[41] Ker-Chau Li. Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association*, 86(414):316–327, 1991. ISSN 01621459. URL http://www.jstor.org/stable/2290563.

[42] Wangyu Luo, Wael Alghamdi, and Yue M Lu. Optimal spectral initialization for signal recovery with applications to phase retrieval. *IEEE Transactions on Signal Processing*, 67(9):2347–2356, 2019.

[43] 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. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/7ec0dbeee45813422897e04ad8424a5e-Paper.pdf.

[44] Antoine Maillard, Bruno Loureiro, Florent Krzakala, and Lenka Zdeborová. Phase retrieval in high dimensions: Statistical and computational phase transitions. *Advances in Neural Information Processing Systems*, 33:11071–11082, 2020.

[45] Marc Mézard. The space of interactions in neural networks: Gardner's computation with the cavity method. *Journal of Physics A: Mathematical and General*, 22(12):2181, 1989.

[46] Marco Mondelli and Andrea Montanari. Fundamental limits of weak recovery with applications to phase retrieval. *Foundations of Computational Mathematics*, 19(3):703–773, Jun 2019. ISSN 1615-3383. doi: 10.1007/s10208-018-9395-y. URL https://doi.org/10.1007/s10208-018-9395-y.

[47] Behrad Moniri, Donghwan Lee, Hamed Hassani, and Edgar Dobriban. A theory of non-linear feature learning with one gradient step in two-layer neural networks. *arXiv preprint arXiv:2310.07891*, 2023.

[48] Andrea Montanari and Yuchen Wu. Statistically optimal first order algorithms: A proof via orthogonalization. *arXiv preprint arXiv:2201.05101*, 2022.

[49] Sundeep Rangan. Generalized approximate message passing for estimation with random linear mixing. In *2011 IEEE International Symposium on Information Theory Proceedings*, pages 2168–2172. IEEE, 2011.

[50] 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.

[51] David Saad and Sara A Solla. On-line learning in soft committee machines. *Physical Review E*, 52(4):4225, 1995.

[52] Rodrigo Veiga, Ludovic Stephan, Bruno Loureiro, Florent Krzakala, and Lenka Zdeborová. Phase diagram of stochastic gradient descent in high-dimensional two-layer neural networks. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 23244–23255. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/939bb847ebfd14c6e4d3b5705e562054-Paper-Conference.pdf.

[53] Ming Yuan. On the identifiability of additive index models. *Statistica Sinica*, 21(4):1901–1911, 2011. ISSN 10170405, 19968507.

[54] Lenka Zdeborová and Florent Krzakala. Statistical physics of inference: Thresholds and algorithms. *Advances in Physics*, 65(5):453–552, 2016.

Appendix / supplemental material

Appendix A. Further related works

Further related work — With its origins on the classical projection pursuit method [32, 33], there is an extensive literature dedicated to designing and analysing efficient algorithms to train multi-index models models, such as Isotronic Regression for single-index [17, 37, 38] and Sliced Inverse Regression in the multi-index case [9, 22, 31, 53].

The case $p = 1$ has seen a lot of interest recently. In terms of gradient descent, [4] has shown that if the link function g is known, one-pass SGD achieves weak recovery in $n = \Theta(d^{\ell-1})$ steps, where ℓ, known as the *information exponent*, is the first non-zero Hermite coefficient of the link function g. In the non-parametric setting —where g is unknown— [14] has shown that a large-width two-layer neural network trained under one-pass SGD can learn an ”easy” single-index target ($\ell = 1$) in $n = \Theta(d)$ steps. This is to be contrasted with full-batch GD, which can achieve weak recovery in $\Theta(1)$ steps with sample complexity $n = \Theta(d)$ even for particular problems with $\ell > 1$ [27] (which is, indeed, the statistical optimal rate [12, 46]). Similar results have also been proven for abstract computational models. [23] has proven a lower bound $n \geq d^{\max(1,\ell/2)}$ under the Correlational Statistical Query (CSQ) model, comprising algorithms that take queries of the type $\mathbb{E}[y\varphi(\boldsymbol{x})]$. More recently, [26] has proven a similar result for the Statistical Query (SQ) model, which allows for more flexible queries of the type $\mathbb{E}[\varphi(\boldsymbol{x}, y)]$ and hence a lower sample complexity $n \geq d^{\max(1,\kappa/2)}$ with $\kappa \leq \ell$ defining the *generative exponent*. This turns out to be equivalent to the optimal computational weak recovery threshold of [12, 42, 44, 46].

While the situation is less understood, the multicase has also whitenessed a surge of recent interest. In particular, it has been used to understand the behavior of gradiend descent algorithm in neural networks. For instance, [1, 2] showed that a certain class of *staircase* functions can be learned by large-width two-layer networks trained under one-pass SGD with sample complexity $n = \Theta(d)$. This is in stark contrast to (embedded) s-sparse parities, which require $n \geq d^{s-1}$ samples [16]. [1, 2] introduced the *leap exponent*, a direction-wise generalisation of the information exponent for multi-index models, and studied the class of *staircase functions* which can be efficiently learned with one-pass SGD. [15] showed that under a particular gradient flow scheme preserving weight orthogonality (a.k.a. *Stiefel gradient flow*), training follow a saddle-to-saddle dynamics, with the characteristic time required to escape a saddle given by the leap exponent. Interestingly, it was then shown that the limit discovered in these set of works could be bypassed by slightly different algorithm, for instance by smoothing the landscapes [25] or reusing batches multi-time [27]. The latter paper, in particular, showing that gradient descent could learn efficiently a larger class of multi-index models that previously believed to be possible. These findings highlight the need of a strict understanding of the limit learnability of these models.

Our approach to establish the limit of computational learnability is based on the study of the approximate message passing algorithm (AMP). Originating from the cavity method in physics [36, 45], AMP [30], and its generalized version GAMP [49] are powerful iterative algorithm to study these high-dimensional setting. These algorithm are widely believed to be optimal between all polynomial algorithms for such high-dimensional problems [5, 10, 11, 28, 54]. In fact, they are provably optimal among all iterative first-order algorithms [18, 48], a very large class of methods that include gradient descent.

While these algorithm were studied in great detail for the single index models, see e.g. [6, 7, 12], and are at the roots of the spectral method that underline the learnability phase transition in this case [44, 46], much less is known in the multi-index case, with the exception of [5], who showed that particular instances of multi-index models known as *committee machines* can be learned with $n = \Theta(d)$ samples.

Different from our approach are [20, 29, 35], who focused also on peculiar form of the multi-index models (using combination of Relu) and derived worst case bound (in terms of the hardest possible function). Instead, we focus on the typical-case learnability of a given, explicit, multi-index model.

Appendix B. Formal definition of State evolution [5, 34]

Let $(\boldsymbol{x}_i, y_i)_{i\in[n]}$ denote n i.i.d. samples from the multi-index model defined in 1. Consider running Algorithm 1 from initial condition $\hat{\boldsymbol{W}}^0 \in \mathbb{R}^{p\times d}$, such that the initial overlap ${}^1\!/\!{}_d\hat{\boldsymbol{W}}^0\boldsymbol{W}^{\star\top}$ converges in probability to a limit $\boldsymbol{M}^0$ as $d \to \infty$. Denote by $\hat{\boldsymbol{W}}^t$ the resulting estimator at time $0 \leq t \leq T$. Then, in the high-dimensional limit where $n, d \to \infty$ with fixed ratio $\alpha = {}^n\!/\!{}_d$ and $T < O(\log d)$, the asymptotic mean-squared error on the label prediction is given by:

$$\lim_{n,d\to\infty} \mathbb{E}[(y - g(\hat{\boldsymbol{W}}^t\boldsymbol{x}))^2] = \mathbb{E}[(Y^t - g(\boldsymbol{Z}))^2], \qquad 0 \leq t \leq T \tag{5}$$

where the expectation is taken over the following effective estimation process:

$$Y^t = g\left((\boldsymbol{I}_p - \boldsymbol{M}^t)^{1/2}\boldsymbol{Z} + \boldsymbol{M}^{t\,1/2}\boldsymbol{\xi}\right), \qquad \boldsymbol{Z} \sim \mathcal{N}(0, \boldsymbol{I}_p) \tag{6}$$

with $\boldsymbol{\xi} \sim \mathcal{N}(0, \boldsymbol{I}_p)$ independently from $\boldsymbol{Z}$, and $\boldsymbol{M}^t$ is given by iterating the following *state evolution equations* from initial condition $\boldsymbol{M}^0$:

$$\boldsymbol{M}^{t+1} = F(\boldsymbol{M}^t) \coloneqq G\left(\alpha\,\mathbb{E}\left[\boldsymbol{g}_{\text{out}}\left(Y^t, \sqrt{\boldsymbol{M}^t}\boldsymbol{\xi}, \boldsymbol{I}_p - \boldsymbol{M}^t\right)\boldsymbol{g}_{\text{out}}\left(Y^t, \sqrt{\boldsymbol{M}^t}\boldsymbol{\xi}, \boldsymbol{I}_p - \boldsymbol{M}^t\right)^\top\right]\right). \tag{7}$$

where $G(\boldsymbol{X}) = (\boldsymbol{I}_p + \boldsymbol{X})^{-1}\boldsymbol{X}$ and the expectation is taken over the effective process eq. (6). Furthermore, $\boldsymbol{M}^t$ is constrained to lie in $\mathcal{S}_p^+$ due to Nishimori-identity (section G.3 in [5]). The limiting overlap ${}^1\!/\!{}_d\hat{\boldsymbol{W}}^t\boldsymbol{W}^{\star\top}$ converge in probability to $\boldsymbol{M}^t$.

Appendix C. Optimality of AMP

Recall that AMP algorithm 1 is tailored to estimate the posterior marginals:

$$p(\boldsymbol{W}|\boldsymbol{X}, \boldsymbol{y}) \propto \prod_{i=1}^{n} \delta(y_i - g(\boldsymbol{W}\boldsymbol{x}_i)) \prod_{k=1}^{p} \mathcal{N}(\boldsymbol{w}_k|0, \boldsymbol{I}_d). \tag{8}$$

But how efficient is it? A priori, making this comparison requires sampling from the posterior distribution, which is computationally prohibitive in the high-dimensional limit of interest here. The second remarkable property of AMP is that its optimality with respect to the Bayesian posterior eq. (8) can be exactly characterised in the high-dimensional regime.

Lemma 5 (Bayes-optimal correlation) *Let $(x_i, y_i)_{i\in[n]}$ denote n i.i.d. samples from the multi-index model defined in 1. Denote by $\hat{\boldsymbol{W}}_{\mathrm{bo}} = \mathbb{E}[W|X,y] \in \mathbb{R}^{p\times d}$ the mean of the posterior marginals eq. (8). Then, in the high-dimensional asymptotic limit where $n, d \to \infty$ with fixed ratio $\alpha = {}^{n}/_{d}$, the asymptotic correlation between the posterior mean and $\boldsymbol{W}^\star$:*

$$\boldsymbol{M}^\star = \lim_{d\to\infty} \mathbb{E}\left[\frac{1}{d}\hat{\boldsymbol{W}}_{\mathrm{bo}}\boldsymbol{W}^{\star\top}\right] \tag{9}$$

is the solution of the following $\sup\inf$ *problem:*

$$\sup_{\hat{\boldsymbol{M}}\in\mathcal{S}_p^+}\inf_{M\in\mathcal{S}_p^+}\left\{-\frac{1}{2}\operatorname{Tr}\boldsymbol{M}\hat{\boldsymbol{M}} - \frac{1}{2}\log\Big(\boldsymbol{I}_p + \hat{\boldsymbol{M}}\Big) + \frac{1}{2}\hat{\boldsymbol{M}} + \alpha H_Y(\boldsymbol{M})\right\} \tag{10}$$

where $H_Y(\boldsymbol{M}) = \mathbb{E}_{\boldsymbol{\xi}\sim\mathcal{N}(0,\boldsymbol{I}_p)}[H_Y(\boldsymbol{m}|\boldsymbol{\xi})]$, with $H_Y(\boldsymbol{M}|\boldsymbol{\xi})$ the the conditional entropy of the effective p-dimensional estimation problem eq. (6).

Note that the state evolution eq. (7) is closely related to the $\sup\inf$ problem in eq. (10). Indeed, remarking that the update function F in eq. (7) is precisely the gradient of the entropy H_Y in eq. (6), one can show that state evolution is equivalent to gradient descent in the objective defined by eq. (10) [5]. This non-trivial fact implies that whenever eq. (10) has a single minima, AMP optimally estimates the posterior marginals.

Finally, note that by construction $\hat{\boldsymbol{W}}_{\mathrm{bo}}$ is the optimal estimator of $\boldsymbol{W}^\star$ given the data (X, y) (in the MMSE sense). Therefore, the rank of $M^\star$ defines the dimension of the statistically optimal subspace reconstruction at sample complexity $\alpha := {}^{n}/_{d}$. The fact that AMP follows state evolution is proven for such problems in [34]

Appendix D. Formalisation of Hierarchical Iterative Denoising

Here we formalise what we mean by Hierarchical Iterative Denoising

Definition 6 *Let $U \in \mathbb{R}^p$ be a subspace of dimensions k. We define $H_E^\star(U)$ to be the subspace spanned by $\boldsymbol{v} \in U^\perp$ such that:*

$$\boldsymbol{v}^\top \boldsymbol{g}_{\mathrm{out}}(Y, \sqrt{\boldsymbol{M}}_U\boldsymbol{\xi}, \boldsymbol{I} - \sqrt{\boldsymbol{M}}_U) = 0, \tag{11}$$

almost surely over $\boldsymbol{\xi} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_p)$ and Y for any $\boldsymbol{M}_U \in \mathcal{S}_p^+$ such that $\operatorname{span}(\boldsymbol{M}_U) = U$*. We define the "trivially-denoising-coupled" subspace $T_U^\star$ for U as the orthogonal complement of $H_T^\star(U)$.*

Analogously, let $H_E^\star(U)$ be the subspace spanned by directions $\boldsymbol{v} \in U^\perp$ such that:

$$\boldsymbol{v}^\top \partial_{\boldsymbol{\omega}} \boldsymbol{g}_{\mathrm{out}}(Y, \sqrt{\boldsymbol{M}_U}\boldsymbol{\xi}, \boldsymbol{I} - \sqrt{\boldsymbol{M}}_U)\boldsymbol{v} = 0, \tag{12}$$

almost surely over bxi and Y for any $\boldsymbol{M}_U \in \mathcal{S}_p^+$ such that $\operatorname{span}(\boldsymbol{M}_U) = U$.

When $\boldsymbol{M}_U$ is additionally a fixed point of $\mathcal{F}_{\boldsymbol{M}}$, we can linearise $\mathcal{F}_{\boldsymbol{M}}$ along the orthogonal complement of U. We define the easy-denoising-coupled subspace $E_U^\star$ for U as the orthogonal complement of $H_E^\star(U)$. Next, suppose that $\boldsymbol{M}_U \in \mathcal{S}_p^+$ with $\operatorname{span}(\boldsymbol{M}_U) = U$ *is further a fixed-point of $\mathcal{F}_{\boldsymbol{M}}$. Let $\mathcal{F}_{\boldsymbol{M}_U}$ denote the linearization of $F(\boldsymbol{M})$ along the orthogonal complement $U^\perp$ at $\boldsymbol{M} = \boldsymbol{M}_U$. We define the iterative denoising critical threshold $\alpha_{hid}(\boldsymbol{M}_U)$ at $\boldsymbol{M} = \boldsymbol{M}_U$ as:*

$$\frac{1}{\alpha_{hid}(\boldsymbol{M}_U)} = \sup_{\boldsymbol{M}^\perp\in U^\perp} \|\mathcal{F}_{\boldsymbol{M}_U}(\boldsymbol{M})\|_F, \tag{13}$$

We denote by $\boldsymbol{M}_U^\star$ the extremiser achieving the above supremum.

The definitions above generalise the notions of *trivial* and *easy* subspaces conditionally on a subspace U that has been previously learned, and characterise the directions whose recovery is enabled upon learning the subspace U. Concretely, upon developing an initial overlap along U, the directions in $T_U^\star, E_U^\star$ can be recovered analogous to the recovery of $T^\star, E^\star$ starting from random initialization. We formalise this below:

Theorem 7 *Let $U \in \mathbb{R}^p$ be a fixed subspace. Suppose the AMP algorithm 1 is initialised such that $\boldsymbol{M}_d^0 = \boldsymbol{M}_U + \epsilon \boldsymbol{A}$, where $\|\boldsymbol{A}\| = 1$ and $\boldsymbol{M}_U$ is a fixed point of $F(\boldsymbol{M})$ in eq. (7) with $\mathrm{span}(\boldsymbol{M}_U) = U$. If $\alpha \geq \alpha_{hid}(\boldsymbol{M}_U)$, $\exists \delta > 0$ such that for any sufficiently small ϵ ,$\boldsymbol{M}^{(t)} \succ \delta \boldsymbol{M}_U^*$ for some $t = \mathcal{O}(\log {}^1\!/_\epsilon)$, where $\boldsymbol{M}_U^*$ is defined as in Definition 6. For $\alpha < \alpha_{hid}(\boldsymbol{M}_U)$ however, there exists an ϵ' such that for $\epsilon < \epsilon'$, $\left\|\boldsymbol{M}_\perp^t\right\| = 0$ as $t \to \infty$, where $\boldsymbol{M}_\perp^t$ denotes the projection of $\boldsymbol{M}^t$ orthogonal to U. Furthermore, suppose that $\boldsymbol{A}$ is full-rank, then there exists an $\alpha > 0$ and a $\delta > 0$ such that is $\boldsymbol{M}_d^t \succ \delta \boldsymbol{M}_{E_U^\star}$, where $\boldsymbol{M}_{E_U^\star} \in \mathcal{S}_p^+$ spans $E_U^\star$*

When $E_U^\star$ is non-empty for some subspace $U \subseteq \mathbb{R}^p$, we say that the target $y = g(\boldsymbol{W}\boldsymbol{x})$ allows learning through *Hierarchical Iterative Denoising*.

Appendix E. Relation with SQ learning

Lemma 3.1 can be related to computational models based on queries, such as SQ learning [39]: the denoiser $\boldsymbol{g}_{out}$ can indeed be interpreted as a non-linear transformation on the labels $y \mapsto \boldsymbol{g}_{\text{out}}(y, \mathbf{0}, \boldsymbol{I}_p)$. From this perspective, the statement on the condition for the existence of a non-empty trivial subspace theorem 3 translates to the following condition:

$$\mathbb{E}\left[\boldsymbol{g}_{\text{out}}(y, \mathbf{0}, \boldsymbol{I}_p)\langle \mathbf{v}^\top \boldsymbol{W}^\star, \mathbf{x}\rangle\right] = \mathbb{E}\left[\mathbb{E}[(\langle \mathbf{v}^\top W^\star, \mathbf{x}\rangle | Y = y])^2\right] \neq 0, \tag{14}$$

where $\boldsymbol{v} \in T^\star$. Indeed, this can be interpreted as a statistical query of the type $\mathbb{E}[\varphi(y)\psi(x)]$ with label pre-processing $\varphi = \boldsymbol{g}_{out}$. The fact that this linear correlation in the transformed labels is non-vanishing implies that one can weakly recover $\mathbf{v}$ through a tailored spectral method [46]. In fact the denoiser $\boldsymbol{g}_{\text{out}}$ is the optimal such transformation in the sense that when $\boldsymbol{g}_{\text{out}}$ fails to obtain a linear correlation along $\mathbf{v}$, i.e when $\mathbf{v} \in H^\star$, then no transformation can:

Lemma 8 *For any $\mathbf{u} \in H^\star$ and any measurable transformation $\mathcal{T} : \mathbb{R} \to \mathbb{R}$:*

$$\mathbb{E}\left[\mathcal{T}(y)\langle \mathbf{u}^\top \boldsymbol{W}^\star, \mathbf{x}\rangle\right] = 0 \tag{15}$$

Note that this is a non-asymptotic statement about the denoising function $\boldsymbol{g}_{\text{out}}$, and the expectation is with respect to the distribution of the labels y (not the effective problem).

Proof The above is a direct consequence of the Tower law of expectation. Specifically, we have:

$$\mathbb{E}\left[\mathcal{T}(y)\langle \mathbf{v}^\top W^\star, \mathbf{x}\rangle\right] = \mathbb{E}\left[y\right] \mathcal{T}(y) \mathbb{E}\left[\boldsymbol{x}\right] \langle \mathbf{v}^\top W^\star, \mathbf{x}\rangle | y.$$

The statement then follows by noting that $\boldsymbol{g}_{out}(y, \mathbf{0}, \boldsymbol{I}_p)^\top W^\star \mathbf{v} = \mathbb{E}\left[\boldsymbol{x}\right] \langle \mathbf{v}^\top W^\star, \mathbf{x}\rangle | y$. ■

Similarly, we can invoke optimality of AMP to translate our result on the optimal denoiser $y \mapsto \boldsymbol{g}_{\text{out}}(y, \mathbf{0}, \boldsymbol{I}_p)$ to a statement for queries on general label pre-processing transformations:

Lemma 9 *For any $\mathbf{v}^\star \in H_E^\star$ and any measurable transformation $\mathcal{T} : \mathbb{R} \to \mathbb{R}$:*

$$\mathbb{E}\left[\mathcal{T}(y)\mathrm{H}e_2(\langle \mathbf{v}^\star, \mathbf{x}\rangle)\right] = 0 \tag{16}$$

Note that this is a non-asymptotic statement about the denoising function $\boldsymbol{g}_{\mathrm{out}}$, and the expectation is with respect to the distribution of the labels y (not the effective problem from eq. (6)).

For an AMP-hard function (that is, when $\alpha_c = \infty$), this implies that, even after applying any transformation to the output (as allowed in SQ) an algorithm like SGD would require at least $O(d^2)$ data, or $O(d^{3/2})$ if we allow smoothing as in [25].

Appendix F. Proofs of the main results

F.1. Linear Approximation

Throughout, we assume that $\boldsymbol{g}_{out} : \mathbb{R}^{k^2+k+1} \to \mathbb{R}^k$ is in $\mathcal{C}^2$.

Our analysis relies on the following result:

Lemma 10 *Let $F(\boldsymbol{M})$ be as defined in Lemma B*

$$F(\boldsymbol{M}) \approx \alpha\mathcal{F}(\delta\boldsymbol{M}) + \mathcal{O}(\alpha\|\delta\boldsymbol{M}\|_F^2), \tag{17}$$

where $\|\|_F$ denotes the Frobenius norm and $\mathcal{F}(\delta\boldsymbol{M})$ is a linear operator on the cone $\mathcal{S}_p^+$ of PSD matrices of dimension p:

$$\mathcal{F}(\boldsymbol{M}) := \mathbb{E}_y\left[\partial_{\boldsymbol{\omega}}\boldsymbol{g}_{\mathrm{out}}(y, 0, \boldsymbol{I}_p - \boldsymbol{M})\,\boldsymbol{M}\partial_{\boldsymbol{\omega}}\boldsymbol{g}_{\mathrm{out}}(y, 0, \boldsymbol{I}_p - \boldsymbol{M})^\top\right]\,, \tag{18}$$

We proceed through an-entry-wise expansion of each term inside the expection in $F(\boldsymbol{M})$ around $\boldsymbol{M} = 0$. Since $\boldsymbol{M}_F \leq p$, the first two derivatives of $\boldsymbol{g}_{\mathrm{out}}$ are uniformly bounded in $\boldsymbol{g}_{out}$ for any fixed $Y, \boldsymbol{\xi}$. Therefore, applying the multivariate Taylor expansion to $\boldsymbol{g}_{out}(Y^t, \Delta_1\boldsymbol{\xi}, I + \Delta_2)$, with $\Delta_1 = \sqrt{M}$ and $\Delta_2 =, \boldsymbol{I}_p - \boldsymbol{M}$ yields:

$$\begin{aligned}
&\boldsymbol{g}_{out}\Big(Y^t, \sqrt{\boldsymbol{M}^t}\boldsymbol{\xi}, \boldsymbol{I}_p - \boldsymbol{M}^t\Big)\,\boldsymbol{g}_{out}\Big(Y^t, \sqrt{\boldsymbol{M}^t}\boldsymbol{\xi}, \boldsymbol{I}_p - \boldsymbol{M}^t\Big)^\top \\
&= \boldsymbol{g}_{out}\big(Y^t, 0, \boldsymbol{I}_p - \boldsymbol{M}^t\big)\,\boldsymbol{g}_{out}\big(Y^t, 0, \boldsymbol{I}_p - \boldsymbol{M}^t\big)^\top + \partial_{\boldsymbol{\omega}}\boldsymbol{g}_{out}\big(Y^t, 0, \boldsymbol{I}_p\big)\,\sqrt{M}\boldsymbol{\xi}\boldsymbol{\xi}^\top\sqrt{M}^\top\partial_{\boldsymbol{\omega}}\boldsymbol{g}_{out}\big(Y^t, 0, \boldsymbol{I}_p\big)^\top \\
&+ \langle\partial_{\boldsymbol{V}}\boldsymbol{g}_{out}\big(Y^t, 0, \boldsymbol{I}_p\big)\,M\rangle\boldsymbol{g}_{out}\big(Y^t, 0, \boldsymbol{I}_p\big)^\top + \boldsymbol{g}_{out}\big(Y^t, 0, \boldsymbol{I}_p\big)\,\langle M, \partial_{\boldsymbol{V}}\boldsymbol{g}_{out}\big(Y^t, 0, \boldsymbol{I}_p\big)\rangle + \alpha C(\boldsymbol{\xi}, y)\mathcal{O}(\|M\|_F^2),
\end{aligned}$$

where $C(\boldsymbol{\xi}, Y)$ is an integrable function in $\boldsymbol{\xi}, Y$. Since $T^\star$ is empty, $\boldsymbol{g}_{out}\big(Y^t, 0, \boldsymbol{I}_p\big)$ vanishes almost surely over y. Therefore, using dominated-convergence theorem, we obtain:

$$\begin{aligned}
F(\boldsymbol{M}) &= \alpha\mathbb{E}\left[\boldsymbol{g}_{out}\Big(Y^t, \sqrt{\boldsymbol{M}^t}\boldsymbol{\xi}, \boldsymbol{I}_p - \boldsymbol{M}^t\Big)\,\boldsymbol{g}_{out}\Big(Y^t, \sqrt{\boldsymbol{M}^t}\boldsymbol{\xi}, \boldsymbol{I}_p - \boldsymbol{M}^t\Big)^\top\right] \\
&= \alpha\mathbb{E}\left[\partial_{\boldsymbol{\omega}}\boldsymbol{g}_{out}\big(Y^t, 0, \boldsymbol{I}_p\big)\,\sqrt{M}\boldsymbol{\xi}\boldsymbol{\xi}^\top\sqrt{M}^\top\partial_{\boldsymbol{\omega}}\boldsymbol{g}_{out}\big(Y^t, 0, \boldsymbol{I}_p\big)^\top\right] + \alpha\mathcal{O}(\|M\|_F^2)
\end{aligned}$$

Since at $\boldsymbol{M}^t = 0$, $\boldsymbol{\xi}$ and Y are independent, the above simplifies to:

$$F(\boldsymbol{M}) = \alpha\,\mathbb{E}_y\left[\partial_{\boldsymbol{\omega}}\boldsymbol{g}_{\mathrm{out}}(y, 0, \boldsymbol{I}_p - \boldsymbol{M})\,\boldsymbol{M}\partial_{\boldsymbol{\omega}}\boldsymbol{g}_{\mathrm{out}}(y, 0, \boldsymbol{I}_p - \boldsymbol{M})^\top\right] + \alpha\mathcal{O}(\|M\|_F^2)$$

F.2. Proof of Lemma 4

Suppose that $M \in \mathcal{S}^+$, we have, for any $\mathbf{v} \in \mathbb{R}^p$:

$$\mathbf{v}^\top \mathcal{F}(\boldsymbol{M})\mathbf{v} = \alpha\, \mathbb{E}_y \left[\mathbf{v}^\top \partial_{\boldsymbol{\omega}} \boldsymbol{g}_{\text{out}}(y, 0, \boldsymbol{I}_p - \boldsymbol{M})\, \boldsymbol{M} \partial_{\boldsymbol{\omega}} \boldsymbol{g}_{\text{out}}(y, 0, \boldsymbol{I}_p - \boldsymbol{M})\mathbf{v}\right], \tag{19}$$

since $\boldsymbol{M} \in \mathcal{S}^+$, each term inside the expectation is non-negative. Therefore:

$$\mathbf{v}^\top \mathcal{F}(\boldsymbol{M})\mathbf{v} \geq 0 \tag{20}$$

Thus, $\mathcal{F}(M)$ is a cone-preserving linear map. The generalised Perron-Frobenius theorem for cone-preserving maps [40] then implies that the operator $\mathcal{F}(\boldsymbol{M})$ admits a unique eigenvector $M^* \in \mathcal{S}_p^+$ corresponding to the largest eigenvalue $\lambda_{\mathcal{F}}$ such that for any $M \in \mathcal{S}_p^+ \perp M^*$:

$$\mathcal{F}(\boldsymbol{M}) < \lambda_{\mathcal{F}} \|\boldsymbol{M}\|_F. \tag{21}$$

Furthermore, all other eigenvalues are strictly smaller than $\lambda_{\mathcal{F}}$. Subsequently, Lemma 10 implies that $F(\boldsymbol{M})$ is stable at $\boldsymbol{M} = 0$ if and only if $\alpha \leq \frac{1}{\lambda_{\mathcal{F}}} = \alpha_c$.

F.3. Proof of Theorem 11

Lemma B allows us to map the behavior of the variable $\boldsymbol{M}^t$ in state-evolution to high-probability statements for the limiting overlaps Suppose that $A \in \mathcal{S}_p^+$ is a full rank matrix as in Theorem 11. Applying Lemmas 10 and 8, we obtain that for $\boldsymbol{M}^0 = \epsilon A$:

$$\left\|\boldsymbol{M}^{t+1}\right\|_F \leq \alpha\lambda_{\mathcal{F}} \left\|\boldsymbol{M}^t\right\|_F + \alpha C \left\|\boldsymbol{M}^t\right\|_F^2, \tag{22}$$

for some constant C. Now, suppose that $\alpha < \frac{1}{\lambda_F}$ and $\epsilon < \frac{1}{C}(1-\alpha\lambda_{\mathcal{F}}-\kappa)$, for some $0 < \kappa < 1-\alpha\lambda_{\mathcal{F}}$ then inductively, we obtain:

$$\left\|\boldsymbol{M}^{t+1}\right\|_F < (1-\kappa)\left\|\boldsymbol{M}^t\right\|_F, \tag{23}$$

for all $t > 0$. Implying that $\left\|\boldsymbol{M}^t\right\|_F \to 0$.

Now, suppose that $\alpha > \alpha_c$ or equivalently $\alpha\lambda_F > 1$. Since A is full-rank admits a decomposition:

$$\operatorname{tr}(A, \boldsymbol{M}^\star) = \theta \tag{24}$$

for some $\theta > 0$. From Lemma 4 and 10, we obtain:

$$\operatorname{tr}(\boldsymbol{M}^{t+1}, \boldsymbol{M}^\star) \geq \alpha\lambda_{\mathcal{F}} \operatorname{tr}(\boldsymbol{M}^t, \boldsymbol{M}^\star) - \alpha C \left\|\boldsymbol{M}^t\right\|_F^2. \tag{25}$$

For $\left\|\boldsymbol{M}^t\right\| \leq \frac{\theta}{C}(\alpha\lambda_{\mathcal{F}} - 1 - \kappa)$ for some $\kappa > 0$, we inductively obtain:

$$\operatorname{tr}(\boldsymbol{M}^{t+1}, \boldsymbol{M}^\star) \geq (1+\kappa) \operatorname{tr}(\boldsymbol{M}^t, \boldsymbol{M}^\star), \tag{26}$$

implying that $\operatorname{tr}(\boldsymbol{M}^{t+1}, \boldsymbol{M}^\star)$ grows as $\omega(e^{\kappa t})$ for sufficiently small ϵ.

Finally, it remains to show that for large enough α, $\boldsymbol{M}^t$ spans $E^\star$.

Notice that for any $\mathbf{v} \in E^\star$:

$$\mathbf{v}^\top \mathcal{F}(\mathbf{v}\mathbf{v}^\top) v = \mathbb{E}_y \left[(\mathbf{v}^\top \partial_{\boldsymbol{\omega}} \boldsymbol{g}_{\text{out}}(y, 0, \boldsymbol{I}_p)\mathbf{v})^2\right] > 0, \tag{27}$$

while for any $A \in \mathcal{S}_p^+$:

$$\mathbf{v}^\top \mathcal{F}(A) v \geq 0, \tag{28}$$

since $\mathcal{F}(A) \in \mathcal{S}_p^+$. Define:

$$\nu_{\mathcal{F}} = \inf \mathbf{v} \in \mathbb{E}, \|v\| = 1 \mathbf{v}^\top \mathcal{F}(\mathbf{v}\mathbf{v}^\top) v. \tag{29}$$

Since $\mathbf{v}$ lies in a compact set, $\nu_{\mathcal{F}} > 0$. Let $A^\star$ denote the projection of A along $E^\star$ and consider its eigendecomposition $A^\star$:

$$A^\star = \sum_{i=1}^{p} \lambda_i \mathbf{v}_i \mathbf{v}_i^\top \tag{30}$$

We

$$\mathbf{v}_i^\top (\boldsymbol{M}^{t+1}) \mathbf{v}_i \geq \alpha \nu_{\mathcal{F}} \mathbf{v}_i^\top (\boldsymbol{M}^t) \mathbf{v}_i - \alpha C \left\| \boldsymbol{M}^t \right\|_F^2. \tag{31}$$

Therefore, for $\alpha \geq \frac{1}{\nu_{\mathcal{F}}}$ and small enough ϵ, $\boldsymbol{M}^t$ expands linearly along each $\mathbf{v}_i$

F.4. Proof of Theorem 7

By assumption, M_U is a fixed point of $F(M)$. Therefore, Equation 7 implies that $\boldsymbol{g}_{out}(y, \boldsymbol{\omega}, \boldsymbol{I} - \sqrt{M}_U)$ almost surely lies in U and thus $T_U^\star$ is empty. The proof of Theorem 7 then follows that of 11 by considering the following linearized operator along $U^\perp$

$$\mathcal{F}_{\boldsymbol{M}_U}(\boldsymbol{M}_\perp) =:= \mathbb{E}_y \left[\partial_{\boldsymbol{\omega}} \boldsymbol{g}_{\mathrm{out}}(y, \sqrt{\boldsymbol{M}_U} \boldsymbol{\xi}, \boldsymbol{I} - \sqrt{M}_U) \boldsymbol{M}_\perp \partial_{\boldsymbol{\omega}} \boldsymbol{g}_{out}(y, \sqrt{\boldsymbol{M}_U} \boldsymbol{\xi}, \boldsymbol{I} - \sqrt{M}_U)^\top \right], \tag{32}$$

where $\boldsymbol{\omega} = \sqrt{M}_U \boldsymbol{\xi}$ with $\boldsymbol{\xi} \sim \mathcal{N}(0, \boldsymbol{I})$. Similar to Lemma 10, the linearization follows by noting that y is independent of $\sqrt{\boldsymbol{M}_\perp} \boldsymbol{\xi}$ and $\sqrt{\boldsymbol{M}_U + \boldsymbol{M}_\perp} = \sqrt{\boldsymbol{M}_U} + \sqrt{\boldsymbol{M}_\perp}$ since $\boldsymbol{M}_\perp \perp \boldsymbol{M}_U$.

F.5. Precise statement on initialisation

Here we describe precisely what are the implications for recovery of easy direction when initialising AMP in the prior at random.

Theorem 11 *Suppose that $T^\star = 0$ and the AMP algorithm 1 is initialised at $\hat{\mathbf{W}}^0$ such that $\boldsymbol{M}_d^0 := {}^1\!/\!_d \hat{\boldsymbol{W}}^0 \boldsymbol{W}^{\star\top} = \epsilon \boldsymbol{A}$, with $\|\boldsymbol{A}\|_F = 1$ and $\boldsymbol{A}$ being full-rank. Then, with high probability as $d \to \infty$, for $\alpha \geq \alpha_c$, $\exists \delta > 0$ such that for any sufficiently small ϵ ,$\boldsymbol{M}_d^t \succ \delta \boldsymbol{M}^\star$ for $t = \mathcal{O}(\log {}^1\!/\!_\epsilon)$, where $\boldsymbol{M}^\star$ is as defined in Lemma 4. For $\alpha < \alpha_c$ however, $\boldsymbol{M}_d^t = 0$ is asympotically stable i.e. there exists an ϵ' such that for $\epsilon < \epsilon'$, $\left\| \boldsymbol{M}^t \right\| \to 0$ as $t \to \infty$. Furthermore, there exists an $\alpha > 0$ and a $\delta > 0$ such that is $\boldsymbol{M}_d^t \succ \delta M_{E^*}$ in $t = \mathcal{O}(\log {}^1\!/\!_\epsilon)$ iterations, where $M_{E^\star} \in \mathcal{S}_p^+$ spans $E^\star$.*

Heuristically, random initialization is equivalent to setting $\epsilon = O({}^1\!/\!\sqrt{d})$ in the above theorem. Therefore, we expect that weak-recovery along $E^\star$ can be achieved in $O(\log d)$ iterations.

Appendix G. A list of examples of the g_{out} function

In this section, we list several problems, their corresponding $\boldsymbol{g}_{\mathrm{out}}$, and analysis of their weak recoverability.

G.1. $g(z_1, z_2, ..., z_p) = \prod_{j=1}^{p} z_j$

For all p, $\boldsymbol{g}_{out}(y, \boldsymbol{\omega}, \boldsymbol{I}_p) = 0$. Otherwise we have two cases We have two cases. If $p = 2$:

$$\mathcal{Z}_{out}(y, \boldsymbol{\omega}, \boldsymbol{I}_p) = \frac{K_0(|y|)}{\pi} \tag{33}$$

and

$$\partial_{\boldsymbol{\omega}} \boldsymbol{g}_{\text{out}}(y, \boldsymbol{\omega}, \boldsymbol{I}_p) = \begin{bmatrix} |y|\frac{K_1(|y|)}{K_0(|y|)} - 1 & y \\ y & |y|\frac{K_1(|y|)}{K_0(|y|)} - 1 \end{bmatrix} \tag{34}$$

where $K_n(y)$ are the modified Bessel functions of the second kind. If $p > 3$ then

$$\mathcal{Z}_{out}(y, \boldsymbol{\omega}, \boldsymbol{I}_p) = \frac{1}{(2\pi)^{p/2}} G_{0,p}^{p,0}\left(\frac{y^2}{2^p} \,\middle|\, \begin{matrix} 0 \\ 0, 0, \ldots, 0 \end{matrix} \right) \tag{35}$$

The matrix $\partial_{\boldsymbol{\omega}} \boldsymbol{g}_{out}(y, \boldsymbol{\omega}, \boldsymbol{I}_p)$ is diagonal, with identical diagonal elements equal to $f(y)$:

$$f(y) = 2G_{0,p}^{p,0}\left(\frac{y^2}{2^p} \,\middle|\, \begin{matrix} 0 \\ 0, 0, \ldots, 0 \end{matrix} \right) / G_{0,p}^{p,0}\left(\frac{y^2}{2^p} \,\middle|\, \begin{matrix} 0 \\ 0, 0, \ldots, 0, 1 \end{matrix} \right) - 1 \tag{36}$$

The alpha critical α_c will thus be

$$\alpha_c = \left[\int_{-\infty}^{\infty} \mathrm{d}y f(y)^2 \mathcal{Z}_{\text{out}}(y) \right]^{-1} \tag{37}$$

G.2. $\frac{1}{p} \sum_{i=1}^{p} z_i^2$

The function is even, so $\boldsymbol{g}_{out}(y, \boldsymbol{\omega}, \boldsymbol{I}_p) = 0$. The trick for the computation is to move to generalised spherical coordinates by choosing $r^2 = z_2^2 + ...z_p^2$. Recall that the area of the unit sphere in $p - 1$ dimensions is

$$\frac{2\pi^{\frac{p-1}{2}}}{\Gamma\left(\frac{p-1}{2}\right)} \tag{38}$$

The matrix $\partial_{\boldsymbol{\omega}} \boldsymbol{g}_{out}(y, \boldsymbol{\omega}, \boldsymbol{I}_p)$ is diagonal because of parity. All the elements on the diagonal are

$$(y-1)^2 \frac{2^{-\frac{p}{2}} e^{-\frac{py}{2}} (py)^{\frac{p}{2}}}{\Gamma\left(\frac{p}{2}\right)} \tag{39}$$

The critical alpha will be

$$\alpha_c = \left[\int \mathrm{d}y (y-1)^2 \frac{2^{-\frac{p}{2}} e^{-\frac{py}{2}} (py)^{\frac{p}{2}}}{\Gamma\left(\frac{p}{2}\right)} \right]^{-1} = \frac{p}{2} \tag{40}$$

Of course this procedure fail at the first step if $p = 1$. On the other hand, one can check that the final result is still correct for all positive integer p.

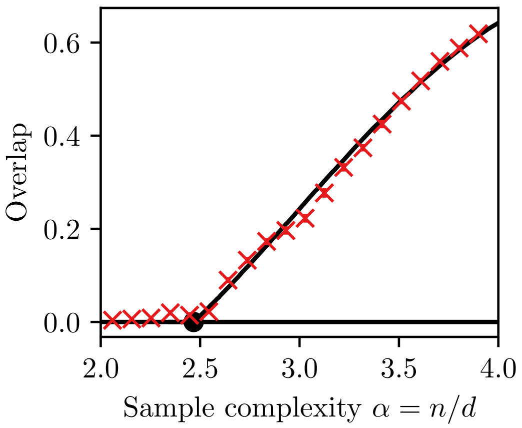

Figure 2: Numerical illustration of the weak learnability phase transition for the 2-sparse parity $g(z_1, z_2) = sign(z_1 z_2)$ that has a phase transition at $\alpha_c(2) = \pi^2/4$. The overlap shows how well the directions z_1 and z_2 are recovered. Given the permutation symmetry in ((b)), we show here and in all the subsequent figures the optimal permutation of the overlap matrix elements reached by AMP. The solid black line is the prediction from the theory. Crosses are averages over 72 runs of AMP algorithm 1 with $d = 500$.

G.3. $g(z_1, \dots, z_p) = \mathrm{sign}(\prod_{j=1}^{p} z_j)$

We define $\boldsymbol{\omega} = (\omega_1, \dots, \omega_p)^\top$ and $v_{ij} := (\boldsymbol{V}^{-1})_{i,j}$. We also introduce the matrix $\boldsymbol{d}(\boldsymbol{V}) \in \mathbb{R}^{(p-1)\times(p-1)}$ such that

$$d_{ij}(\boldsymbol{V}) := \det\begin{pmatrix} v_{i,j} & v_{i,p} \\ v_{j,p} & v_{p,p} \end{pmatrix}, \quad i, j \in 1, \dots, p, \tag{41}$$

the function

$$E_s(\boldsymbol{z} \in \mathbb{R}^{p-1}, y, \boldsymbol{\omega}, \boldsymbol{V}) = 1 - s\,\mathrm{erf}\left(\frac{v_{1,p}(\sum_{j=1}^{p-1} v_{j,p}(z_j - \omega_j) - v_{pp}\omega_p}{\sqrt{2v_{pp}}}\right),$$

the Gaussian probability density

$$\rho(\boldsymbol{z} \in \mathbb{R}^p, \boldsymbol{V}) := \frac{1}{(2\pi)^{p/2}\sqrt{\det(\boldsymbol{V})}} e^{-\frac{1}{2}\boldsymbol{z}^\top \boldsymbol{V}^{-1}\boldsymbol{z}}, \tag{42}$$

and finally

$$I_s(\boldsymbol{z} \in \mathbb{R}^{p-1}, y, \boldsymbol{\omega}, \boldsymbol{V}) := E_s(\boldsymbol{z}, y, \boldsymbol{\omega}, \boldsymbol{V})\rho\left(\begin{pmatrix} z_1 - \omega_2 \\ \dots \\ z_{p-1} - \omega_{p-1} \end{pmatrix}^\top, v_{pp}\boldsymbol{d}(\boldsymbol{V})^{-1}\right).$$

Then, defining $\mathbb{R}_{+1} = [0, +\infty)$ and $\mathbb{R}_{-1} = (-\infty, 0]$, we have that

$$Z_{\mathrm{out}}(y, \boldsymbol{\omega}\boldsymbol{V}) = \sum_{s_1, \dots, s_p = \pm 1} \frac{\delta_{y,(\prod_j s_j)}}{2} \int_{\mathbb{R}_{s_1}} \mathrm{d}z_1 \dots \int_{\mathbb{R}_{s_{p-1}}} \mathrm{d}z_{p-1} I_{s_p}(\boldsymbol{z}, y, \boldsymbol{\omega}, \boldsymbol{V}). \tag{43}$$

The components of $\boldsymbol{g}_{out}(y, \boldsymbol{\omega}, \boldsymbol{V})$ are given by

$$(\boldsymbol{g}_{\text{out}}(y, \boldsymbol{\omega}, \boldsymbol{V}))_k = \sum_{s_1,\dots,s_p=\pm 1} \frac{s_k \delta_{y,(\prod_j s_j)}}{2} \int_{\mathbb{R}_{s_1}} \mathrm{d}z_1 \dots \int_{\mathbb{R}_{s_{p-1}}} \mathrm{d}z_{p-1} I_{s_p}(\boldsymbol{z}, y, \boldsymbol{\omega}, \boldsymbol{V}) \delta(z_k - 0)$$

Stability of the fixed point We can now focus on studying the stability

$$Z_{\text{out}}(y, \mathbf{0}, \boldsymbol{I}) = \frac{1}{(2\pi)^{p/2}} \int \mathrm{d}\boldsymbol{z} e^{-\boldsymbol{z}^\top \boldsymbol{z}/2} \delta_{y,\text{sign}(z_1 \dots z_p)} = \frac{1}{2} \tag{44}$$

In order to compute $\nabla_\omega \boldsymbol{g}_{out}(y, \mathbf{0}, \boldsymbol{I})$, we note at first that

$$\frac{1}{(2\pi)^{p/2}} \int \mathrm{d}\boldsymbol{z} (z_k^2 - 1) e^{-\boldsymbol{z}^\top \boldsymbol{z}/2} \delta_{y,\text{sign}(z_1 \dots z_p)} = 0, \quad \forall k \in \{1, \dots, p\}. \tag{45}$$

For $p = 2$, consider the integral

$$\frac{1}{2\pi} \int \mathrm{d}z_1 \mathrm{d}z_2 \, z_1 z_2 e^{-\frac{z_1^2}{2} - \frac{z_2^2}{2}} \delta_{y,\text{sign}(z_1 \dots z_p)} \tag{46}$$

$$= \frac{2}{2\pi} \left(\delta_{y,1} \int_0^\infty \mathrm{d}z_1 \int_0^\infty \mathrm{d}z_2 \, z_1 z_2 e^{-\frac{z_1^2}{2} - \frac{z_2^2}{2}} + \delta_{y,-1} \int_0^\infty \mathrm{d}z_1 \int_{-\infty}^0 \mathrm{d}z_2 \, z_1 z_2 e^{-\frac{z_1^2}{2} - \frac{z_2^2}{2}} \right) \tag{47}$$

$$= \frac{sign(y)}{\pi} \tag{48}$$

Using (4) we obtain that α_c is the inverse of the largest eigenvalues of

$$\frac{4}{\pi^2} \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \tag{49}$$

So $\alpha_c = \pi^2/4$. We can see that this is consistent with Figure 2. For $p \geq 3$, defining $\mathbb{R}^p_{(\pm)} = \{\boldsymbol{z} \in \mathbb{R}^p | \, \text{sign}(z_1 \dots z_p) = \pm 1\}$, we need to compute integrals of the type

$$\frac{1}{(2\pi)^{p/2}} \int_{\mathbb{R}^p_{(+)}} \mathrm{d}z_1 \dots \mathrm{d}z_p z_1 z_2 e^{-\frac{z_1^2}{2}} \dots e^{-\frac{z_p^2}{2}}$$

$$= \frac{2}{2^{p-2}(2\pi)^{p/2}} \left(\sum_{j=0}^{\lfloor \frac{p-2}{2} \rfloor} \begin{pmatrix} p-2 \\ 2j \end{pmatrix} \int_{\mathbb{R}^2_{(+)}} \mathrm{d}z_1 \mathrm{d}z_2 \, z_1 z_2 e^{-\frac{z_1^2}{2} - \frac{z_2^2}{2}} + \sum_{j=1}^{\lfloor \frac{p-1}{2} \rfloor} \begin{pmatrix} p-2 \\ 2j-1 \end{pmatrix} \int_{\mathbb{R}^2_{(-)}} \mathrm{d}z_1 \mathrm{d}z_2 \, z_1 z_2 e^{-\frac{z_1^2}{2} - \frac{z_2^2}{2}} \right)$$

$$= \frac{2^{3-p}}{(2\pi)^{(p+2)/2}} \sum_{j=0}^{p-2} \begin{pmatrix} p-2 \\ j \end{pmatrix} (-1)^j = 0$$

In the same way it is possible to prove that the integral over $\mathbb{R}^p_{(-)}$ is also vanishing. This implies that $\partial_{\boldsymbol{\omega}} \boldsymbol{g}_{out} = 0$, so using (4), we obtain that $\alpha_c = +\infty$. This model cannot be learned with $n = \mathcal{O}(d)$ samples for $p \geq 3$.

G.4. $g(z_1, \dots, z_p) = z_1^2 + \text{sign}\left(\prod_{j=1}^{p} z_j\right)$

We define $\boldsymbol{\omega} = (\omega_1, \dots, \omega_p)^\top$ and $v_{ij} := (\boldsymbol{V}^{-1})_{i,j}$. We also introduce the matrix $\boldsymbol{d}(\boldsymbol{V}) \in \mathbb{R}^{(p-1)\times(p-1)}$ such that

$$d_{ij}(\boldsymbol{V}) := \det \begin{pmatrix} v_{i,j} & v_{i,p} \\ v_{j,p} & v_{p,p} \end{pmatrix}, \quad i, j \in 1, \dots, p, \tag{50}$$

the function

$$E_{s,s_1,s_p}(\boldsymbol{z} \in \mathbb{R}^{p-1}, y, \boldsymbol{\omega}, \boldsymbol{V}) = 1 - s_p \,\text{erf}\left(\frac{v_{1,p}(s_1\sqrt{y-s} - \omega_1) + \sum_{j=2}^{p-1} v_{j,p}(z_j - \omega_j) - v_{pp}\omega_p}{\sqrt{2v_{pp}}}\right),$$

the Gaussian probability density

$$\rho(\boldsymbol{z} \in \mathbb{R}^p, \boldsymbol{V}) := \frac{1}{(2\pi)^{p/2}\sqrt{\det(\boldsymbol{V})}} e^{-\frac{1}{2}\boldsymbol{z}^\top \boldsymbol{V}^{-1}\boldsymbol{z}}, \tag{51}$$

and finally

$$I_{s,s_1,s_{p-1}}(\boldsymbol{z} \in \mathbb{R}^{p-1}, y, \boldsymbol{\omega}, \boldsymbol{V}) := \frac{E_{s,s_1,s_p}(\boldsymbol{z}, y, \boldsymbol{\omega}, \boldsymbol{V})}{4\sqrt{y-s}} \rho\left(\begin{pmatrix} s_1\sqrt{y-s} - \omega_1 \\ z_2 - \omega_2 \dots \\ z_{p-1} - \omega_{p-1} \end{pmatrix}, v_{pp}\boldsymbol{d}(\boldsymbol{V})^{-1}\right).$$

Then, defining $\mathbb{R}_{+1} = [0, +\infty)$ and $\mathbb{R}_{-1} = (-\infty, 0]$, we have that

$$Z_{\text{out}}(y, \boldsymbol{\omega}\boldsymbol{V}) = \sum_{s,s_1,\dots,s_p=\pm 1} \delta_{s,(\prod_j s_j)} \mathbf{1}_{y>s} \int_{\mathbb{R}_{s_2}} \mathrm{d}z_2 \dots \int_{\mathbb{R}_{s_{p-1}}} \mathrm{d}z_{p-1} I_{s,s_1,s_p}(\boldsymbol{z}, y, \boldsymbol{\omega}, \boldsymbol{V}), \tag{52}$$

where $\mathbf{1}_{x>0}$ is the Heaviside step function. The components of $\boldsymbol{g}_{out}(y, \boldsymbol{\omega}, \boldsymbol{V})$ are given by

$$(\boldsymbol{g}_{out}(y, \boldsymbol{\omega}, \boldsymbol{V}))_1 = \sum_{s,s_1,\dots,s_p=\pm 1} \frac{\delta_{s,(\prod_j s_j)} \mathbf{1}_{y>s}}{Z_{out}} \int_{\mathbb{R}_{s_2}} \mathrm{d}z_2 \dots \int_{\mathbb{R}_{s_{p-1}}} \mathrm{d}z_{p-1} \frac{\partial}{\partial \omega_1} I_{s,s_1,s_p}(\boldsymbol{z}, y, \boldsymbol{\omega}, \boldsymbol{V})$$

$$(\boldsymbol{g}_{out}(y, \boldsymbol{\omega}, \boldsymbol{V}))_{k\neq 1} = \sum_{s,s_1,\dots,s_p=\pm 1} \frac{s_k \delta_{s,(\prod_j s_j)} \mathbf{1}_{y>s}}{Z_{out}} \int_{\mathbb{R}_{s_2}} \mathrm{d}z_2 \dots \int_{\mathbb{R}_{s_{p-1}}} \mathrm{d}z_{p-1} I_{s,s_1,s_p}(\boldsymbol{z}, y, \boldsymbol{\omega}, \boldsymbol{V}) \delta(z_k - 0)$$

Stability of the fixed point The result is dependent on a number of integrals which we can compute analytically. First we have

$$Z_{out}(y, \mathbf{0}, \boldsymbol{I}) = \frac{1}{(2\pi)^{p/2}} \int \mathrm{d}z_1 \dots \mathrm{d}z_p \delta(y - z_1^2 - sign(z_1 \dots z_p)) e^{-\frac{z_1^2}{2} - \dots - \frac{z_p^2}{2}} \tag{53}$$

$$= \frac{1}{2(2\pi)^{p/2}} \sum_{s_1,\dots,s_p=\pm 1} \frac{\mathbf{1}_{y>s}}{\sqrt{y - \prod_j s_j}} \int_{\mathbb{R}_{s_2}} \mathrm{d}z_2 \dots \int_{\mathbb{R}_{s_{p-1}}} \mathrm{d}z_{p-1} e^{-\frac{y - s_1 \dots s_p}{2} - \frac{z_2^2}{2} - \dots - \frac{z_p^2}{2}} \tag{54}$$

$$= \frac{1}{2\sqrt{2\pi}} \sum_{s=\pm 1} \mathbf{1}_{y>s} \frac{e^{-\frac{y-s}{2}}}{\sqrt{y-s}} \tag{55}$$

$$= \frac{e^{-y/2}}{2\sqrt{2\pi e}} \left(\frac{1}{\sqrt{y+1}} + \mathbf{1}_{y>1} \frac{e}{\sqrt{y-1}}\right) \tag{56}$$

From this it is straightforward to see that $\boldsymbol{g}^0_{\text{out}}(y) = 0$. In order to compute $\nabla_\omega \boldsymbol{g}_{\text{out}}(y, 0, 1)$ we need to consider the additional integrals

$$1)\ \frac{1}{(2\pi)^{p/2}} \int \mathrm{d}z_1 \dots \mathrm{d}z_p (z_1^2 - 1) e^{-\frac{z_1^2}{2} - \dots - \frac{z_p^2}{2}} \delta(y - z_1^2 - \mathrm{s}ign(z_1 \dots z_p)) \tag{57}$$

$$= \frac{1}{2\sqrt{2\pi}} \sum_{s=\pm 1} \mathbf{1}_{y>s} \frac{e^{-\frac{y-s}{2}}(y - s - 1)}{\sqrt{y-s}} \tag{58}$$

$$= \frac{e^{-y}}{2\sqrt{2\pi e}} \left(\frac{y}{\sqrt{y+1}} + \mathbf{1}_{y>1} \frac{e(y-2)}{\sqrt{y-1}} \right) \tag{59}$$

$$2)\ \frac{1}{(2\pi)^{p/2}} \int \mathrm{d}z_1 \dots \mathrm{d}z_p (z_{k\neq 1}^2 - 1) e^{-\frac{z_1^2}{2} - \dots - \frac{z_p^2}{2}} \delta(y - z_1^2 - \mathrm{s}ign(z_1 \dots z_p)) = 0 \tag{60}$$

$$3)\ \frac{1}{(2\pi)^{p/2}} \int \mathrm{d}z_1 \dots \mathrm{d}z_p z_j z_k e^{-\frac{z_1^2}{2} - \dots - \frac{z_p^2}{2}} \delta(y - z_1^2 - \mathrm{s}ign(z_1 \dots z_p)) \tag{61}$$

$$\propto \sum_{s, s_j, s_k = \pm 1} \mathbf{1}_{y>s} s_j s_k e^{-\frac{y-s}{2}} = 0 \tag{62}$$

$$\tag{63}$$

This shows that

$$\partial_{\boldsymbol{\omega}} \boldsymbol{g}_{\text{out}}(y) = \begin{bmatrix} C(y) & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \tag{64}$$

Where $C(y)$ is

$$C(y) = \begin{cases} y & -1 < y < 1 \\ y - 2e\frac{\sqrt{y+1}}{e\sqrt{y+1} + \sqrt{y-1}} - 1 & y > 1 \end{cases} \tag{65}$$

The critical alpha is simply found by one last integral

$$\alpha_c = \left[\int \mathrm{d}y C(y)^2 Z_{\text{out}}(y) \right]^{-1} \approx 0.575166 \tag{66}$$

G.5. $g(z_1, \dots, z_p) = \sum_{j=1}^p \mathrm{sign}(z_p)$

We define $\boldsymbol{\omega} = (\omega_1, \dots, \omega_p)^\top$ and $v_{ij} := (\boldsymbol{V}^{-1})_{i,j}$. We also introduce the matrix $\boldsymbol{d}(\boldsymbol{V}) \in \mathbb{R}^{(p-1)\times(p-1)}$ such that

$$d_{ij}(\boldsymbol{V}) := \det \begin{pmatrix} v_{i,j} & v_{i,p} \\ v_{j,p} & v_{p,p} \end{pmatrix}, \quad i, j \in 1, \dots, p, \tag{67}$$

the function

$$E_s(\boldsymbol{z} \in \mathbb{R}^{p-1}, y, \boldsymbol{\omega}, \boldsymbol{V}) = 1 - s \operatorname{erf}\left(\frac{\sum_{j=1}^{p-1} v_{j,p}(z_j - \omega_j) - v_{pp}\omega_p}{\sqrt{2 v_{pp}}} \right), \tag{68}$$

the Gaussian probability density

$$\rho(\boldsymbol{z} \in \mathbb{R}^p, \boldsymbol{V}) := \frac{1}{(2\pi)^{p/2}\sqrt{\det(\boldsymbol{V})}} e^{-\frac{1}{2}\boldsymbol{z}^\top \boldsymbol{V}^{-1} \boldsymbol{z}}, \tag{69}$$

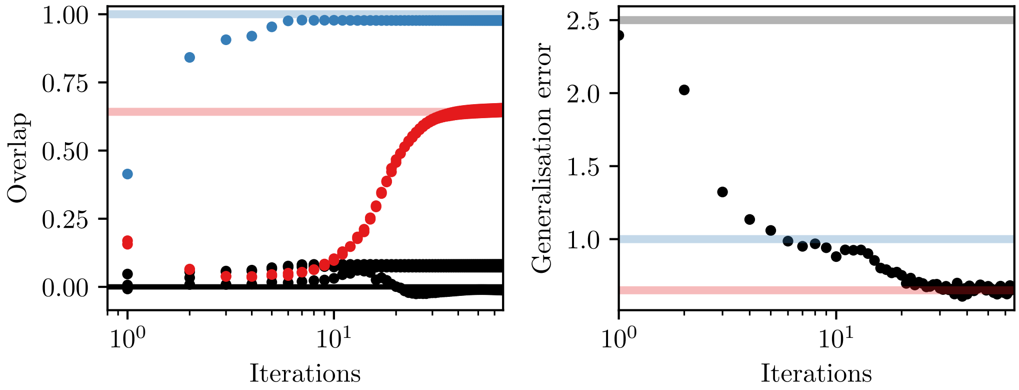

Figure 3: Dynamics for a single finite-size run of AMP with $d = 500$ at $\alpha = 4$ for $g(z_1, z_2, z_3) = z_1^2 + sign(z_1 z_2 z_3)$. (**Left**) Evolution of the overlaps. We display M_{11} in blue, $1/2(M_{22} + M_{33})$ in red, and the off-diagonal overlaps in black. (**Right**) Evolution of the generalisation error.

and finally

$$I_s(\boldsymbol{z} \in \mathbb{R}^{p-1}, y, \boldsymbol{\omega}, \boldsymbol{V}) := E_s(\boldsymbol{z}, y, \boldsymbol{\omega}, \boldsymbol{V})\rho\left(\begin{pmatrix} z_1 - \omega_2 \\ \dots \\ z_{p-1} - \omega_{p-1} \end{pmatrix}^\top, v_{pp}\boldsymbol{d}(\boldsymbol{V})^{-1}\right).$$

Then, defining $\mathbb{R}_{+1} = [0, +\infty)$, $\mathbb{R}_{-1} = (-\infty, 0]$ and $S_a = \{\boldsymbol{s} = (s_1, ..., s_p) | \exists \boldsymbol{j} = \{j_1, ..., j_a\} \subseteq \{1, ..., n\}$ s.t $s_j = 1$ for $j \in \boldsymbol{j}$ and $s_j = -1$ otherwise$\}$, we have that

$$Z_{\text{out}} = \sum_{a=1}^{p} \frac{\delta_{y,2a-p}}{2} \sum_{\boldsymbol{s} \in S_a} \int_{\mathbb{R}_{s_1}} \mathrm{d}z_1 \dots \int_{\mathbb{R}_{s_{p-1}}} \mathrm{d}z_{p-1} I_{s_p}(\boldsymbol{z}, y, \boldsymbol{\omega}, \boldsymbol{V}), \tag{70}$$

and the k^{th} component of $\boldsymbol{g}_{\text{out}}(y, \boldsymbol{\omega}, \boldsymbol{V})$ is

$$(\boldsymbol{g}_{\text{out}}(y, \boldsymbol{\omega}, \boldsymbol{V}))_k = \frac{1}{Z_{\text{out}}} \sum_{a=1}^{p} \frac{\delta_{y,2a-p}}{2} \sum_{\boldsymbol{s} \in S_a} s_k \int_{\mathbb{R}_{s_1}} \mathrm{d}z_1 \dots \int_{\mathbb{R}_{s_{p-1}}} \mathrm{d}z_{p-1} I_{s_p}(\boldsymbol{z}, y, \boldsymbol{\omega}, \boldsymbol{V}) \delta(z_k - 0)$$

Appendix H. Further numerical observations and details

Here we give more details about the numerical implementation of State Evolution and AMP. Both approaches require to compute $\boldsymbol{g}_{out}$. All the examples we implemented are detailed in Appendix G.

The integrals in g_{out} are performed with the quadrature package in Scipy. In order to avoid instabilities we regularize the interval of integration by replacing ∞ with Λ. We typically choose $\Lambda \approx 10$. Similarly, we add $\epsilon \approx 10^{-4}$ to the diagonal of V. In State Evolution we need to integrate over functions of $\boldsymbol{g}_{out}$. We do these integrals using a simple Monte Carlo approach. For Figures 2 and 1 we used a total 72000 samples. Computing such integrals is the numerical bottleneck. In order to

Algorithm 1 Multi-index AMP

```
Input: Data $\boldsymbol{X} \in \mathbb{R}^{n\times d}$, $\boldsymbol{y} \in \mathbb{R}^n$
Initialize $\hat{\boldsymbol{W}}^{t=0} \in \mathbb{R}^{p\times d}$, $\hat{\boldsymbol{C}}_k^{t=0} \in \mathcal{S}_p^+$ for $k \in [d]$, $\boldsymbol{g}^{t=0} \in \mathbb{R}^{n\times p}$.
for $t \leq T$ do
    /* Update likelihood mean and variance
    $\boldsymbol{V}_i^t = \sum_{k=1}^d X_{ik}^2 \hat{\boldsymbol{C}}_k \in \mathbb{R}^{p\times p}$;    $\omega_i^t = \sum_{k=1}^d X_{ik} \hat{\boldsymbol{W}}_k^t - \boldsymbol{V}_i^t \boldsymbol{g}_i^{t-1} \in \mathbb{R}^p$, $i \in [n]$;
    $\boldsymbol{g}_i^t = g_{\text{out}}(y_i, \boldsymbol{\omega}_i^t, \boldsymbol{V}_i^t) \in \mathbb{R}^p$ ;    $\partial \boldsymbol{g}_i^t = \partial_{\boldsymbol{\omega}} g_{\text{out}}(y_i, \boldsymbol{\omega}_i^t, \boldsymbol{V}_i^t) \in \mathbb{R}^{p\times p}$ ; $i \in [n]$
    /* Update prior first and second moments
    $\boldsymbol{A}_k^t = -\sum_{i=1}^n X_{ik}^2 \partial \boldsymbol{g}_i^t \in \mathbb{R}^{p\times p}$ ;    $\boldsymbol{b}_k^t = \sum_{i\in[n]} X_{ik} \boldsymbol{g}_i^t + \boldsymbol{A}_k^t \hat{\boldsymbol{W}}_k^t$;    $k \in [d]$
    $\hat{\boldsymbol{W}}_k^{t+1} = (\boldsymbol{I}_p + \boldsymbol{A}_k^t)^{-1} \boldsymbol{b}_k^t \in \mathbb{R}^p$ ;    $\hat{\boldsymbol{C}}_k^{t+1} = (\boldsymbol{I}_p + \boldsymbol{A}_k^t)^{-1} \in \mathbb{R}^{p\times p}$,    $k \in [d]$
end for
Return: Estimators $\hat{\boldsymbol{W}}_{amp} \in \mathbb{R}^{p\times d}$, $\hat{\boldsymbol{C}}_{amp} \in \mathbb{R}^{d\times p\times p}$
```

make this part faster we parallelised the MCMC: for each iteration of State Evolution we make every worker in our pool estimate the integral, and then average the estimation and then average among workers. In the cases in which $\boldsymbol{M} = 0$ is a fixed point, we initialize $\boldsymbol{M}$ with the empirical overlap of AMP at the beginning of the iteration. We describe in detail the AMP iteration in 1. In both the AMP and State Evolution implementation we used some damping: the overlap $\boldsymbol{M}$ or the $\hat{\boldsymbol{W}}$ at the new iterations are averaged with the current value, with a weight δ for the new one, where typically $0.6 < \delta < 0.9$. We display the evolution of the overlaps in a typical run of AMP for the model $z_1^2 + \mathrm{s}ign(z_1 z_2 z_3)$ in Figure 3. We already displayed the values of the overlap and generalisation error at convergence in Figure 1. We can see how AMP has a saddle-to-saddle dynamic, where the algorithm alternates plateaus for $\mathcal{O}(\log d)$ iterations to fast drops in generalisation error, which are associated with new directions being learned.

As stated in the main text, models non-trivial subspaces are associated with symmetries. This also implies that the associated overlaps are have the same invariances. In order to make the plots readable we remove all such symmetries by hand. For this reason we take the absolute value of the overlap if the model is even, and we impose a specific inequality in the overlap if there is invariance under permutation. As a general idea we want to always have the ”best” possible overlap. Meaning we want $\boldsymbol{M}$ to be as diagonal as possible. We list what this mean for the examples in the figures:

- $\mathrm{s}ign(z_1 z_2)$: Because of invariance under exchange of z_1 and z_2 we always have $\boldsymbol{M}_{11} = \boldsymbol{M}_{22}$. Here AMP can reach 2 equivalent configurations: either the diagonal or the anti-diagonal is zero. We choose the configuration where the anti-diagonal is zero.

- $z_1^2 + \mathrm{s}ign(z_1 z_2 z_3)$: This model will first learn just z_1. We fix $\boldsymbol{M}_{11} > 0$. For the rest of the components we are reduced to the case above.

- $\mathrm{s}ign(z_1) + \mathrm{s}ign(z_2) + \mathrm{s}ign(z_3)$: Because of the invariance under permutation each row of $\boldsymbol{M}$ will have either the same element in all the entries (in which case we don't need to do anything) or two are the same and one is bigger. We permute the matrix such that the largest entry is on the diagonal.