Fluctuations, Bias, Variance & Ensemble of Learners: Exact Asymptotics for Convex Losses in High-Dimension

Bruno Loureiro [1 *] **Cédric Gerbelot** [2 *] **Maria Refinetti** [2] **Gabriele Sicuro** [3 *] **Florent Krzakala** [1]

Abstract

From the sampling of data to the initialisation of parameters, randomness is ubiquitous in modern Machine Learning practice. Understanding the statistical fluctuations engendered by the different sources of randomness in prediction is therefore key to understanding robust generalisation. In this manuscript we develop a quantitative and rigorous theory for the study of fluctuations in an ensemble of generalised linear models trained on different, but correlated, features in high-dimensions. In particular, we provide a complete description of the asymptotic joint distribution of the empirical risk minimiser for generic convex loss and regularisation in the high-dimensional limit. Our result encompasses a rich set of classification and regression tasks, such as the lazy regime of over-parametrised neural networks, or equivalently the random features approximation of kernels. While allowing to study directly the mitigating effect of ensembling (or bagging) on the bias-variance decomposition of the test error, our analysis also helps disentangle the contribution of statistical fluctuations, and the singular role played by the interpolation threshold that are at the roots of the "double-descent" phenomenon.

1. Introduction

Randomness is ubiquitous in Machine Learning. It is present in the data (e.g., noise in acquisition and annotation), in commonly used statistical models (e.g., random features (Rahimi & Recht, 2007)), or in the algorithms used to train them (e.g., in the choice of initialisation of weights of neural networks (Narkhede et al., 2021), or when sampling a mini-batch in Stochastic Gradient Descent (Bottou, 2012)). Strikingly, fluctuations associated to different sources of randomness can have a major impact in the generalisation performance of a model. For instance, this is the case in least-squares regression with random features, where it has been shown (Geiger et al., 2020; D'Ascoli et al., 2020; Jacot et al., 2020) that the variance associated with the random projections matrix is responsible for poor generalisation near the interpolation peak (Advani & Saxe, 2017; Spigler et al., 2019; Belkin et al., 2020). As a consequence, this *double-descent* behaviour can be mitigated by averaging over a large *ensemble* of learners, effectively suppressing this variance. Indeed, considering an ensemble (sometimes also refereed to as a committee (Drucker et al., 1994)) of independent learners provide a natural framework to study the contribution of the variance of prediction in the estimation accuracy. In this manuscript we leverage this idea to provide an exact asymptotic characterisation of the statistics of fluctuations in empirical risk minimisation with generic convex losses and penalties in high-dimensional models. We focus on the case of synthetic datasets, and we apply our results to random feature learning in particular.

Setting Let $(\boldsymbol{x}^\mu, y^\mu) \in \mathbb{R}^d \times \mathcal{Y}$, $\mu \in [n] \coloneqq \{1, \dots, n\}$, denote a labelled data set composed of n independent samples from a joint density $p(\boldsymbol{x}, y)$ (e.g., $\mathcal{Y} = \{-1, 1\}$ for a binary classification problem). In this manuscript we are interested in studying an ensemble of K parametric predictors, each of them depending on a vector of parameters $\boldsymbol{w}_k \in \mathbb{R}^p$, $k \in [K]$, and independently trained on the dataset $\{(\boldsymbol{x}^\mu, y^\mu)\}_{\mu \in [n]}$. Note that even if the vectors of parameters $\{\boldsymbol{w}_k\}_{k \in [K]}$ are trained independently, they correlate through the training data. Statistical fluctuations in the learnt parameters can then arise for different reasons. For instance, a common practice is to initialise the parameters randomly during optimisation, which will induce statistical variability between the different predictors. Alternatively, each predictor could be trained on a subsample of the data, as it is commonly done in bagging (Breiman, 1996). The statistical model can also be inherently stochastic, e.g., the

[*]Equal contribution [1]IdePHICS lab, École Polytechnique Fédérale de Lausanne, Rte Cantonale 1015 Lausanne, Switzerland [2]Laboratoire de Physique de l'ENS, Université PSL, CNRS, Sorbonne Université, 24 Rue Lhomond, 75005 Paris, France [3]Department of Mathematics, King's College London, Strand WC2R 2LS, London, United Kingdom. Correspondence to: Bruno Loureiro <bruno.loureiro@epfl.ch>, Gabriele Sicuro <gabriele.sicuro@kcl.ac.uk>.

Proceedings of the 39th International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022.

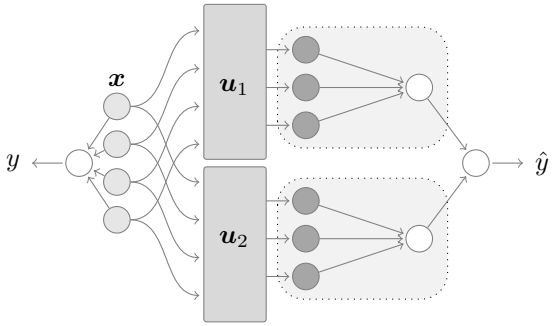

Figure 1. Pictorial representation of the model considered in the paper for $K = 2$. Two learners with the same architecture (in gray) receive a correlated input generated from the same vector $\boldsymbol{x} \sim \mathcal{N}(\mathbf{0}_d, \boldsymbol{I}_d)$. The output $\hat{y}$ is an average of their outputs. While the study of an ensemble of learners is already interesting *per se*, it is also pivotal to study the fluctuation between learners, and the error steaming from the difference in the weights in random features and lazy training.

random features approximation for kernel methods (Rahimi & Recht, 2007). Finally, the predictors could also be jointly trained, e.g., coupling them through the loss or penalty as it is done in boosting (Schapire, 1990).

Our goal in this work is to provide a sharp characterisation of the statistical fluctuations of the ensemble of parameters $\{\boldsymbol{w}_k\}_{k\in[K]}$ in a particular, mathematically tractable, class of predictors: *generalised linear models* (GLMs),

$$\hat{y}(\boldsymbol{x}) = \hat{f}\left(\frac{\hat{\boldsymbol{w}}_1^\top \boldsymbol{u}_1(\boldsymbol{x})}{\sqrt{p}}, \dots, \frac{\hat{\boldsymbol{w}}_K^\top \boldsymbol{u}_K(\boldsymbol{x})}{\sqrt{p}}\right) \tag{1}$$

where $\boldsymbol{u}_k : \mathbb{R}^d \to \mathbb{R}^p$, $k \in [K]$ is an ensemble of possibly correlated features and $\hat{f} \colon \mathbb{R}^K \to \mathcal{Y}$ is an activation function. For most of this work, we discuss the case in which the predictors are *independently trained* through regularised empirical risk minimisation:

$$\hat{\boldsymbol{w}}_k = \underset{\boldsymbol{w}\in\mathbb{R}^p}{\operatorname{argmin}} \left[\frac{1}{n}\sum_{\mu=1}^{n} \ell\left(y^\mu, \frac{\boldsymbol{w}^\top \boldsymbol{u}_k(\boldsymbol{x}^\mu)}{\sqrt{p}}\right) + \frac{\lambda}{2}\|\boldsymbol{w}\|_2^2\right] \tag{2}$$

with a convex loss function $\ell \colon \mathcal{Y} \times \mathbb{R} \to \mathbb{R}$ (e.g., the logistic loss) and ridge penalty whose strength is given by $\lambda \in \mathbb{R}^+$. However, our analysis also includes the case in which the learners are jointly trained with a generic convex penalty. This case will be further discussed in Sec. 4. In what follows we will also concentrate in the random features case where $\boldsymbol{u}_k(\boldsymbol{x}) = \phi(\boldsymbol{F}_k \boldsymbol{x})$ with $\phi : \mathbb{R} \to \mathbb{R}$ an activation function acting component-wise and $\boldsymbol{F}_k \in \mathbb{R}^{p\times d}$ a family of independently sampled random matrices. Besides being an efficient approximation for kernels (Rahimi & Recht, 2007), random features are often studied as a simple model for neural networks in the lazy and neural tangent kernel regimes of deep neural networks (Chizat et al., 2019; Jacot et al., 2018), in which case the matrices $\boldsymbol{F}_k$ correspond to different random initialisation of hidden-layer weights. Moreover, the random features model displays some of the exotic behaviours of high-dimensional overparametrised models, such as double-descent (Mei & Montanari, 2021; Gerace et al., 2020) and benign overfitting (Bartlett et al., 2020), therefore providing an ideal playground to study the interplay between fluctuations and overparametrisation. A broader class of features maps is also discussed in Sec. 4.

To provide an exact characterisation of the statistics of the estimators in eq. (2), we shall assume data is generated from a target

$$y = f_0\left(\frac{\boldsymbol{\theta}^\top \boldsymbol{x}}{\sqrt{d}}\right) \tag{3}$$

with $f_0 : \mathbb{R} \to \mathcal{Y}$ and $\boldsymbol{\theta} \sim \mathcal{N}(\mathbf{0}_d, \rho \boldsymbol{I}_d)$, $\rho \in \mathbb{R}_0^+$, $\boldsymbol{I}_d$ d-dimensional identity matrix. The dataset is then constructed generating *i.i.d.* n vectors $\boldsymbol{x}^\mu \sim \mathcal{N}(\mathbf{0}_d, \boldsymbol{I}_d)$, $\mu \in [n]$.

An illustration summary of the setting considered here in given in Figure 1. Note that such architecture can be interpreted as a two-layer tree neural network, also known in some contexts as the *tree-committee* or *parity machine* (Schwarze & Hertz, 1992). Although our focus on generalised linear models might appear restrictive, the investigation of this hypothesis class has proven fruitful in the current understanding of overparametrisation in learning, capturing the phenomenology of more complex architectures which is the ultimate goal of our contribution. Indeed, the popularity of the considered setting in theoretical works is justified by its connection with the lazy training limit of neural networks (Chizat et al., 2019; Jacot et al., 2018) and to kernel methods more generally, together with the fact that it displays a similar phenomenology to many non-convex networks (for instance, benign overfitting (Bartlett et al., 2020)). From this perspective, our analysis highlights the role played by overfitting the fluctuations due to random initialisation in the degradation of the generalisation of lazy deep neural networks (Geiger et al., 2020).

Main contributions The results in this manuscript can be listed as follows.

- We provide a sharp asymptotic characterisation of the joint statistics of the ensemble of empirical risk minimisers $\{\hat{\boldsymbol{w}}_k\}_{k\in[K]}$ in the high-dimensional limit where $p, n \to +\infty$ with n/p kept constant, for any convex loss and penalty. In particular, we show that the pre-activations $\{\hat{\boldsymbol{w}}_k^\top \boldsymbol{u}_k\}_{k\in[K]}$ are jointly Gaussian, with sufficient statistics obeying a set of explicit closed-form equations. Note that the analysis of ensembling with non-square losses is out of the grasp of the most commonly adopted theoretical tools (e.g., random matrix theory). Therefore, our proof method based on recent progress on Approximate Message Passing techniques (Berthier et al., 2020; Gerbelot & Berthier, 2021)

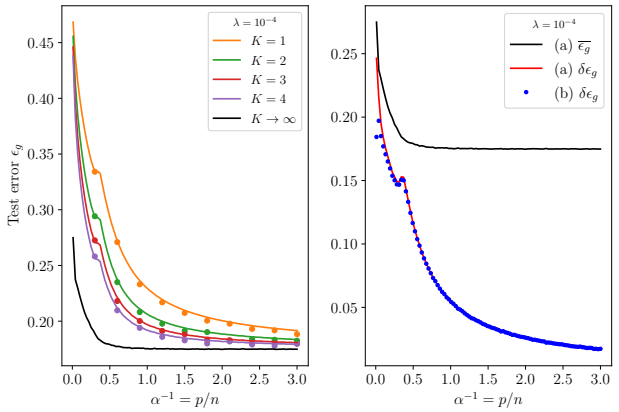

Figure 2. Left. Test error for logistic regression with $\lambda = 10^{-4}$ and different values of K as function of $p/n = 1/\alpha$ with $n/d = 2$ and $\rho = 1$. Dots represent the average of the outcomes of 10^3 numerical experiments. Here we adopted $\phi(x) = \mathrm{erf}(x)$ and estimator $\hat{f}(\boldsymbol{v}) = \mathrm{sign}(\sum_k v_k)$. *Right.* Decomposition of the $K = 1$ test error $\epsilon_g = \overline{\epsilon_g} + \delta\epsilon_g$ for the estimator **(a)**, with $n/d = 2$ and $\lambda = 10^{-4}$. We plot also the contribution $\delta\epsilon_g$ corresponding to the estimator **(b)**: we numerically observed that such decomposition coincides in the two cases. Note also the presence of a kink in $\delta\epsilon_g$ at the interpolation transition.

is of independent interest. Different versions of our theorem are discussed throughout the manuscript. First, in Sec. 2 for the particular case of independently trained learners on random features (Theorem 2.1). Later, in Sec. 4 for the general case of jointly trained learners on correlated Gaussian covariates (Theorem 4.1).

• We discuss the role played by fluctuations in the non-monotonic behaviour of the generalisation performance of interpolators (a.k.a. double-descent behaviour). In particular —as discussed in (Geiger et al., 2020; d'Ascoli et al., 2021) for the ridge case— the interpolation peak arises from the model overfitting the particular realisation of the random weights. We show the test error can be decomposed $\epsilon_g(K = 1) = \overline{\epsilon_g} + \delta\epsilon_g$ in terms of a fluctuation-free term $\overline{\epsilon_g}$ and a fluctuation term $\delta\epsilon_g$ responsible for the double-descent behavior, see Fig. 2 for the case of max-margin classification.

• In the context of classification, we discuss how *majority vote* and *score averaging*, two popular ensembling procedures, compare in terms of generalisation performance. More specifically, we show that in the setting we study score averaging consistently outperforms the majority vote predictor. However, for a large number of learners $K \gg 1$ these two predictors agree, see Fig. 6 (right).

• Finally, we discuss how ensembling can be used as a tool for uncertainty quantification. In particular, we connect the correlation between two learners to the probability of disagreement, and show that it decreases with overparametrisation, see Fig. 6 (left). We provide a full characterisation of the joint probability density of the confidence score between two independent learners, see Fig. 5.

Related works The idea of reducing the variance of a predictor by averaging over independent learners is quite old in Machine Learning (Hansen & Salamon, 1990; Perrone & Cooper, 1993; Perrone, 1994; Krogh & Vedelsby, 1995; Krogh & Sollich, 1997), and a variety of ensembling methods have been proposed (Opitz & Maclin, 1999). In a very inspiring work, Geiger et al. (2020) carried out an extensive series of experiments in order to shed light on the generalisation properties of neural networks, and reported many observations and empirical arguments about the role of the variance due to the random initialisation of the weights in the double-descent curve using an ensemble of learners. This was a major motivation for the present work. Closest to our setting is the work of Neal et al. (2018); D'Ascoli et al. (2020); Jacot et al. (2020) which disentangles the various sources of variance in the process of training deep neural networks. Indeed, here we adopt the model defined by D'Ascoli et al. (2020), and provide a rigorous justification of their results for the case of ridge regression. A slightly finer decomposition of the variance in terms of the different sources of randomness in the problem was later proposed by Adlam & Pennington (2020a). Lin & Dobriban (2021) show that such decomposition is not unique, and can be more generally understood from the point of view of the *analysis of variance* (ANOVA) framework. Interestingly, subsequent papers were able to identity a series of triple (and more) descent, e.g., (d'Ascoli et al., 2021; Adlam & Pennington, 2020b; Chen et al., 2020). (LeJeune et al., 2020) studies ensembles of ordinary least-squares learned from subsamples of a common data matrix, and shows its equivalence to an implicit ridge regularization.

The Random Features (RF) model was introduced in the seminal work of Rahimi & Recht (2007) as an efficient approximation for kernel methods. Drawing from early ideas of Karoui (2010), Pennington & Worah (2017) showed that the empirical distribution of the Gram matrix of RF is asymptotically equivalent to a linear model with matched second statistics, and characterised in this way memorisation with RF regression. The learning problem was first analysed by Mei & Montanari (2021), who provided an exact asymptotic characterisation of the training and generalisation errors of RF regression. This analysis was later extended to generic convex losses by Gerace et al. (2020) using the heuristic replica method, and later proved by Dhifallah & Lu (2020) using convex Gaussian inequalities.

The aforementioned asymptotic equivalence between the RF model and a Gaussian model with matched moments

has been named the *Gaussian Equivalence Principle* (GEP) (Goldt et al., 2020). Rigorous proofs in the memorisation and learning setting with square loss appeared in (Pennington & Worah, 2017; Mei & Montanari, 2021), and for general convex penalties in (Goldt et al., 2021; Hu & Lu, 2020). Goldt et al. (2021) and Loureiro et al. (2021b) provided extensive numerical evidence that the GEP holds for more generic feature maps, including features stemming from trained neural networks.

Most of the previously mentioned works deriving exact asymptotics for the RF model in the proportional limit use either Random Matrix Theory techniques or Convex Gaussian inequalities. While these tools have been recently used in many different contexts, they ultimately fall short when considering an ensemble of predictors with generic convex loss and regularisation, along with structured design matrices. Therefore, to prove the results herein we employ an *Approximate Message Passing* (AMP) proof technique (Bayati & Montanari, 2011a; Donoho & Montanari, 2016), leveraging on recently introduced progresses in (Loureiro et al., 2021b; Gerbelot & Berthier, 2021) which enables to capture the full complexity of the problem and obtain the asymptotic joint distribution of the ensemble of predictors.

2. Learning with an ensemble of random features

In this section give a first formulation of our main result, namely the exact asymptotic characterisation of the statistics of the ensembling estimator introduced in eq. (1). We prove that, in the proportional high dimensional limit, the statistics of the arguments of the activation function in eq. (1) is simply given by a multivariate Gaussian, whose covariance matrix we can completely specify. This result holds for any convex loss, any convex regularisation, and for all models of generative networks $\boldsymbol{u}_k \colon \mathbb{R}^d \to \mathbb{R}^p$, as we will show in full generality in Sec. 4. However, for simplicity, in this section and in the following we focus on the setting described in Sec. 1, in which the statistician averages over an independent ensemble of random features, i.e., $\boldsymbol{u}_k(\boldsymbol{x}) = \phi(\boldsymbol{F}_k\boldsymbol{x})$. In this case, our result can be formulated as follows:

Theorem 2.1 (Simplified version)**.** *Assume that in the high-dimensional limit where $d, p, n \to +\infty$ with $\alpha := {}^{n}\!/_{p}$ and $\gamma := {}^{d}\!/_{p}$ kept $\Theta(1)$ constants, the Wishart matrix $\boldsymbol{F}\boldsymbol{F}^\intercal$ has a well-defined asymptotic spectral distribution. Then in this limit, for any pseudo-Lispchitz function of order 2 $\varphi \colon \mathbb{R} \times \mathbb{R}^K \to \mathbb{R}$, we have*

$$\mathbb{E}_{(\boldsymbol{x},y)}\left[\varphi\left(y, \frac{\hat{\boldsymbol{w}}_1^\intercal \boldsymbol{u}_1}{\sqrt{p}}, \dots, \frac{\hat{\boldsymbol{w}}_K^\intercal \boldsymbol{u}_K}{\sqrt{p}}\right)\right] \xrightarrow{\mathrm{P}} \mathbb{E}_{(\nu,\boldsymbol{\mu})}[\varphi(f_0(\nu), \boldsymbol{\mu})], \tag{4}$$

where $(\nu, \boldsymbol{\mu}) \in \mathbb{R}^{K+1}$ is a jointly Gaussian vector $(\nu, \boldsymbol{\mu}) \sim \mathcal{N}(\boldsymbol{0}_{K+1}, \boldsymbol{\Sigma})$ with covariance

$$\boldsymbol{\Sigma} = \begin{pmatrix} \rho & m\boldsymbol{1}_K^\intercal \\ m\boldsymbol{1}_K & \boldsymbol{Q} \end{pmatrix}, \qquad \boldsymbol{Q} := (q_0 - q_1)\boldsymbol{I}_K + q_1 \boldsymbol{1}_{K,K}, \tag{5}$$

with $\boldsymbol{1}_{K,K} \in \mathbb{R}^{K\times K}$ and $\boldsymbol{1}_K \in \mathbb{R}^K$ are a matrix and a vector of ones respectively. The entries of $\boldsymbol{\Sigma}$ are solutions of a set of self-consistent equations given in Corollary 4.3.

Note that pseudo-Lipschitz functions of order 2 roughly amounts to functions with polynomially bounded increment, see Definition B.6 in the Appendix. As discussed in the introduction, the asymptotic statistics of the *single* learner has been studied in (Gerace et al., 2020; Dhifallah & Lu, 2020; Loureiro et al., 2021b). Their result amounts to the analysis of the estimator solving the empirical risk minimisation problem in eq. (2) and it is recovered imposing $K = 1$ in the theorem above. For $K = 1$, $(\nu, \mu) \in \mathbb{R}^2$ is jointly Gaussian with zero mean and covariance $\boldsymbol{\Sigma} = \left(\begin{smallmatrix} \rho & m \\ m & q_0 \end{smallmatrix}\right)$.

However, such result is not enough to quantify the correlation between different learners, induced by the training on the same dataset, which is required to compute, e.g., the test error associated with an ensembling predictor as in eq. (1). For example, in the simple case where $f_0(u) = u$ and $\hat{f}(\boldsymbol{v}) = \frac{1}{K}\sum_k v_k$, the mean-squared error on the labels is given by $\epsilon_g = \mathbb{E}_{(\boldsymbol{x},y)}[(y - \hat{y}(\boldsymbol{x}))^2] = \rho + (q_0 - q_1)K^{-1} + q_1 - 2m$, which crucially depends on the average correlation between two independent learners $q_1 := \frac{1}{p}\mathbb{E}[\hat{\boldsymbol{w}}_1^\intercal \hat{\boldsymbol{w}}_2]$. Our main result is precisely an exact asymptotic characterisation of this correlation in the proportional limit of the previous theorem. Once m, q_0 and q_1 have been determined, the generalisation error can be computed as

$$\epsilon_g := \mathbb{E}_{(\boldsymbol{x},y)}[\Delta(y, \hat{y}(\boldsymbol{x}))] \xrightarrow{n\to+\infty} \mathbb{E}_{(\nu,\boldsymbol{\mu})}\left[\Delta\left(f_0(\nu), \hat{f}(\boldsymbol{\mu})\right)\right] \tag{6}$$

for any error measure $\Delta \colon \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}^+$.

Suppose now that

$$\hat{f}(\boldsymbol{v}) \equiv \hat{f}_0\left(\tfrac{1}{K}\textstyle\sum_k v_k\right) \tag{7}$$

for some $\hat{f}_0 \colon \mathbb{R} \to \mathcal{Y}$ activation function of the single learner. In this case we can introduce the random variable $\hat{\mu} \overset{\mathrm{d}}{=} \lim_{K\to+\infty} \frac{1}{K}\sum_k \mu_k$. It is not difficult to see that the joint probability $p(\nu, \hat{\mu}) \sim \mathcal{N}(\boldsymbol{0}_2, \hat{\boldsymbol{\Sigma}})$ where $\hat{\boldsymbol{\Sigma}} = \left(\begin{smallmatrix} \rho & m \\ m & q_1 \end{smallmatrix}\right)$. This formally coincides with the joint distribution for the activation fields for $K = 1$ (Gerace et al., 2020), but with q_0 replaced by $q_1 \leq q_0$. The smaller variance is due to the fact that the fluctuations of the activation fields are averaged out by the ensembling process. The test error in the $K \to +\infty$ limit is then

$$\overline{\epsilon_g} := \mathbb{E}_{(\nu,\hat{\mu})}[\Delta(f_0(\nu), \hat{f}_0(\hat{\mu}))], \tag{8}$$

so that the fluctuation contribution to the test error for $K = 1$

can be defined as

$$\delta\epsilon_g := \mathbb{E}_{(\nu,\mu)}[\Delta(f_0(\nu), \hat{f}_0(\mu))] - \overline{\epsilon_g}. \tag{9}$$

The term $\delta\epsilon_g$ is by definition the contribution suppressed by ensembling and corresponds to the *ambiguity* introduced by Krogh & Vedelsby (1995) for the square loss. This contribution expresses the variance in the ensemble and it is responsible for the non-monotonic behaviour in the test error of interpolators, also known as the double-descent behavior.

3. Applications

We will consider now two relevant examples of separable losses, namely a ridge loss and a logistic loss. In both cases, it is possible to derive the explicit expression of the training loss and generalisation error in terms of the elements of the correlation matrix introduced above.

3.1. Ridge regression

If we assume $f_0(x) = x$, $\hat{f}(\boldsymbol{v}) = \frac{1}{K}\sum_k v_k$, and a quadratic loss of the type $\ell(y, x) = \frac{1}{2}(y - x)^2$, it is possible to write down simple recursive equations for m, q_0 and q_1 (see Appendix A.3). Taking $\Delta(y, \hat{y}) = (y - \hat{y})^2$, the generalisation error is easily computed as

$$\epsilon_g = \rho + \frac{q_0 - q_1}{K} + q_1 - 2m \xrightarrow{K\to+\infty} \rho + q_1 - 2m \equiv \overline{\epsilon_g}. \tag{10}$$

Note that in this case the $\lambda \to 0^+$ limit gives the minimum ℓ_2-norm interpolator. In Fig. 3 we compare our theoretical prediction with numerical results for $\lambda = 10^{-6}$ and various values of K. It is evident that the divergence of the generalisation error at $\alpha = 1$ is only due to the divergence of q_0, whereas the contribution $\overline{\epsilon_g}$, which is independent on q_0, is smooth everywhere. Alongside with the interpolation divergence, $\delta\epsilon_g = q_0 - q_1$ has an additional bump at $p/n = d/n$, which corresponds to the "linear peak" discussed by d'Ascoli et al. (2021). In the plot we present also the so-called kernel limit, corresponding to the limit $n/p = \alpha \to 0$ at fixed n/d. An explicit manipulation (see Appendix A.3) shows that $q_1 = q_0 \equiv q$ in this limit. This implies that in the kernel limit ϵ_g^{k} does not depend on K, being equal to $\epsilon_g^{\mathrm{k}} \equiv \rho + q - 2m$. The generalisation error obtained in the kernel limit coincides with $\overline{\epsilon_g}$ for $p > n$: this is expected as in $\overline{\epsilon_g}$ the fluctuations amongst learners are averaged out, effectively recovering the cost obtained in the case of an infinite number of parameters.

3.2. Binary classification

Suppose now that we are considering a classification task, such that $\mathcal{Y} = \{-1, 1\}$. For this task we consider $f_0(x) = \mathrm{sign}(x)$. A popular choice of loss in this classification task

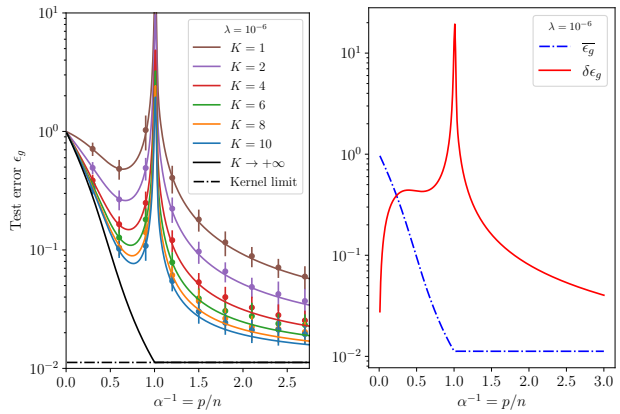

Figure 3. Left. Test error for ridge regression with $\lambda = 10^{-6}$ and different values of K as function of $p/n = 1/\alpha$ with $n/d = 2$ and $\rho = 1$. Dots represent the average of the outcomes of 50 numerical experiments in which the parameters of the neurons are estimated using $\min(d, p) = 200$. Here we adopted $\phi(x) = \mathrm{erf}(x)$. *Right.* Decomposition of $\epsilon_g = \overline{\epsilon_g} + \delta\epsilon_g$ in the $K = 1$ case

is the logistic loss,

$$\ell(y, x) = \ln\big(1 + \mathrm{e}^{-yx}\big), \tag{11}$$

although other choices, e.g, hinge loss, can be considered. Since both the logistic and hinge losses depend only on the *margin* $y\boldsymbol{w}^\intercal\boldsymbol{u}$, the empirical risk minimiser for $\lambda \to 0^+$ in both cases give the max-margin interpolator (Rosset et al., 2004). We write down the explicit saddle-point equations associated to the logistic and hinge loss in Appendix A.3, but we will focus our attention on the logistic case for the sake of brevity. For this choice of the loss, we obtained the values of m, q_0 and q_1 showed in Fig. 4. Using these values, a number of relevant questions can be addressed.

Alignment of learners – Assuming that the predictor of the learner k is $\hat{y}_k(\boldsymbol{x}) = \mathrm{sign}(\hat{\boldsymbol{w}}_k^\intercal \boldsymbol{u}_k(\boldsymbol{x}))$, in Fig. 6 (left) we estimate the probability that two learners give opposite classification. This is analytically given by

$$\mathbb{P}[\hat{y}_1(\boldsymbol{x}) \neq \hat{y}_2(\boldsymbol{x})] = \mathbb{P}[\mu_1\mu_2 < 0] = \frac{1}{\pi}\arccos\Big(\frac{q_1}{q_0}\Big). \tag{12}$$

Note that by definition the ratio q_1/q_0 is a cosine similarity between two learners in the norm induced by the feature space. Therefore, this provides an interesting interpretation of these sufficient statistics in terms of the probability of disagreement. In particular, as illustrated in Fig. 6 (left) over-parametrisation promotes agreement between the learners, therefore suppressing uncertainty. More generally, ensembling can be used as a technique for uncertainty estimation (Lakshminarayanan et al., 2017). In the context of logistic regression, the pre-activation to the sign function is often interpreted as a *confidence score*. Indeed, introducing the

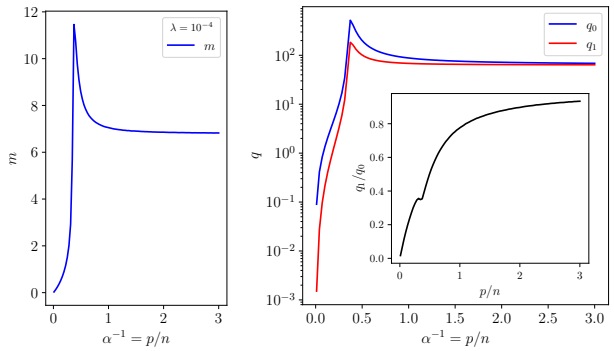

Figure 4. Analytical estimation of the order parameters m (left), q_0 and q_1 (right) in a classification task using logistic loss with ridge penalty with $\lambda = 10^{-4}$ at fixed $n/d = 2$ as function of p/n. In the inset, ratio q_1/q_0, quantifying the correlation between two learners. In all parameters the interpolation kink is clearly visible.

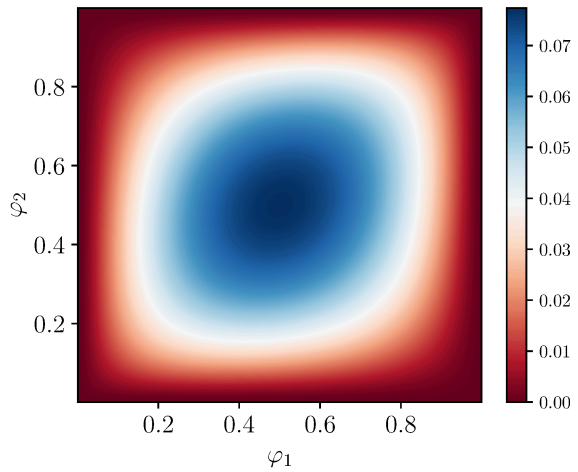

Figure 5. Joint probability density of the confidence score $\varphi_i(\boldsymbol{x}) = (1 + \exp(-p^{-1/2}\hat{\boldsymbol{w}}_i^\intercal \boldsymbol{u}_i(\boldsymbol{x})))^{-1}$ of two learners for $p/n \simeq 0.13$.

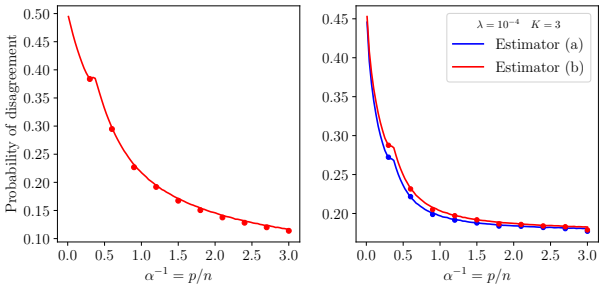

Figure 6. Left. Probability that two learners give discordant predictions using logistic regression as function of $p/n = 1/\alpha$ with $n/d = 2$, $\rho = 1$, and $\lambda = 10^{-4}$. *Right.* Test error for logistic regression using the estimators in eq. (13) and $K = 3$, with the same parameters. We adopted $\phi(x) = \mathrm{erf}(x)$. We observe that the test error obtained using **(a)** is always smaller than the one obtained using **(b)**. (*Left and right*) Dots represent the average of the outcomes of 10^3 numerical experiments.

logistic function $\varphi_k(\boldsymbol{x}) = (1+\exp(-p^{-1/2}\hat{\boldsymbol{w}}_k^\intercal \boldsymbol{u}_k(\boldsymbol{x})))^{-1}$, it expresses the confidence of the kth classifier in associating $\hat{y} = 1$ to the input $\boldsymbol{x}$. Therefore, it is reasonable to ask how reliable is the logistic score as a confidence measure. For instance, what is the variance of the confidence among different learners? This can be quantified by the joint probability density $\rho(\varphi_1, \varphi_2) := \mathbb{E}_{\boldsymbol{x}}[\delta(\varphi_1 - \varphi_1(\boldsymbol{x}))\delta(\varphi_2 - \varphi_2(\boldsymbol{x}))]$, which can be readily computed using our Theorem 2.1. Fig. 5 shows one example at fixed p/n and vanishing λ.

Ensemble predictors – In the previous two points, we discussed how ensembling can be used as a tool to quantify fluctuations. However, ensembling methods are also used in practical settings in order to mitigate fluctuations, e.g., (Breiman, 1996). An important question in this context is: given an ensemble of predictors $\{\hat{\boldsymbol{w}}_k\}_{k\in[K]}$, what is the best way of combining them to produce a point estimate? In our setting, this amounts to choosing the function $\hat{f} : \mathbb{R}^K \to \mathcal{Y}$. Let us consider two popular choices for the estimator $\hat{f}$ in eq. (1) used in practice:

$$\textbf{(a)} \qquad \hat{f}(\boldsymbol{v}) = \mathrm{sign}\left(\textstyle\sum_k v_k\right), \tag{13a}$$

$$\textbf{(b)} \qquad \hat{f}(\boldsymbol{v}) = \mathrm{sign}\left(\textstyle\sum_k \mathrm{sign}(v_k)\right). \tag{13b}$$

In a sense, **(a)** provides an estimator based on the average of the output fields, whereas **(b)**, which corresponds to a majority rule if K is odd (Hansen & Salamon, 1990), is a function of the average of the estimators of the single learners. For both choices of the estimator we use $\Delta(y, \hat{y}) = \delta_{\hat{y},y}$ to measure the test error. In Fig. 6 (right) we compare the test error obtained using **(a)** and **(b)** for $K = 3$ with vanishing regularisation $\lambda = 10^{-4}$. It is observed that the estimator **(a)** has better performances than the estimator **(b)**. As previously discussed, in this case logistic regression is equivalent to max-margin estimation, and in this case the error **(a)** can be intuitively understood in terms of a robust max-margin estimation obtained by averaging the margins associated to different draws of the random features. In the case **(a)** it is easy to show that the generalisation error takes the form

$$\epsilon_g = \frac{1}{\pi}\arccos\left(\frac{\sqrt{K}m}{\sqrt{\rho(q_0 - q_1 + Kq_1)}}\right) \xrightarrow{K\to\infty} \frac{1}{\pi}\arccos\left(\frac{m}{\sqrt{\rho q_1}}\right) \equiv \overline{\epsilon_g}. \tag{14}$$

This formula is in agreement with numerical experiments, see Fig. 2 (left). Unfortunately, we did not find a similar closed-form expression in case **(b)**. However, we can observe that in the $K \to +\infty$ limit the generalisation error in case **(a)** coincides with the generalisation error in case **(b)**, see Fig. 2 (right). By comparing with the results in Fig. 6 (left), it is evident that the benefit of ensembling in reducing the test error correlates with the tendency of learners to disagree, i.e., for small values of p/n, as stressed by Krogh & Vedelsby (1995). Finally, we observe a constant value of

$\overline{\epsilon_g}$ beyond the interpolation threshold, compatibly with the numerical results of Geiger et al. (2020).

4. The case of general loss and regularisation

In this Section we generalise our results in Sec. 2 relaxing the hypothesis on the loss, on the regularisation and on the properties of the feature maps.

In the general setting we are going to consider, we denote $P_y^0(y|x)$ the probabilistic law by which y is generated. For example, in Sec. 2, $P_y^0(y|x) = \delta(y - f_0(x))$. In the treatment given here, we allow for more general cases (e.g., the presence of noise in the label generation). We make weak assumptions on the generative networks $\boldsymbol{u}_k$, so that the information about the first layer is contained in the following tensors,

$$\boldsymbol{\Omega} := \mathbb{E}_{\boldsymbol{x}}[\boldsymbol{U}(\boldsymbol{x}) \otimes \boldsymbol{U}(\boldsymbol{x})] \in \mathbb{R}^{p\times p} \otimes \mathbb{R}^{K\times K}, \tag{15}$$

$$\hat{\boldsymbol{\Phi}} := \mathbb{E}_{\boldsymbol{x}}[\boldsymbol{U}(\boldsymbol{x})\boldsymbol{x}^\intercal \boldsymbol{\theta}] \in \mathbb{R}^{p\times K}, \tag{16}$$

$$\boldsymbol{\Theta} = \hat{\boldsymbol{\Phi}} \otimes \hat{\boldsymbol{\Phi}} \in \mathbb{R}^{p\times p} \otimes \mathbb{R}^{K\times K}. \tag{17}$$

In the equations above, $\boldsymbol{U}(\boldsymbol{x}) \in \mathbb{R}^{p\times K}$ is the matrix having as concatenated columns $\boldsymbol{u}_k(\boldsymbol{x})$: we will assume that, for each k, $\boldsymbol{\Omega}_{kk} \in \mathbb{R}^{p\times p}$ has bounded spectrum with strictly positive eigenvalues. We aim at learning a rule as in eq. (1), adopting a general convex loss $\hat{\ell}\colon \mathcal{Y} \times \mathbb{R}^K \to \mathbb{R}$, so that the weights are estimated as

$$\hat{\boldsymbol{W}} = \underset{\boldsymbol{W}\in\mathbb{R}^{p\times K}}{\operatorname{argmin}} \left[\frac{1}{n}\sum_{\mu=1}^{n} \hat{\ell}\left(y^\mu, \frac{\operatorname{diag}(\boldsymbol{W}^\intercal \boldsymbol{U}^\mu)}{\sqrt{p}}\right) + \lambda r(\boldsymbol{W}) \right] \tag{18}$$

where $r\colon \mathbb{R}^{p\times K} \to \mathbb{R}$ is a convex regularisation, $\boldsymbol{U}^\mu \equiv \boldsymbol{U}(\boldsymbol{x}^\mu)$ and $\hat{\boldsymbol{W}} \in \mathbb{R}^{p\times K}$ matrix of the concatenated columns $\{\hat{\boldsymbol{w}}_k\}$. Here, since the optimization problem defining the estimator may be non strictly convex, the solution may not be unique. We then denote with $\hat{\boldsymbol{W}}$ the unique least ℓ_2 norm solution of Eq.(18). As anticipated for Corollary 4.2, both $\hat{\ell}$ and r are assumed to be bounded by order 2 polynomials, or, more precisely, to be pseudo-Lipschitz functions of order 2 with respect to their arguments.

In the most general case, the statistical properties of $\hat{\boldsymbol{W}}$ are captured by a finite set of finite-dimensional order parameters, namely $\boldsymbol{V}, \hat{\boldsymbol{V}}, \boldsymbol{Q}, \hat{\boldsymbol{Q}} \in \mathbb{R}^{K\times K}$ and $\boldsymbol{m}, \hat{\boldsymbol{m}} \in \mathbb{R}^K$. These order parameters satisfy a set of fixed-point equations. To avoid a proliferation of indices in our formulas, let us introduce some notation. Let $\mathsf{A} = (A_{kk'}^{ij})_{k,k'\in[K]}^{i,j\in[p]} \in \mathbb{R}^{p\times p} \otimes \mathbb{R}^{K\times K}$ be a tensor, and $\boldsymbol{X} = (X_k^i)_{k\in[K]}^{i\in[p]}$, $\boldsymbol{Y} = (Y_k^i)_{k\in[K]}^{i\in[p]}$, $\boldsymbol{X}, \boldsymbol{Y} \in \mathbb{R}^{p\times K}$ two matrices. We will denote

$$\langle\!\langle \mathsf{A} \rangle\!\rangle := (\textstyle\sum_i A_{kk'}^{ii})_{kk'} \in \mathbb{R}^{K\times K}, \tag{19a}$$

$$\langle\!\langle \boldsymbol{X}|\mathsf{A}|\boldsymbol{Y} \rangle\!\rangle := (\textstyle\sum_{ij} X_k^i A_{kk'}^{ij} Y_{k'}^j)_{kk'} \in \mathbb{R}^{K\times K}, \tag{19b}$$

$$\langle\!\langle \boldsymbol{X}|\boldsymbol{Y} \rangle\!\rangle := (\textstyle\sum_{ij} X_k^i Y_k^i)_k \in \mathbb{R}^K, \tag{19c}$$

$$\langle \boldsymbol{X}|\mathsf{A}|\boldsymbol{Y} \rangle := \textstyle\sum_{ijk} X_k^i A_{kk}^{ij} Y_k^j \in \mathbb{R} \tag{19d}$$

$$\langle \boldsymbol{X}|\boldsymbol{Y} \rangle := \textstyle\sum_{ik} X_k^i Y_k^i \in \mathbb{R}. \tag{19e}$$

Given a second tensor $\mathsf{B} \in \mathbb{R}^{p\times p} \otimes \mathbb{R}^{K\times K}$, we write

$$\mathsf{AB} := (\textstyle\sum_{i'\kappa} A_{k\kappa}^{ii'} B_{\kappa k'}^{i'j})_{kk'}^{ij} \in \mathbb{R}^{p\times p} \otimes \mathbb{R}^{K\times K}, \tag{19f}$$

$$\mathsf{A} \circ \mathsf{B} := (\textstyle\sum_{i'} A_{kk'}^{ii'} B_{k'k}^{i'j})_{kk'}^{ij} \in \mathbb{R}^{p\times p} \otimes \mathbb{R}^{K\times K}, \tag{19g}$$

$$\mathsf{A} \odot \mathsf{B} := (A_{kk'}^{ij} B_{kk'}^{ij})_{kk'}^{ij} \in \mathbb{R}^{p\times p} \otimes \mathbb{R}^{K\times K}. \tag{19h}$$

We can now state our general result.

Theorem 4.1. *Let us consider the random quantities $\boldsymbol{\xi} \in \mathbb{R}^K$ and $\boldsymbol{\Xi} \in \mathbb{R}^{K\times K}$ with entries distributed as $\mathcal{N}(0,1)$. Assume that in the high-dimensional limit where $d, p, n \to +\infty$ with $\alpha := n/p$ and $\gamma := d/p$ kept $\Theta(1)$ constants. Then in this limit, for any pseudo-Lispchitz functions of order 2 $\varphi\colon \mathbb{R}\times\mathbb{R}^K \to \mathbb{R}$ and $\tilde{\varphi}\colon \mathbb{R}^{K\times p} \to \mathbb{R}$, the estimator $\hat{\boldsymbol{W}}$ verifies*

$$\begin{aligned}
&\mathbb{E}_{(y,\boldsymbol{x})}\left[\varphi\left(y, \frac{\langle\!\langle \hat{\boldsymbol{W}}|\boldsymbol{U} \rangle\!\rangle}{\sqrt{p}}\right)\right] \xrightarrow{\mathrm{P}} \int_{\mathcal{Y}} \mathrm{d}y \mathbb{E}_{(\nu,\boldsymbol{\mu})}\left[P_y^0(y|\nu)\varphi(y,\boldsymbol{\mu})\right],\\
&\frac{1}{n}\sum_{\mu=1}^{n}\varphi\left(y^\mu, \frac{\langle\!\langle \hat{\boldsymbol{W}}|\boldsymbol{U}^\mu \rangle\!\rangle}{\sqrt{p}}\right) \xrightarrow{\mathrm{P}} \int_{\mathcal{Y}} \mathrm{d}y \mathbb{E}_{\boldsymbol{\xi}}\left[\mathcal{Z}^0(y,\omega_0,\sigma_0)\,\varphi(y,\boldsymbol{h})\right],\\
&\tilde{\varphi}(\hat{\boldsymbol{W}}) \xrightarrow{\mathrm{P}} \mathbb{E}_{\boldsymbol{\Xi}}\left[\tilde{\varphi}(\boldsymbol{G})\right],
\end{aligned} \tag{20}$$

where $\boldsymbol{U} \equiv \boldsymbol{U}(\boldsymbol{x})$, $(\nu,\boldsymbol{\mu}) \in \mathbb{R}^{1+K}$ are jointly Gaussian random variables with zero mean and covariance matrix

$$(\nu,\boldsymbol{\mu}) \sim \mathcal{N}\left(\boldsymbol{0}_{1+K}, \begin{pmatrix} \rho & \boldsymbol{m}^\intercal \\ \boldsymbol{m} & \boldsymbol{Q} \end{pmatrix}\right), \tag{21}$$

and we have introduced the proximals for the loss and the regularisation:

$$\begin{aligned}
\boldsymbol{h} &:= \underset{\boldsymbol{u}}{\operatorname{argmin}}\left[\frac{(\boldsymbol{u}-\boldsymbol{\omega})\boldsymbol{V}^{-1}(\boldsymbol{u}-\boldsymbol{\omega})}{2} + \hat{\ell}(y,\boldsymbol{u})\right],\\
\boldsymbol{G} &:= \underset{\boldsymbol{U}}{\operatorname{argmin}}\left[\frac{\langle \boldsymbol{U}|(\boldsymbol{1}_{p,p}\otimes\hat{\boldsymbol{V}})\odot\boldsymbol{\Omega}|\boldsymbol{U}\rangle}{2} - \langle \boldsymbol{B}|\boldsymbol{U}\rangle + \lambda r(\boldsymbol{U})\right],
\end{aligned} \tag{22}$$

with $\boldsymbol{\omega} := \boldsymbol{Q}^{1/2}\boldsymbol{\xi}$ and $\boldsymbol{B} := (\boldsymbol{1}_p \otimes \hat{\boldsymbol{m}}^\intercal) \odot \hat{\boldsymbol{\Phi}} + ((\boldsymbol{1}_{p,p} \otimes \hat{\boldsymbol{Q}}) \odot \boldsymbol{\Omega})^{\frac{1}{2}}\boldsymbol{\Xi}$. We have also introduced the auxiliary function

$$\mathcal{Z}^0(y,\mu,\sigma) := \int \frac{P_y^0(y|x)\mathrm{d}x}{\sqrt{2\pi\sigma}}\, \mathrm{e}^{-\frac{(x-\mu)^2}{2\sigma}}. \tag{23}$$

and the scalar quantities $\omega_0 := \boldsymbol{m}^\intercal\boldsymbol{Q}^{-1/2}\boldsymbol{\xi}$ and $\sigma_0 := \rho - \boldsymbol{m}^\intercal\boldsymbol{Q}^{-1}\boldsymbol{m}$. The order parameters satisfy the saddle-point equations

$$\begin{aligned}
\hat{\boldsymbol{V}} &= -\alpha \int_{\mathcal{Y}} \mathrm{d}y\, \mathbb{E}_{\boldsymbol{\xi}}\left[\mathcal{Z}^0(y,\omega_0,\sigma_0)\,\partial_{\boldsymbol{\omega}}\boldsymbol{f}\right],\\
\hat{\boldsymbol{Q}} &= \alpha \int_{\mathcal{Y}} \mathrm{d}y\, \mathbb{E}_{\boldsymbol{\xi}}\left[\mathcal{Z}^0(y,\omega_0,\sigma_0)\,\boldsymbol{f}\boldsymbol{f}^\intercal\right],\\
\hat{\boldsymbol{m}} &= \tfrac{\alpha}{\sqrt{\gamma}} \int_{\mathcal{Y}} \mathrm{d}y\, \mathbb{E}_{\boldsymbol{\xi}}\left[\partial_\mu \mathcal{Z}^0(y,\omega_0,\sigma_0)\boldsymbol{f}\right],
\end{aligned} \tag{24}$$

and

$$\begin{aligned} \boldsymbol{V} &= \frac{2}{p}\mathbb{E}_{\boldsymbol{\Xi}}\left\langle \boldsymbol{G}\middle| D_{\hat{\boldsymbol{Q}}}((\mathbf{1}_{p,p}\otimes\hat{\boldsymbol{Q}})\odot\boldsymbol{\Omega})^{1/2}\middle|\boldsymbol{\Xi}\right\rangle \\ \boldsymbol{Q} &= \tfrac{1}{p}\mathbb{E}_{\boldsymbol{\Xi}}\langle\!\langle \boldsymbol{G}|\boldsymbol{\Omega}|\boldsymbol{G}\rangle\!\rangle, \\ \boldsymbol{m} &= \tfrac{1}{\sqrt{\gamma}p}\mathbb{E}_{\boldsymbol{\Xi}}\langle\!\langle \hat{\boldsymbol{\Phi}}|\boldsymbol{G}\rangle\!\rangle. \end{aligned} \tag{25}$$

In the equation above we have introduced the short-hand notation $\boldsymbol{f} := \boldsymbol{V}^{-1}(\boldsymbol{h}-\boldsymbol{\omega})$.

In the theorem above, for a tensor $\hat{\mathbf{A}} \in \mathbb{R}^{p\times p}\otimes\mathbb{R}^{K\times K}$, then $[\mathrm{D}_{\hat{\boldsymbol{Q}}}\hat{\mathbf{A}}]_{ij}^{kk',\kappa\kappa'} \equiv \partial_{\hat{Q}_{\kappa\kappa'}}\hat{A}_{ij}^{kk'}$: in the formula, the contractions involve latin indices only. Eqs. (24) are typically called *channel equations*, because depend on the form of the loss $\hat{\ell}$. Eqs. (25), instead, are usually called *prior equations*, because of their dependence on the prior, i.e., r. In the following Corollary, we specify their expression for a ridge regularisation, $r(\boldsymbol{W}) = \frac{1}{2}\|\boldsymbol{W}\|_{\mathrm{F}}^2$.

Corollary 4.2 (Ridge regularisation)**.** *In the hypotheses of Theorem 4.1, if* $r(\boldsymbol{W}) = \frac{1}{2}\|\boldsymbol{W}\|_{\mathrm{F}}^2$*, then the prior equations are*

$$\begin{aligned} \boldsymbol{V} &= \tfrac{1}{p}\langle\!\langle \boldsymbol{\Omega}\circ\mathbf{A}\rangle\!\rangle, \\ \boldsymbol{Q} &= \tfrac{1}{p}\langle\!\langle \boldsymbol{\Omega}\circ\Big(\mathbf{A}\Big((\mathbf{1}_{p,p}\otimes\hat{\boldsymbol{m}}\otimes\hat{\boldsymbol{m}}^{\intercal})\odot\boldsymbol{\Theta}+(\mathbf{1}_{p,p}\otimes\hat{\boldsymbol{Q}})\odot\boldsymbol{\Omega}\Big)\mathbf{A}\Big)\rangle\!\rangle, \\ \boldsymbol{m} &= \tfrac{1}{\sqrt{\gamma}p}\langle\!\langle \mathbf{A}((\mathbf{1}_{p,p}\otimes\hat{\boldsymbol{m}}\otimes\mathbf{1}_K^{\intercal})\odot\boldsymbol{\Theta})\rangle\!\rangle. \end{aligned} \tag{26}$$

In the equation above, we have used the auxiliary tensor $\mathbf{A} \equiv \mathbf{A}(\hat{\boldsymbol{V}};\lambda,\boldsymbol{\Omega}) := (\lambda\boldsymbol{I}_p\otimes\boldsymbol{I}_K + (\mathbf{1}_{p,p}\otimes\hat{\boldsymbol{V}})\odot\boldsymbol{\Omega})^{-1} \in \mathbb{R}^{p\times p}\otimes\mathbb{R}^{K\times K}$.

As shown in Appendix B, the proof of Theorem 4.1 requires a novel decomposition of the problem to obtain an appropriate AMP iteration scheme. Moreover, its specification in Corollary 4.2 required the application of asymptotic freeness properties of the involved random matrices, as discussed in further detail in Appendix A.3.

4.1. The random feature case and the kernel limit

Theorem 4.1 is given in a very general setting, and, in particular, no assumptions are made on the features $\boldsymbol{u}_k$. We have anticipated in Sec. 2 that, in the case of random features, the structure of the order parameters highly simplifies and the covariance matrix $\boldsymbol{\Sigma}$ is fully specified by only three scalar order parameters for any $K > 1$. Here will adapt therefore Theorem 4.1 to the random feature setting in Sec. 2, using the notation therein. The motivation of this section is to explicitly present the self-consistent equations that are required to produce the results given in the paper.

Corollary 4.3. *Assume that in the high-dimensional limit where* $d,p,n \to +\infty$ *with* $\alpha := {}^{n}\!/_{p}$ *and* $\gamma := {}^{d}\!/_{p}$ *kept* $\Theta(1)$ *constants, the Wishart matrix* $\boldsymbol{F}\boldsymbol{F}^{\intercal}$ *has a well-defined asymptotic spectral distribution. Then in this limit, for any pseudo-Lispchitz function of finite order* $\varphi\colon \mathbb{R}\times\mathbb{R}^K \to \mathbb{R}$*, the estimator* $\hat{\boldsymbol{W}}$ *verifies*

$$\mathbb{E}_{(\boldsymbol{x},y)}\left[\varphi\left(y, \frac{\langle\!\langle \hat{\boldsymbol{W}}|\boldsymbol{U}\rangle\!\rangle}{\sqrt{p}}\right)\right] \xrightarrow{\mathrm{P}} \mathbb{E}_{(\nu,\boldsymbol{\mu})}[\varphi(f_0(\nu),\boldsymbol{\mu})], \tag{27}$$

where $(\nu,\boldsymbol{\mu}) \in \mathbb{R}^{K+1}$ *is a jointly Gaussian vector with covariance*

$$(\nu,\boldsymbol{\mu}) \sim \mathcal{N}\left(\mathbf{0}_{K+1}, \begin{pmatrix} \rho & m\mathbf{1}_K^{\intercal} \\ m\mathbf{1}_K & \boldsymbol{Q} \end{pmatrix}\right), \tag{28}$$

and $\boldsymbol{Q} := (q_0-q_1)\boldsymbol{I}_K + q_1\mathbf{1}_{K,K}$*. The collection of parameters* (q_0,q_1,m) *is obtained solving a set of fixed point equations involving the auxiliary variables* $(\hat{q}_0,\hat{q}_1,\hat{m},v,\hat{v})$*, namely:*

$$\hat{v} = -\alpha\int_{\mathcal{Y}}\mathrm{d}y\,\mathbb{E}_{\omega}\left[\mathcal{Z}^0\left(y,\tfrac{m\omega}{q_0},\rho-\tfrac{m^2}{q_0}\right)\partial_{\omega}f\right], \tag{29a}$$

$$\hat{m} = \tfrac{\alpha}{\sqrt{\gamma}}\int_{\mathcal{Y}}\mathrm{d}y\,\mathbb{E}_{\omega}\left[\partial_{\mu}\mathcal{Z}^0\left(y,\tfrac{m\omega}{q_0},\rho-\tfrac{m^2}{q_0}\right)f\right], \tag{29b}$$

$$\hat{q}_0 = \alpha\int_{\mathcal{Y}}\mathrm{d}y\,\mathbb{E}_{\omega}\left[\mathcal{Z}^0\left(y,\tfrac{m\omega}{q_0},\rho-\tfrac{m^2}{q_0}\right)f^2\right], \tag{29c}$$

$$\hat{q}_1 = \alpha\int_{\mathcal{Y}}\mathrm{d}y\mathbb{E}_{\omega,\omega'}\left[\mathcal{Z}^0\left(y,m\tfrac{\omega+\omega'}{q_0+q_1},\rho-\tfrac{2m^2}{q_0+q_1}\right)ff'\right], \tag{29d}$$

$$v = \int\tfrac{s\varrho(s)\mathrm{d}s}{\lambda+s\hat{v}}, \tag{29e}$$

$$m = \tfrac{\hat{m}}{\sqrt{\gamma}}\int\tfrac{s-\kappa_*^2}{\lambda+\hat{v}s}\varrho(s)\mathrm{d}s, \tag{29f}$$

$$q_0 = \int\tfrac{(\hat{q}_0+\hat{m}^2)s^2-\hat{m}^2\kappa_*^2 s}{(\lambda+\hat{v}s)^2}\varrho(s)\mathrm{d}s, \tag{29g}$$

$$q_1 = \left(1+\tfrac{\hat{q}_1}{\hat{m}^2}\right)m^2. \tag{29h}$$

where ω *and* ω' *are two correlated Gaussian random variables of zero mean and* $\mathbb{E}[\omega^2] = \mathbb{E}[\omega'^2] = q_0$*,* $\mathbb{E}[\omega\omega'] = q_1$*. Moreover, we have introduced the proximals*

$$f = \frac{\mathrm{Prox}_{v\ell(y,\bullet)}(\omega)-\omega}{v}, \quad f' = \frac{\mathrm{Prox}_{v\ell(y,\bullet)}(\omega')-\omega'}{v}, \tag{30}$$

with

$$\mathrm{Prox}_{v\ell(y,\bullet)}(\omega) := \arg\min_x\left[\frac{(x-\omega)^2}{2v}+\ell(y,x)\right]. \tag{31}$$

Finally, $\varrho(s)$ *is the asymptotic spectral density of the features covariance matrix* $\boldsymbol{\Omega} \equiv \mathrm{Var}(\boldsymbol{u}) = \kappa_0^2\mathbf{1}_{p,p} + \frac{\kappa_1^2}{d}\boldsymbol{F}\boldsymbol{F}^{\intercal} + \kappa_*^2\boldsymbol{I}_p$ *and the coefficients are given by* $\kappa_0 := \mathbb{E}_{\zeta}[\phi(\zeta)]$*,* $\kappa_1 := \mathbb{E}_{\zeta}[\zeta\phi(\zeta)]$*,* $\kappa_* := \mathbb{E}_{\zeta}[\phi^2(\zeta)] - \kappa_0^2 - \kappa_1^2$ *with* $\zeta \sim \mathcal{N}(0,1)$*.*

The previous corollary recovers the results of Gerace et al. (2020), Dhifallah & Lu (2020), and Loureiro et al. (2021b) when restricted to the $K = 1$ case by marginalisation. Note that if all K learners have identical features, then all learners have identical performances and $q_0 = q_1$, see Appendix A.3, and in this case there is obviously no gain from ensembling.

The kernel limit – The so-called kernel limit is obtained by taking the limit of infinite number of parameters so that $\gamma \to 0$ (i.e., $p \gg d$ and $p \gg n$), but with a fixed ratio $\alpha/\gamma = n/d$. To balance the loss term and the regularisation it is convenient to rescale $\lambda \mapsto \alpha\lambda$. We can simplify the saddle-point equation in this special limit introducing $\hat{q}_0 \mapsto \alpha\hat{q}_0$, $\hat{q}_1 \mapsto \alpha\hat{q}_1$, $\hat{m} \mapsto \sqrt{\alpha}\hat{m}$, $\hat{v} \mapsto \alpha\hat{v}$. The channel equations keep a simple form,

$$\hat{v} = -\int_{\mathcal{Y}} \mathrm{d}y \mathbb{E}_\zeta \left[\mathcal{Z}^0\left(y, \omega_0, \sigma_0\right) \partial_\omega f\right], \tag{32a}$$

$$\hat{m} = \sqrt{\delta} \int_{\mathcal{Y}} \mathrm{d}y \mathbb{E}_\zeta \left[f \partial_\mu \mathcal{Z}^0\left(y, \omega_0, \sigma_0\right)\right], \tag{32b}$$

$$\hat{q}_0 = \int_{\mathcal{Y}} \mathrm{d}y \mathbb{E}_\zeta \left[\mathcal{Z}^0\left(y, \omega_0, \sigma_0\right) f^2\right], \tag{32c}$$

$$\hat{q}_1 = \int_{\mathcal{Y}} \mathrm{d}y \mathbb{E}_{\omega,\omega'} \left[\mathcal{Z}^0\left(y, m\tfrac{\omega+\omega'}{q_0+q_1}, \rho - \tfrac{2m^2}{q_0+q_1}\right) f f'\right]. \tag{32d}$$

The $p \to +\infty$ limit in the prior equations depends on the spectral density $\varrho(s)$. For example, if $\boldsymbol{F}$ has random Gaussian entries with zero mean and unit variance, then $\varrho(s)$ is a shifted Marchenko–Pastur distribution,

$$\varrho(s) = \nu(s - \kappa_*^2; \alpha^{-1}\delta, \kappa_1), \tag{33}$$

where, if $[x]_+ = x\theta(x)$,

$$\nu(x;b,a) = \frac{\sqrt{[(a_+ - x)(x - a_-)]_+}}{2ab^2\pi^2 x} + \left[1 - \frac{1}{a}\right]_+ \delta(x), \tag{34}$$

with $a_\pm := b^2(1 \pm \sqrt{a})^2$. By means of a series of algebraic manipulation, we obtain in the end at the first order in α

$$\begin{aligned} v &= \frac{\lambda(\kappa_1^2 + \kappa_*^2) + \delta^2\kappa_1^2\kappa_*^2\hat{v}}{\lambda(\lambda + \delta\kappa_1^2\hat{v})}, & q_0 &= \frac{\delta\kappa_1^4\left(\hat{q}_0 + \delta\hat{m}^2\right)}{\left(\lambda + \delta\kappa_1^2\hat{v}\right)^2}, \\ m &= \frac{\sqrt{\delta}\kappa_1^2\hat{m}}{\lambda + \delta\kappa_1^2\hat{v}}, & q_1 &= \frac{\delta\kappa_1^4\left(\hat{q}_1 + \delta\hat{m}^2\right)}{\left(\lambda + \delta\kappa_1^2\hat{v}\right)^2}, \end{aligned} \tag{35}$$

which complete our set of equations for the kernel limit.

Acknowledgements

We thank Ali Bereyhi, Giulio Biroli, Stéphane d'Ascoli, Justin Ko for discussions, and Francesca Mignacco for sharing code with us. We acknowledge funding from the French National Research Agency grants ANR-17-CE23-0023-01 PAIL and ANR-19P3IA-0001 PRAIRIE.

References

Adlam, B. and Pennington, J. Understanding double descent requires a fine-grained bias-variance decomposition. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F., and Lin, H. (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 11022–11032. Curran Associates, Inc., 2020a.

Adlam, B. and Pennington, J. The neural tangent kernel in high dimensions: Triple descent and a multi-scale theory of generalization. In *International Conference on Machine Learning*, pp. 74–84. PMLR, 2020b.

Advani, M. S. and Saxe, A. M. High-dimensional dynamics of generalization error in neural networks, 2017.

Bartlett, P. L., Long, P. M., Lugosi, G., and Tsigler, A. Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, 117(48):30063–30070, 2020.

Bauschke, H., Combettes, P., and Noll, D. Joint minimization with alternating bregman proximity operators. *Pacific Journal of Optimization*, 2006.

Bauschke, H. H., Borwein, J. M., and Combettes, P. L. Bregman monotone optimization algorithms. *SIAM Journal on control and optimization*, 42(2):596–636, 2003.

Bauschke, H. H., Combettes, P. L., et al. *Convex analysis and monotone operator theory in Hilbert spaces*, volume 408. Springer, 2011.

Bayati, M. and Montanari, A. The dynamics of message passing on dense graphs, with applications to compressed sensing. *IEEE Transactions on Information Theory*, 57 (2):764–785, 2011a.

Bayati, M. and Montanari, A. The LASSO risk for Gaussian matrices. *IEEE Transactions on Information Theory*, 58 (4):1997–2017, 2011b.

Belkin, M., Hsu, D., Ma, S., and Mandal, S. Reply to loog et al.: Looking beyond the peaking phenomenon. *Proceedings of the National Academy of Sciences*, 117 (20):10627–10627, 2020.

Berthier, R., Montanari, A., and Nguyen, P.-M. State evolution for approximate message passing with non-separable functions. *Information and Inference: A Journal of the IMA*, 9(1):33–79, 2020.

Bottou, L. Stochastic gradient descent tricks. In *Neural networks: Tricks of the trade*, pp. 421–436. Springer, 2012.

Breiman, L. Bagging predictors. *Machine Learning*, 24: 123–140, 1996.

Celentano, M., Montanari, A., and Wei, Y. The Lasso with general Gaussian designs with applications to hypothesis testing. *Preprint arXiv:2007.13716*, 2020.

Chen, L., Min, Y., Belkin, M., and Karbasi, A. Multiple descent: Design your own generalization curve. *arXiv preprint arXiv:2008.01036*, 2020.

Chizat, L., Oyallon, E., and Bach, F. On lazy training in differentiable programming. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

D'Ascoli, S., Refinetti, M., Biroli, G., and Krzakala, F. Double trouble in double descent: Bias and variance(s) in the lazy regime. In Daumé III, H. and Singh, A. (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 2280–2290. PMLR, 13–18 Jul 2020.

d'Ascoli, S., Sagun, L., and Biroli, G. Triple descent and the two kinds of overfitting: where and why do they appear? *Journal of Statistical Mechanics: Theory and Experiment*, 2021(12):124002, dec 2021.

Dhifallah, O. and Lu, Y. M. A precise performance analysis of learning with random features. *arXiv:2008.11904*, 2020.

Donoho, D. and Montanari, A. High dimensional robust m-estimation: Asymptotic variance via approximate message passing. *Probability Theory and Related Fields*, 166 (3):935–969, 2016.

Drucker, H., Cortes, C., Jackel, L. D., LeCun, Y., and Vapnik, V. Boosting and other ensemble methods. *Neural computation*, 6(6):1289–1301, 1994.

Geiger, M., Jacot, A., Spigler, S., Gabriel, F., Sagun, L., d'Ascoli, S., Biroli, G., Hongler, C., and Wyart, M. Scaling description of generalization with number of parameters in deep learning. *Journal of Statistical Mechanics: Theory and Experiment*, 2020(2):023401, 2020.

Gerace, F., Loureiro, B., Krzakala, F., Mezard, M., and Zdeborova, L. Generalisation error in learning with random features and the hidden manifold model. In III, H. D. and Singh, A. (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 3452–3462. PMLR, 13–18 Jul 2020.

Gerbelot, C. and Berthier, R. Graph-based approximate message passing iterations. *arXiv preprint arXiv:2109.11905*, 2021.

Goldt, S., Mézard, M., Krzakala, F., and Zdeborová, L. Modeling the influence of data structure on learning in neural networks: The hidden manifold model. *Physical Review X*, 10:041044, Dec 2020.

Goldt, S., Loureiro, B., Reeves, G., Krzakala, F., Mézard, M., and Zdeborová, L. The gaussian equivalence of generative models for learning with shallow neural networks. *Proceedings of Machine Learning Research*, 145:1–46, 2021.

Hansen, L. K. and Salamon, P. Neural network ensembles. *IEEE transactions on pattern analysis and machine intelligence*, 12(10):993–1001, 1990.

Hu, H. and Lu, Y. M. Universality laws for high-dimensional learning with random features. *arXiv:2009.07669*, 2020.

Jacot, A., Gabriel, F., and Hongler, C. Neural tangent kernel: Convergence and generalization in neural networks. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

Jacot, A., Simsek, B., Spadaro, F., Hongler, C., and Gabriel, F. Implicit regularization of random feature models. In *International Conference on Machine Learning*, pp. 4631–4640. PMLR, 2020.

Javanmard, A. and Montanari, A. State evolution for general approximate message passing algorithms, with applications to spatial coupling. *Information and Inference: A Journal of the IMA*, 2(2):115–144, 2013.

Karoui, N. E. The spectrum of kernel random matrices. *The Annals of Statistics*, 38(1):1 – 50, 2010.

Krogh, A. and Sollich, P. Statistical mechanics of ensemble learning. *Phys. Rev. E*, 55:811–825, Jan 1997.

Krogh, A. and Vedelsby, J. Neural network ensembles, cross validation, and active learning. In Tesauro, G., Touretzky, D., and Leen, T. (eds.), *Advances in Neural Information Processing Systems*, volume 7. MIT Press, 1995.

Lakshminarayanan, B., Pritzel, A., and Blundell, C. Simple and scalable predictive uncertainty estimation using deep ensembles, 2017.

LeJeune, D., Javadi, H., and Baraniuk, R. The implicit regularization of ordinary least squares ensembles. In *International Conference on Artificial Intelligence and Statistics*, pp. 3525–3535. PMLR, 2020.

Lin, L. and Dobriban, E. What Causes the Test Error? Going Beyond Bias-Variance via ANOVA. *Journal of Machine Learning Research*, 22(155):1–82, 2021.

Loureiro, B., Gerbelot, C., Cui, H., Goldt, S., Krzakala, F., Mézard, M., and Zdeborová, L. Capturing the learning curves of generic features maps for realistic data sets with a teacher-student model. *arXiv:2102.08127*, 2021a.

Loureiro, B., Sicuro, G., Gerbelot, C., Pacco, A., Krzakala, F., and Zdeborová, L. Learning gaussian mixtures with generalised linear models: Precise asymptotics in high-dimensions. *arXiv:2106.03791*, 2021b.

Mei, S. and Montanari, A. The generalization error of random features regression: Precise asymptotics and the double descent curve. *Communications on Pure and Applied Mathematics*, 2021.

Narkhede, M. V., Bartakke, P. P., and Sutaone, M. S. A review on weight initialization strategies for neural networks. *Artificial intelligence review*, pp. 1–32, 2021.

Neal, B., Mittal, S., Baratin, A., Tantia, V., Scicluna, M., Lacoste-Julien, S., and Mitliagkas, I. A modern take on the bias-variance tradeoff in neural networks. *arXiv preprint arXiv:1810.08591*, 2018.

Opitz, D. and Maclin, R. Popular ensemble methods: An empirical study. *Journal of Artificial Intelligence Research*, 11:169–198, 8 1999.

Parikh, N. and Boyd, S. Proximal algorithms. *Foundations and Trends in optimization*, 1(3):127–239, 2014.

Pennington, J. and Worah, P. Nonlinear random matrix theory for deep learning. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

Perrone, M. Putting it all together: Methods for combining neural networks. In Cowan, J., Tesauro, G., and Alspector, J. (eds.), *Advances in Neural Information Processing Systems*, volume 6. Morgan-Kaufmann, 1994.

Perrone, M. P. and Cooper, L. N. When networks disagree: Ensemble methods for hybrid neural networks. pp. 126–142. Chapman and Hall, 1993.

Rahimi, A. and Recht, B. Random Features for Large-Scale Kernel Machines. In *NIPS*, pp. 1177–1184, 2007.

Rosset, S., Zhu, J., and Hastie, T. Margin maximizing loss functions. In Thrun, S., Saul, L., and Schölkopf, B. (eds.), *Advances in Neural Information Processing Systems*, volume 16. MIT Press, 2004.

Schapire, R. E. The strength of weak learnability. *Machine learning*, 5(2):197–227, 1990.

Schwarze, H. and Hertz, J. Generalization in a large committee machine. *Europhysics Letters (EPL)*, 20(4):375–380, oct 1992.

Spigler, S., Geiger, M., d'Ascoli, S., Sagun, L., Biroli, G., and Wyart, M. A jamming transition from under- to over-parametrization affects generalization in deep learning. *Journal of Physics A: Mathematical and Theoretical*, 52 (47):474001, oct 2019.

Thrampoulidis, C., Abbasi, E., and Hassibi, B. Precise error analysis of regularized m-estimators in high dimensions. *IEEE Transactions on Information Theory*, 64(8):5592–5628, 2018.

Vershynin, R. Introduction to the non-asymptotic analysis of random matrices. *arXiv:1011.3027*, 2010.

Vershynin, R. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

Zdeborová, L. and Krzakala, F. Statistical physics of inference: Thresholds and algorithms. *Advances in Physics*, 65(5):453–552, 2016.

Appendix

A. The replica approach

A.1. Notation

We introduce here some notation that will help us to keep the expressions in this appendix more compact. Given two tensors $\mathbf{A} = (A^{ij}_{kk'})^{ij}_{kk'} \in \mathbb{R}^{p\times p} \otimes \mathbb{R}^{K\times K}$ and $\mathbf{B} = (B^{ij}_{kk'})^{ij}_{kk'} \in \mathbb{R}^{p\times p} \otimes \mathbb{R}^{K\times K}$, $i, j \in [p]$, $k, k' \in [K]$, then

$$\hat{\mathbf{C}} = \mathbf{AB} \Leftrightarrow \hat{C}^{ij}_{kk'} \coloneqq \sum_{r,\kappa} A^{ir}_{k\kappa} B^{rj}_{\kappa k'} \in \mathbb{R}^{p\times p} \otimes \mathbb{R}^{K\times K} \tag{36}$$

$$\mathbf{C} = \mathbf{A} \odot \mathbf{B} \Leftrightarrow C^{ij}_{kk'} \coloneqq A^{ij}_{kk'} B^{ij}_{kk'} \in \mathbb{R}^{p\times p} \otimes \mathbb{R}^{K\times K} \tag{37}$$

$$\tilde{\mathbf{C}} = \mathbf{A} \circ \mathbf{B} \Leftrightarrow \tilde{C}^{ij}_{kk'} \coloneqq \sum_{r} A^{ir}_{kk'} B^{rj}_{k'k} \in \mathbb{R}^{p\times p} \otimes \mathbb{R}^{K\times K}. \tag{38}$$

Also, if $\boldsymbol{X} = (X^i_k)^i_k \in \mathbb{R}^{p\times K}$ and $\boldsymbol{Y} = (Y^i_k)^i_k \in \mathbb{R}^{p\times K}$, we write

$$\boldsymbol{X} \otimes \boldsymbol{Y} = (X^i_k Y^j_{k'})^{ij}_{kk'} \in \mathbb{R}^{p\times p} \otimes \mathbb{R}^{K\times K} \tag{39}$$

$$\mathbf{A}|\boldsymbol{X}\rangle = \left(\textstyle\sum_{j\kappa} A^{ij}_{k\kappa} X^j_\kappa\right)^i_k \in \mathbb{R}^{p\times K} \tag{40}$$

$$\langle \boldsymbol{X}|\boldsymbol{Y}\rangle = \textstyle\sum_{ik} X^i_k Y^i_k \in \mathbb{R} \tag{41}$$

$$\langle \boldsymbol{X}|\mathbf{A}|\boldsymbol{Y}\rangle = \textstyle\sum_{ijkk'} X^i_k A^{ij}_{kk'} Y^j_{k'} \in \mathbb{R}, \tag{42}$$

$$\langle\!\langle \boldsymbol{X}|\boldsymbol{Y}\rangle\!\rangle = \left(\textstyle\sum_{ij} X^i_k Y^j_k\right)_k \in \mathbb{R}^K \tag{43}$$

$$\langle\!\langle \boldsymbol{X}|\mathbf{A}|\boldsymbol{Y}\rangle\!\rangle = \left(\textstyle\sum_{ij} X^i_k A^{ij}_{kk'} Y^j_{k'}\right)_{kk'} \in \mathbb{R}^{K\times K}, \tag{44}$$

$$\langle\!\langle \mathbf{A}\rangle\!\rangle = \left(\textstyle\sum_{i} A^{ii}_{kk'}\right)_{kk'} \in \mathbb{R}^{K\times K}. \tag{45}$$

In other words, the double brakets $\langle\!\langle\bullet\rangle\!\rangle$ express the contraction of the upper indices only. This means for example that $\langle\!\langle \boldsymbol{X}|\boldsymbol{Y}\rangle\!\rangle = \mathrm{diag}(\boldsymbol{X}^\intercal \boldsymbol{Y}) \in \mathbb{R}^K$. Finally, if $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^K$ and $\boldsymbol{A} \in \mathbb{R}^{K\times K}$, $\langle \boldsymbol{u}|\boldsymbol{A}|\boldsymbol{v}\rangle \coloneqq \boldsymbol{u}^\intercal \boldsymbol{A}\boldsymbol{v} = \sum_{kk'} u_k A_{kk'} v_{k'} \in \mathbb{R}$. We will adopt the same notation in the simple $K = 1$ case.

A.2. The replica computation

The replica computation relies on the treatment of a Gibbs measure over the weights $\boldsymbol{W}$ which concentrates on the weights $\hat{\boldsymbol{W}}$ that minimize a certain loss $\hat{\ell}$ when a fictitious 'inverse temperature' parameter is sent to infinity. Such measure reads

$$\mu_\beta(\boldsymbol{W}) = \frac{P_{\boldsymbol{W}}(\boldsymbol{W})}{\mathcal{Z}(\beta)} \prod_{\mu=1}^{n} \exp\left[-\beta\hat{\ell}\left(y^\mu, \frac{\langle\!\langle \boldsymbol{W}|\boldsymbol{U}^\mu\rangle\!\rangle}{\sqrt{p}}\right)\right], \quad \mathcal{Z}(\beta) \coloneqq \int \mathrm{d}\boldsymbol{W} P_{\boldsymbol{W}}(\boldsymbol{W}) \prod_{\mu=1}^{n} \exp\left[-\beta\hat{\ell}\left(y^\mu, \frac{\langle\!\langle \boldsymbol{W}|\boldsymbol{U}^\mu\rangle\!\rangle}{\sqrt{p}}\right)\right], \tag{46}$$

where $P_{\boldsymbol{W}}(\boldsymbol{W}) = \mathrm{e}^{-\beta\lambda r(\boldsymbol{W})}$ is the prior on the weights $\boldsymbol{W} = (W^i_k)^{i\in[p]}_{k\in[K]} \in \mathbb{R}^{p\times K}$, possibly containing the regularisation. The dataset $(y^\mu, \boldsymbol{U}^\mu)_\mu$ is obtained from a set of n samples $\boldsymbol{x}^\mu \sim \mathcal{N}(\mathbf{0}_d, \boldsymbol{I}_d)$, $\mu \in [n]$. For each μ, the label y^μ has distribution $P^0_y\left(y|d^{-1/2}\langle\boldsymbol{\theta}|\boldsymbol{x}^\mu\rangle\right)$ for some fixed $\boldsymbol{\theta} \sim \mathcal{N}(\mathbf{0}_d, \rho\boldsymbol{I}_d)$. The array of features $\boldsymbol{U}^\mu$, instead, is obtained as function of the vector $\boldsymbol{x}^\mu$ via a law $\boldsymbol{U}\colon \mathbb{R}^d \to \mathbb{R}^{p\times K}$ such that $\boldsymbol{U}^\mu \coloneqq \boldsymbol{U}(\boldsymbol{x}^\mu) \in \mathbb{R}^{p\times K}$. As we will show below, the tensors

$$\boldsymbol{\Omega} \coloneqq \mathbb{E}_{\boldsymbol{x}}[\boldsymbol{U}(\boldsymbol{x}) \otimes \boldsymbol{U}(\boldsymbol{x})] \in \mathbb{R}^{p\times p} \otimes \mathbb{R}^{K\times K}, \tag{47}$$

$$\hat{\boldsymbol{\Phi}} \coloneqq \mathbb{E}_{\boldsymbol{x}}[\boldsymbol{U}(\boldsymbol{x})\langle\boldsymbol{x}|\boldsymbol{\theta}\rangle] \in \mathbb{R}^{p\times K}, \tag{48}$$

will incorporate the information about the action of the law $\boldsymbol{U}$. We denote for brevity

$$P_y(y|\boldsymbol{u}) \propto \exp\left[-\beta\hat{\ell}(y, \boldsymbol{u})\right], \tag{49}$$

and we proceed computing the *free entropy* $\Phi \coloneqq \mathbb{E}_{(y^\mu,\boldsymbol{x}^\mu)_\mu}[\ln \mathcal{Z}(\beta)]$ using the replica trick, i.e., the fact that $\mathbb{E}[\ln \mathcal{Z}(\beta)] = \lim_{s\to 0} \frac{1}{s} \ln \mathbb{E}[\mathcal{Z}^s(\beta)]$

$$\mathbb{E}_{(y^\mu,\boldsymbol{x}^\mu)_\mu}[\mathcal{Z}^s(\beta)] = \prod_{a=1}^{s} \int \mathrm{d}\boldsymbol{W}^a P_{\boldsymbol{W}}(\boldsymbol{W}^a) \left(\mathbb{E}_{(y,\boldsymbol{x})}\left[P^0_y\left(y\middle|\frac{\langle\boldsymbol{x}|\boldsymbol{\theta}\rangle}{\sqrt{d}}\right) \prod_{a=1}^{s} P_y\left(y\middle|\frac{\langle\!\langle \boldsymbol{W}^a|\boldsymbol{U}(\boldsymbol{x})\rangle\!\rangle}{\sqrt{p}}\right)\right]\right)^n. \tag{50}$$

Denoting by $\boldsymbol{\mu}^a \equiv (\mu_k^a)_{k\in[K]}$, if we now consider

$$\mathbb{E}_{\boldsymbol{x}}\left[P_y^0\left(y\middle|\frac{\langle \boldsymbol{x}|\boldsymbol{\theta}\rangle}{\sqrt{d}}\right)\prod_{a=1}^{s}P_y\left(y\middle|\frac{\langle\!\langle \boldsymbol{W}^a|\boldsymbol{U}(\boldsymbol{x})\rangle\!\rangle}{\sqrt{d}}\right)\right]$$
$$=\int \mathrm{d}\nu P_y^0(y|\nu)\prod_{a=1}^{s}\left[\int \mathrm{d}\boldsymbol{\mu}^a P_y(y|\boldsymbol{\mu}^a)\right]\underbrace{\mathbb{E}_{\boldsymbol{x}}\left[\delta\left(\nu-\frac{\langle \boldsymbol{x}|\boldsymbol{\theta}\rangle}{\sqrt{d}}\right)\prod_{a=1}^{s}\delta\left(\boldsymbol{\mu}^a-\frac{\langle\!\langle \boldsymbol{W}^a|\boldsymbol{U}(\boldsymbol{x})\rangle\!\rangle}{\sqrt{p}}\right)\right]}_{P(\nu,\boldsymbol{\mu})}. \quad (51)$$

We apply now the *Gaussian Equivalence Principle* (Goldt et al., 2021), i.e., we assume that $P(\nu,\boldsymbol{\mu})$ is a Gaussian with covariance matrix given by

$$\boldsymbol{\Sigma}(\boldsymbol{W})=\begin{pmatrix}\rho & \mathbf{m}^\intercal\\ \mathbf{m} & \mathbf{Q}\end{pmatrix}, \quad (52)$$

where $\mathbf{m}=(\boldsymbol{m}^a)_{a\in[s]}\in\mathbb{R}^{sK}$ and $\mathbf{Q}=(\boldsymbol{Q}^{ab})_{a,b\in[s]}\in\mathbb{R}^{sK\times sK}$. Here, for each $a,b\in[n]$, $\boldsymbol{m}^a\in\mathbb{R}^K$ and $\boldsymbol{Q}^{ab}\in\mathbb{R}^{K\times K}$ and are defined as

$$\rho \coloneqq \mathbb{E}[\nu^2]=\frac{\|\boldsymbol{\theta}\|_2^2}{d}, \quad (53)$$
$$\boldsymbol{m}^a \coloneqq \mathbb{E}[\boldsymbol{\mu}^a\nu]=\frac{\langle\!\langle \boldsymbol{W}^a|\hat{\boldsymbol{\Phi}}\rangle\!\rangle}{\sqrt{pd}}, \quad (54)$$
$$\boldsymbol{Q}^{ab} \coloneqq \mathbb{E}[\boldsymbol{\mu}^a\boldsymbol{\mu}^{b\intercal}]=\frac{\langle\!\langle \boldsymbol{W}^a|\boldsymbol{\Omega}|\boldsymbol{W}^b\rangle\!\rangle}{p}. \quad (55)$$

In the end

$$\mathbb{E}_{(y^\mu,\boldsymbol{x}^\mu)_\mu}[\mathcal{Z}^s(\beta)]=\mathbb{E}_{\boldsymbol{\theta}}\left[\prod_{a=1}^{s}\int \mathrm{d}\boldsymbol{W}^a P_{\boldsymbol{W}}(\boldsymbol{W}^a)\left(\int \mathrm{d}y\int \mathrm{d}\nu P_y^0(y|\nu)\prod_{a=1}^{s}\left[\int \mathrm{d}\boldsymbol{\mu}^a P_y(y|\boldsymbol{\mu}^a)\right]P(\nu,\boldsymbol{\mu})\right)^n\right]$$
$$=\left(\prod_{a=1}^{s}\iint\frac{\mathrm{d}\mathbf{m}^a\mathrm{d}\hat{\mathbf{m}}^a}{(2\pi)^K}\right)\left(\prod_{a=1}^{s}\iint\frac{\mathrm{d}\mathbf{Q}^{ab}\mathrm{d}\hat{\mathbf{Q}}^{ab}}{(2\pi)^{K^2}}\right)\int\frac{\mathrm{d}\rho\mathrm{d}\hat{\rho}}{2\pi}\exp\left(p\Phi^{(s)}(\rho,\mathbf{m},\mathbf{Q},\hat{\rho},\hat{\mathbf{m}},\hat{\mathbf{Q}})\right). \quad (56)$$

Absorbing the factor $-i$ in the integrals, and denoting by $n/p=\alpha$ and $d/p=\gamma$,

$$\Phi^{(s)}(\rho,\mathbf{m},\mathbf{Q},\hat{\rho},\hat{\mathbf{m}},\hat{\mathbf{Q}})=-\gamma\rho\hat{\rho}-\sqrt{\gamma}\sum_{a=1}^{s}\langle\hat{\boldsymbol{m}}^a|\boldsymbol{m}^a\rangle-\sum_{a\le b}\langle\hat{\boldsymbol{Q}}^{ab}|\boldsymbol{Q}^{ab}\rangle+\alpha\Psi_y^{(s)}(\rho,\mathbf{m},\mathbf{Q})+\Psi_{\boldsymbol{w}}^{(s)}(\hat{\rho},\hat{\mathbf{m}},\hat{\mathbf{Q}}). \quad (57)$$

Here we have introduced

$$\Psi_y^{(s)}(\rho,\mathbf{m},\mathbf{Q}) \coloneqq \ln\left[\int \mathrm{d}y\int \mathrm{d}\nu P_y^0(y|\nu)\prod_{a=1}^{s}\left[\int \mathrm{d}\boldsymbol{\mu}^a P_y(y|\boldsymbol{\mu}^a)\right]P(\nu,\boldsymbol{\mu})\right] \quad (58)$$
$$\Psi_{\boldsymbol{w}}^{(s)}(\hat{\rho},\hat{\mathbf{m}},\hat{\mathbf{Q}}) \coloneqq \frac{1}{p}\ln\left[\mathrm{e}^{\hat{\rho}\|\boldsymbol{\theta}\|_2^2}\left(\prod_{a=1}^{s}\int P(\boldsymbol{W}^a)\,\mathrm{e}^{\langle(\mathbf{1}_p\otimes\hat{\boldsymbol{m}}^{a\intercal})\odot\hat{\boldsymbol{\Phi}})|\boldsymbol{W}^a\rangle}\right)\exp\left(\sum_{a\le b}\langle\boldsymbol{W}^a|(\mathbf{1}_{p,p}\otimes\hat{\boldsymbol{Q}}^{ab})\odot\boldsymbol{\Omega}|\boldsymbol{W}^b\rangle\right)\right] \quad (59)$$

so that in the high dimensional limit the desired average is the extremum of the functional $1s\Phi^{(s)}$ in the $s\to 0$ limit,

$$\mathbb{E}_{(y^\mu,\boldsymbol{x}^\mu)_\mu}[\ln\mathcal{Z}(\beta)]=\lim_{s\to 0}\frac{1}{s}\mathrm{Ext}\,\Phi^{(s)}(\rho,\mathbf{m},\mathbf{Q},\hat{\rho},\hat{\mathbf{m}},\hat{\mathbf{Q}}). \quad (60)$$

Replica symmetric ansatz – In order to take the limit, let us assume as usual a *replica symmetric* (RS) ansatz:

$$\boldsymbol{m}^a\equiv\boldsymbol{m}\quad a\in[s],\qquad \hat{\boldsymbol{m}}^a\equiv\hat{\boldsymbol{m}}\quad a\in[s],$$
$$\boldsymbol{Q}^{ab}\equiv\begin{cases}\boldsymbol{R} & \text{if } a=b,\\ \boldsymbol{Q} & \text{if } a\neq b,\end{cases}\qquad \hat{\boldsymbol{Q}}^{ab}\equiv\begin{cases}-\frac{1}{2}\hat{\boldsymbol{R}} & \text{if } a=b,\\ \hat{\boldsymbol{Q}} & \text{if } a\neq b.\end{cases} \quad (61)$$

Observe that $\lim_{s\to 0}\Phi^{(s)}=0$ by construction, meaning that $\hat{\rho}=0$ fixing $\rho=\frac{1}{d}\mathbb{E}_{\boldsymbol{\theta}}[\|\boldsymbol{\theta}\|_2^2]$. Before proceeding further, we note that the matrix $\mathbf{Q}$ in the RS ansatz can be written as $\mathbf{Q}=\boldsymbol{I}_s\otimes(\boldsymbol{R}-\boldsymbol{Q})+\mathbf{1}_{s,s}\otimes\boldsymbol{Q}$, where $\mathbf{1}_{s,s}$ is the $s\times s$ matrix of 1. Similarly, $\mathbf{m}=\mathbf{1}_s\otimes\boldsymbol{m}$, where $\mathbf{1}_s$ is the column vector of s elements equal to 1. Following similar steps to the ones detailed, e.g., in (Loureiro et al., 2021b), we obtain

$$\Psi_y(\rho,\boldsymbol{m},\boldsymbol{Q},\boldsymbol{R}) \coloneqq \lim_{s\to 0}\frac{1}{s}\Psi_y^{(s)}(\rho,\mathbf{m},\mathbf{Q}) = \int_{\mathcal{Y}}\mathrm{d}y\,\mathbb{E}_{\boldsymbol{\xi}}\left[\mathcal{Z}^0\Big(y,\langle\boldsymbol{m}|\boldsymbol{Q}^{-1/2}|\boldsymbol{\xi}\rangle,\rho-\langle\boldsymbol{m}|\boldsymbol{Q}^{-1}|\boldsymbol{m}\rangle\Big)\ln\mathcal{Z}\left(y,\sqrt{\boldsymbol{Q}}\boldsymbol{\xi},\boldsymbol{V}\right)\right], \tag{62}$$

where $\boldsymbol{V}\coloneqq\boldsymbol{R}-\boldsymbol{Q}$, $\boldsymbol{\xi}\sim\mathcal{N}(\mathbf{0}_K,\boldsymbol{I}_K)$ and we have introduced

$$\mathcal{Z}(y,\boldsymbol{\mu},\boldsymbol{\Sigma}) \coloneqq \int\frac{P_y(y|\boldsymbol{\mu})\mathrm{d}\boldsymbol{\mu}}{\sqrt{\det(2\pi\boldsymbol{\Sigma})}}\,\mathrm{e}^{-\frac{1}{2}\langle\boldsymbol{u}-\boldsymbol{\mu}|\boldsymbol{\Sigma}^{-1}|\boldsymbol{u}-\boldsymbol{\mu}\rangle}, \tag{63}$$

$$\mathcal{Z}^0(y,\mu,\sigma) \coloneqq \int\frac{P_y^0(y|x)\mathrm{d}x}{\sqrt{2\pi\sigma}}\,\mathrm{e}^{-\frac{(x-\mu)^2}{2\sigma}}. \tag{64}$$

Similarly, defining $\hat{\boldsymbol{V}}=\hat{\boldsymbol{R}}+\hat{\boldsymbol{Q}}$, we can write down the prior channel. We can write then

$$\begin{aligned}\Psi_{\boldsymbol{w}}(\hat{\boldsymbol{m}},\hat{\boldsymbol{Q}},\hat{\boldsymbol{R}}) &\coloneqq \lim_{s\to 0}\frac{1}{s}\Psi_{\boldsymbol{w}}^{(s)}(0,\hat{\mathbf{m}},\hat{\mathbf{Q}})\\ &= \lim_{s\to 0}\frac{1}{sp}\ln\left[\left(\prod_{a=1}^{s}\int\mathrm{d}\boldsymbol{W}^a P_{\boldsymbol{W}}(\boldsymbol{W}^a)\,\mathrm{e}^{\langle(\mathbf{1}_p\otimes\hat{\boldsymbol{m}}^\intercal)\odot\hat{\boldsymbol{\Phi}}|\boldsymbol{W}^a\rangle-\frac{1}{2}\langle\boldsymbol{W}^a|(\mathbf{1}_{p,p}\otimes\hat{\boldsymbol{R}})\odot\boldsymbol{\Omega}|\boldsymbol{W}^a\rangle}\right)\prod_{a<b}\mathrm{e}^{\langle\boldsymbol{W}^a|(\mathbf{1}_{p,p}\otimes\hat{\boldsymbol{Q}})\odot\boldsymbol{\Omega})^{1/2}|\boldsymbol{W}^b\rangle}\right]\\ &= \frac{1}{p}\mathbb{E}_{\boldsymbol{\Xi}}\left[\ln\left(\int\mathrm{d}\boldsymbol{W}\,\mathrm{e}^{-\lambda\beta r(\boldsymbol{W})+\langle(\mathbf{1}_p\otimes\hat{\boldsymbol{m}}^\intercal)\odot\hat{\boldsymbol{\Phi}}|\boldsymbol{W}\rangle+\langle\boldsymbol{\Xi}|(\mathbf{1}_{p,p}\otimes\hat{\boldsymbol{Q}})\odot\boldsymbol{\Omega})^{1/2}|\boldsymbol{W}\rangle-\frac{1}{2}\langle\boldsymbol{W}|(\mathbf{1}_{p,p}\otimes\hat{\boldsymbol{V}})\odot\boldsymbol{\Omega}|\boldsymbol{W}\rangle}\right)\right]\end{aligned} \tag{65}$$

where we have performed a Hubbard-Stratonovich transformation and $\boldsymbol{\Xi}\equiv(\Xi_k^i)_k^i\in\mathbb{R}^{p\times K}$ has $\Xi_k^i\sim\mathcal{N}(0,1)$ for all i,k. The free entropy is then

$$\Phi \coloneqq \lim_{s\to 0}\frac{1}{s}\Phi^{(s)} = -\sqrt{\gamma}\langle\hat{\boldsymbol{m}}|\boldsymbol{m}\rangle + \frac{\langle\hat{\boldsymbol{V}}|\boldsymbol{V}\rangle+\langle\hat{\boldsymbol{V}}|\boldsymbol{Q}\rangle-\langle\hat{\boldsymbol{Q}}|\boldsymbol{V}\rangle}{2} + \alpha\Psi_y(\rho,\boldsymbol{m},\boldsymbol{Q},\boldsymbol{V}) + \Psi_{\boldsymbol{w}}(\hat{\boldsymbol{m}},\hat{\boldsymbol{Q}},\hat{\boldsymbol{V}}). \tag{66}$$

We are interested in the extremum of this quantity, and therefore we have to find the order parameters that maximise it by means of a set of saddle-point equations. Defining for brevity

$$\boldsymbol{\omega}=\boldsymbol{Q}^{1/2}\boldsymbol{\xi},\qquad \omega_0=\langle\boldsymbol{m}|\boldsymbol{Q}^{-1}|\boldsymbol{\omega}\rangle,\qquad \sigma_0=\rho-\langle\boldsymbol{m}|\boldsymbol{Q}^{-1}|\boldsymbol{m}\rangle, \tag{67}$$

a first set of saddle-point equation is

$$\hat{\boldsymbol{V}} = -\alpha\int_{\mathcal{Y}}\mathrm{d}y\,\mathbb{E}_{\boldsymbol{\xi}}\left[\mathcal{Z}^0(y,\omega_0,\sigma_0)\partial_{\boldsymbol{\omega}}\boldsymbol{f}\right], \tag{68a}$$

$$\hat{\boldsymbol{Q}} = \alpha\int_{\mathcal{Y}}\mathrm{d}y\,\mathbb{E}_{\boldsymbol{\xi}}\left[\mathcal{Z}^0(y,\omega_0,\sigma_0)\boldsymbol{f}\boldsymbol{f}^\intercal\right], \tag{68b}$$

$$\hat{\boldsymbol{m}} = \frac{\alpha}{\sqrt{\gamma}}\int_{\mathcal{Y}}\mathrm{d}y\,\mathbb{E}_{\boldsymbol{\xi}}\left[\mathcal{Z}^0(y,\omega_0,\sigma_0)f^0\boldsymbol{f}\right] \tag{68c}$$

where

$$f^0 \equiv \partial_{\omega_0}\ln\mathcal{Z}^0(y,\omega_0,\sigma_0) \tag{69}$$

$$\boldsymbol{f} \equiv \partial_{\boldsymbol{\omega}}\ln\mathcal{Z}(y,\boldsymbol{\omega},\boldsymbol{V}). \tag{70}$$

Zero temperature state evolution equations – To obtain a nontrivial $\beta\to+\infty$ limit we rescale $\hat{\boldsymbol{V}}\mapsto\beta\hat{\boldsymbol{V}}$, $\boldsymbol{V}\mapsto\beta^{-1}\boldsymbol{V}$, $\hat{\boldsymbol{Q}}\mapsto\beta^2\hat{\boldsymbol{Q}}$, $\hat{\boldsymbol{m}}\mapsto\beta\hat{\boldsymbol{m}}$. After this change of variable, Eqs. (79) remain formally identical. To complete the set of saddle-point equations, let us observe that, defining

$$\mathcal{L}(y,\boldsymbol{u}) = \frac{1}{2}\langle\boldsymbol{u}-\boldsymbol{\omega}|\boldsymbol{V}^{-1}|\boldsymbol{u}-\boldsymbol{\omega}\rangle + \hat{\ell}(y,\boldsymbol{u}) \tag{71}$$

then after the rescaling

$$\ln \mathcal{Z}(y, \boldsymbol{\omega}, \beta^{-1}\boldsymbol{V}) = \ln \int \frac{\mathrm{e}^{-\beta \mathcal{L}(y,\boldsymbol{u})}\, \mathrm{d}\boldsymbol{u}}{\sqrt{\det(2\pi \boldsymbol{V})}} \xrightarrow{\beta \gg 1} -\beta \mathcal{L}(y, \boldsymbol{h}) \quad \text{with } \boldsymbol{h} = \arg\min_{\boldsymbol{u}} \mathcal{L}(y, \boldsymbol{u}). \tag{72}$$

In this way the remaining saddle-point equations keep the form (68) but with

$$\boldsymbol{f} \coloneqq \boldsymbol{V}^{-1}(\boldsymbol{h} - \boldsymbol{\omega}). \tag{73}$$

In the $\beta \to +\infty$ limit, we can write also

$$\begin{aligned}\Psi_{\boldsymbol{w}}(\hat{\boldsymbol{m}}, \hat{\boldsymbol{Q}}, \hat{\boldsymbol{R}}) &= \frac{1}{p}\mathbb{E}_{\boldsymbol{\Xi}}\left[\ln\left(\int \mathrm{d}\boldsymbol{W}\, \mathrm{e}^{-\lambda\beta r(\boldsymbol{W}) + \beta\langle \boldsymbol{B}|\boldsymbol{W}\rangle - \frac{\beta}{2}\langle \boldsymbol{W}|(\mathbf{1}_{p,p}\otimes\hat{\boldsymbol{V}})\odot\boldsymbol{\Omega}|\boldsymbol{W}\rangle}\right)\right] \\ &= -\frac{\beta}{p}\mathbb{E}_{\boldsymbol{\Xi}}\left[\frac{\langle \boldsymbol{G}|(\mathbf{1}_{p,p}\otimes\hat{\boldsymbol{V}})\odot\boldsymbol{\Omega}|\boldsymbol{G}\rangle}{2} - \langle \boldsymbol{B}|\boldsymbol{G}\rangle + \lambda r(\boldsymbol{G})\right]\end{aligned} \tag{74}$$

where

$$\boldsymbol{B} \coloneqq (\mathbf{1}_p \otimes \hat{\boldsymbol{m}}^{\intercal}) \odot \hat{\boldsymbol{\Phi}} + \left((\mathbf{1}_{p,p}\otimes\hat{\boldsymbol{Q}})\odot\boldsymbol{\Omega}\right)^{1/2}\boldsymbol{\Xi} \tag{75}$$

and

$$\boldsymbol{G} \coloneqq \arg\min_{\boldsymbol{U}}\left[\frac{\langle \boldsymbol{U}|(\mathbf{1}_{p,p}\otimes\hat{\boldsymbol{V}})\odot\boldsymbol{\Omega}|\boldsymbol{U}\rangle}{2} - \langle \boldsymbol{B}|\boldsymbol{U}\rangle + \lambda r(\boldsymbol{U})\right]. \tag{76}$$

As a result, the remaining saddle point equations are

$$\boldsymbol{V} = \frac{2}{p}\mathbb{E}_{\boldsymbol{\Xi}}\left\langle \boldsymbol{G}\middle|\frac{\mathrm{D}\left((\mathbf{1}_{p,p}\otimes\hat{\boldsymbol{Q}})\odot\boldsymbol{\Omega}\right)^{1/2}}{\mathrm{D}\hat{\boldsymbol{Q}}}\middle|\boldsymbol{\Xi}\right\rangle \tag{77a}$$

$$\boldsymbol{Q} = \frac{1}{p}\mathbb{E}_{\boldsymbol{\Xi}}\langle\!\langle \boldsymbol{G}|\boldsymbol{\Omega}|\boldsymbol{G}\rangle\!\rangle, \tag{77b}$$

$$\boldsymbol{m} = \frac{1}{\sqrt{\gamma}p}\mathbb{E}_{\boldsymbol{\Xi}}\langle\!\langle \hat{\boldsymbol{\Phi}}|\boldsymbol{G}\rangle\!\rangle. \tag{77c}$$

In the first equation, the derivative produce in general a 6-index tensor. Denoting $\hat{\mathbf{A}} \in \mathbb{R}^{p\times p}\otimes\mathbb{R}^{K\times K}$, then $[\frac{\mathrm{D}\hat{\mathbf{A}}}{\mathrm{D}\hat{\boldsymbol{Q}}}]_{ij}^{kk',\kappa\kappa'} \equiv \frac{\partial \hat{A}_{ij}^{kk'}}{\partial \hat{Q}_{\kappa\kappa'}}$. In the formula, the contractions involve latin indices only.

The case of ℓ_2 regularisation – If we assume an $\hat{\ell}_2$ regularisation $r(\boldsymbol{W}) = \frac{1}{2}\|\boldsymbol{W}\|_{\mathrm{F}}^2$, $\Psi_{\boldsymbol{w}}$ is a Gaussian integral that can be explicitly computed before the rescaling in β. Denoting

$$\mathbf{A} \coloneqq \left[\lambda \boldsymbol{I}_K \otimes \boldsymbol{I}_p + (\mathbf{1}_{p,p}\otimes\hat{\boldsymbol{V}})\odot\boldsymbol{\Omega}\right]^{-1} \quad \text{and} \quad \boldsymbol{\Theta} \coloneqq \hat{\boldsymbol{\Phi}}\otimes\hat{\boldsymbol{\Phi}} \in \mathbb{R}^{p\times p}\otimes\mathbb{R}^{K\times K} \tag{78}$$

we obtain the following saddle-point equations for $\boldsymbol{V}$, $\boldsymbol{Q}$ and $\boldsymbol{m}$,

$$\boldsymbol{V} = \frac{\langle\!\langle \mathbf{A}\circ\boldsymbol{\Omega}\rangle\!\rangle}{p}, \tag{79a}$$

$$\boldsymbol{Q} = \frac{1}{p}\langle\!\langle \boldsymbol{\Omega}\circ\left(\mathbf{A}\left((\mathbf{1}_{p,p}\otimes\hat{\boldsymbol{m}}\otimes\hat{\boldsymbol{m}}^{\intercal})\odot\boldsymbol{\Theta} + (\mathbf{1}_{p,p}\otimes\hat{\boldsymbol{Q}})\odot\boldsymbol{\Omega}\right)\mathbf{A}\right)\rangle\!\rangle \tag{79b}$$

$$\boldsymbol{m} = \frac{1}{\sqrt{\gamma}p}\langle\!\langle \mathbf{A}\left((\mathbf{1}_{p,p}\otimes\hat{\boldsymbol{m}}\otimes\mathbf{1}_K^{\intercal})\odot\boldsymbol{\Theta}\right)\rangle\!\rangle. \tag{79c}$$

Training loss and generalisation error – The order parameters introduced to solve the problem allow us to reach our ultimate goal of computing the average errors of the learning process. We have

$$\epsilon_{\hat{\ell}} \equiv \frac{1}{n}\sum_{\nu=1}^{n}\hat{\ell}\left(y^{\nu}, \frac{\langle\!\langle \hat{\boldsymbol{W}}|\boldsymbol{U}^{\nu}\rangle\!\rangle}{\sqrt{p}}\right) \xrightarrow{n\to+\infty} -\partial_{\beta}\Psi_y = \int \mathrm{d}y \mathbb{E}_{\boldsymbol{\xi}}\left[\frac{\mathcal{Z}^0(y,\omega_0,\sigma_0)}{\mathcal{Z}(y,\boldsymbol{\omega},\boldsymbol{V})}\int \frac{\hat{\ell}(y,\boldsymbol{u})\mathrm{e}^{-\frac{1}{2}\langle \boldsymbol{\omega}-\boldsymbol{\mu}|\boldsymbol{V}^{-1}|\boldsymbol{\omega}-\boldsymbol{u}\rangle - \beta\hat{\ell}(y,\boldsymbol{u})}\mathrm{d}\boldsymbol{u}}{\sqrt{\det(2\pi\boldsymbol{V})}}\right]$$
$$\xrightarrow[\boldsymbol{V}\mapsto\beta^{-1}\boldsymbol{V}]{\beta\to+\infty} \int \mathrm{d}y\, \mathbb{E}_{\boldsymbol{\xi}}\left[\mathcal{Z}^0(y,\omega_0,\sigma_0)\,\hat{\ell}(y,\boldsymbol{h})\right], \quad (80)$$

where $\boldsymbol{h}$ is the proximal introduced above and all overlaps have to be intended computed at the fixed point.

We can also study the generalisation error

$$\epsilon_g = \mathbb{E}_{(y,\boldsymbol{U})}\left[\Delta\left(y, \hat{y}\left(\frac{\langle\!\langle \hat{\boldsymbol{W}}|\boldsymbol{U}\rangle\!\rangle}{\sqrt{p}}\right)\right)\right] = \int \mathrm{d}\boldsymbol{\mu}\int \mathrm{d}\nu \int \mathrm{d}y \Delta(y,\hat{y}(\boldsymbol{\mu}))P_y^0(y|\nu)\mathbb{E}_{\boldsymbol{x}}\left[\delta\left(\boldsymbol{\mu} - \frac{\langle\!\langle \hat{\boldsymbol{W}}|\boldsymbol{U}\rangle\!\rangle}{\sqrt{p}}\right)\delta\left(\nu - \frac{\langle \boldsymbol{x}|\boldsymbol{\theta}\rangle}{\sqrt{d}}\right)\right]$$
$$= \int \mathbb{E}_{(\nu,\boldsymbol{\mu})}\left[\Delta\left(y,\hat{y}(\boldsymbol{\mu})\right)P_y^0(y|\nu)\right]\mathrm{d}y \quad (81)$$

where $(\nu, \boldsymbol{\mu})$ are jointly Gaussian with covariance

$$\boldsymbol{\Sigma} = \begin{pmatrix} \rho & \boldsymbol{m}^{\intercal} \\ \boldsymbol{m} & \boldsymbol{Q} \end{pmatrix} \quad (82)$$

In particular, if $P_y^0(y|\nu) = \delta(y - f_0(\nu))$, then $\epsilon_g = \mathbb{E}_{(\nu,\boldsymbol{\mu})}\left[\Delta\left(f_0(\nu), \hat{y}(\boldsymbol{\mu})\right)\right]$, which corresponds to (6).

A.3. Separable loss with ridge regularisation

Let us focus now on the case of separable losses, i.e., losses in the form $\hat{\ell}(y,\boldsymbol{u}) = \sum_k \ell(y,u_k)$, which is a crucial special case in the analysis of our contribution. We will assume a ridge regularisation $r(\boldsymbol{W}) = \frac{1}{2}\|\boldsymbol{W}\|_{\mathrm{F}}^2$. Let us also assume that the K generative networks are statistically equivalent. This implies a specific structure in the tensors $\boldsymbol{\Theta}$ and $\boldsymbol{\Omega}$,

$$\boldsymbol{\Omega}_{kk'} = \boldsymbol{\Omega}_{k'k}^{\intercal} \overset{\mathrm{d}}{=} \begin{cases} \boldsymbol{\Omega} & \text{for } k = k', \\ \hat{\boldsymbol{\Omega}} & \text{for } k < k', \end{cases} \quad (83)$$

$$\boldsymbol{\Theta}_{kk'} = \boldsymbol{\Theta}_{k'k}^{\intercal} \overset{\mathrm{d}}{=} \begin{cases} \boldsymbol{\Theta} & \text{for } k = k', \\ \hat{\boldsymbol{\Theta}} & \text{for } k < k'. \end{cases} \quad (84)$$

Here by $\overset{\mathrm{d}}{=}$ we mean that the equalities hold in distribution. Observe $\boldsymbol{\Omega}_{kk}$ and $\boldsymbol{\Omega}_{kk'}$ are not uncorrelated quantities. For reasons of symmetry reasons, we impose therefore the ansatz

$$\begin{aligned} \boldsymbol{V} &= v\boldsymbol{I}_K, & \hat{\boldsymbol{V}} &= \hat{v}\boldsymbol{I}_K, \\ \boldsymbol{m} &= m\mathbf{1}_K, & \hat{\boldsymbol{m}} &= \hat{m}\mathbf{1}_K, \\ \boldsymbol{Q} &= (q_0 - q_1)\boldsymbol{I}_K + q_1\mathbf{1}_{K,K}, & \hat{\boldsymbol{Q}} &= (\hat{q}_0 - \hat{q}_1)\boldsymbol{I}_K + \hat{q}_1\mathbf{1}_{K,K}. \end{aligned} \quad (85)$$

It is easily seen that

$$\boldsymbol{Q}^{1/2} = \sqrt{q_0 - q_1}\boldsymbol{I}_K + \frac{\sqrt{q_0 + (K-1)q_1} - \sqrt{q_0 - q_1}}{K}\mathbf{1}_{K,K}, \quad (86)$$

$$\mathbf{A}_{kk'} = (\lambda\boldsymbol{I}_p + \boldsymbol{\Omega}_{kk})^{-1}\delta_{kk'}. \quad (87)$$

Plugging this ansatz in our equations, and introducing $\eta = {}^{q_1}/_{q_0}$, we obtain

$$\begin{aligned}
\hat{v} &= -\alpha \int_{\mathcal{Y}} \mathrm{d}y\, \mathbb{E}_\zeta \left[\mathcal{Z}^0(y, \omega_0, \sigma_0)\, \partial_\omega f \right], \\
\hat{m} &= \frac{\alpha}{\sqrt{\gamma}} \int_{\mathcal{Y}} \mathrm{d}y\, \mathbb{E}_\zeta \left[f \partial_\mu \mathcal{Z}^0(y, \omega_0, \sigma_0) \right], \\
\hat{q}_0 &= \alpha \int_{\mathcal{Y}} \mathrm{d}y\, \mathbb{E}_\zeta \left[\mathcal{Z}^0(y, \omega_0, \sigma_0)\, f^2 \right], \\
\hat{q}_1 &= \alpha \mathbb{E}_{y,\zeta,\zeta'} \left[\mathcal{Z}^0\left(y, \frac{m}{\sqrt{q_0}} \frac{\zeta + \zeta'}{1+\eta}, \rho - \frac{2m^2}{q_0 + q_1} \right) f f' \right],
\end{aligned} \qquad \text{with } \omega_0 := \frac{m\zeta}{\sqrt{q_0}}, \quad \sigma_0 := \rho - \frac{m^2}{q_0}. \tag{88}$$

The new variables ζ_1 and ζ_2 are obtained by a linear transformation from the old ones. In particular, they are distributed as two components of a vector

$$\boldsymbol{\zeta} = \left(\sqrt{1-\eta}\, \boldsymbol{I}_K + \frac{\sqrt{1+(K-1)\eta} - \sqrt{1-\eta}}{K} \mathbf{1}_{K,K} \right) \boldsymbol{\xi}, \tag{89}$$

where $\boldsymbol{\xi} \sim \mathcal{N}(\mathbf{0}_K, \boldsymbol{I}_K)$. It follows that $\zeta, \zeta' \sim \mathcal{N}(0,1)$ but they are correlated as

$$\mathbb{E}[\zeta \zeta'] = \eta. \tag{90}$$

Moreover, we have introduced the proximal

$$f = \frac{h - \omega}{v} \qquad \text{where} \qquad h = \arg\min_x \left[\frac{(x - \sqrt{q_0}\zeta)^2}{2v} + \ell(y, x) \right] \tag{91}$$

and the corresponding f' obtained using ζ'. The remaining equations read

$$v = \frac{\mathrm{tr}\big[(\lambda \boldsymbol{I}_p + \hat{v} \boldsymbol{\Omega})^{-1} \boldsymbol{\Omega} \big]}{p}, \tag{92a}$$

$$m = \frac{\hat{m}}{\sqrt{\gamma}} \frac{\mathrm{tr}\big[\boldsymbol{\Theta} (\lambda \boldsymbol{I}_p + \hat{v} \boldsymbol{\Omega})^{-1} \big]}{p} \tag{92b}$$

$$q_0 = \frac{\mathrm{tr}\big[(\lambda \boldsymbol{I}_p + \hat{v} \boldsymbol{\Omega})^{-1} \left(\hat{m}^2 \boldsymbol{\Theta} + \hat{q}_0 \boldsymbol{\Omega} \right) (\lambda \boldsymbol{I}_p + \hat{v} \boldsymbol{\Omega})^{-1} \boldsymbol{\Omega} \big]}{p} \tag{92c}$$

$$q_1 = \frac{\mathrm{tr}\Big[(\lambda \boldsymbol{I}_p + \hat{v} \boldsymbol{\Omega})^{-1} \left(\hat{m}^2 \hat{\boldsymbol{\Theta}} + \hat{q}_1 \hat{\boldsymbol{\Omega}} \right) (\lambda \boldsymbol{I}_p + \hat{v} \boldsymbol{\Omega}')^{-1} \hat{\boldsymbol{\Omega}}^\intercal \Big]}{p}. \tag{92d}$$

As a side note, observe that in the case of *identical* learners, then $\boldsymbol{\Omega} = \hat{\boldsymbol{\Omega}} = \boldsymbol{\Omega}'$. In this case, the equations above are solved by $q \equiv q_0 = q_1 \Leftrightarrow \eta = 1$, as, under this assumption all components of $\boldsymbol{\zeta}$ are equal to $\frac{1}{\sqrt{K}} \sum_i \xi_i$, and therefore $\zeta = \zeta'$ and

$$\hat{q}_1 = \alpha \mathbb{E}_{y,\zeta} \left[\mathcal{Z}^0\left(y, \frac{m}{\sqrt{q}} \zeta, \rho - \frac{m^2}{q} \right) f^2 \right] \equiv \hat{q}_0 \equiv \hat{q}. \tag{93}$$

This express the simple fact that *identical* learners have identical performances (due to the convexity of the problem) given by the result in (Gerace et al., 2020), and the overlap between two of them is equal to the self-overlap of a single learner.

The random-features model for the generative networks – To further simplify these expressions suppose now that our generative networks are such that

$$\boldsymbol{u}_k(\boldsymbol{x}) = \phi\left(\frac{\boldsymbol{F}_k \boldsymbol{x}}{\sqrt{d}} \right), \quad k \in [K], \tag{94}$$

where $\boldsymbol{F}_k \in \mathbb{R}^{p \times d}$ are (fixed) random matrices extracted from some given distribution and ϕ is a nonlinearity acting elementwise. As anticipated in the main text, we can use the fact that each generative network is equivalent to the following Gaussian model (Mei & Montanari, 2021)

$$\boldsymbol{u}_k(\boldsymbol{x}) \mapsto \kappa_0 \mathbf{1}_p + \kappa_1 \frac{\boldsymbol{F}_k \boldsymbol{x}}{\sqrt{d}} + \kappa_* \boldsymbol{z}_k. \tag{95}$$

for some coefficients κ_0, κ_1 and κ_* depending on ϕ (see Therorem 4.3), and $\boldsymbol{z}^\mu \sim \mathcal{N}(\mathbf{0}_p, \boldsymbol{I}_p)$. Assuming now, for the sake of simplicity, to the $\kappa_0 = 0$ case and that $\mathbb{E}_{\boldsymbol{\theta}}[\boldsymbol{\theta}\boldsymbol{\theta}^\intercal] = \boldsymbol{I}_d$, then

$$\begin{aligned} \boldsymbol{\Omega} &\stackrel{\mathrm{d}}{=} \frac{\kappa_1^2}{d}\boldsymbol{F}^\intercal\boldsymbol{F} + \kappa_*^2\boldsymbol{I}_p, \\ \boldsymbol{\Theta} &\stackrel{\mathrm{d}}{=} \frac{\kappa_1^2}{d}\boldsymbol{F}^\intercal\boldsymbol{F}, \end{aligned} \qquad \hat{\boldsymbol{\Theta}} \stackrel{\mathrm{d}}{=} \hat{\boldsymbol{\Omega}} \stackrel{\mathrm{d}}{=} \frac{\kappa_1^2}{d}\boldsymbol{F}^\intercal\boldsymbol{F}', \quad \boldsymbol{F} \stackrel{\mathrm{d}}{=} \boldsymbol{F}' \stackrel{\mathrm{d}}{=} \boldsymbol{F}_k \ \forall k \in [K]. \tag{96}$$

Once the spectral density $\varrho(s)$ of $\boldsymbol{\Omega}$ is introduced, it is immediate to see that the equations for q_0, m and v take the forms given in the main text. The equation for q_1 requires an additional step. If we introduce the symmetric random matrix

$$\hat{\boldsymbol{F}} := \frac{\kappa_1^2}{d}\boldsymbol{F}\left((\lambda + \hat{v}\kappa_*^2)\boldsymbol{I}_p + \frac{\hat{v}\kappa_1^2}{d}\boldsymbol{F}^\intercal\boldsymbol{F}\right)^{-1}\boldsymbol{F}^\intercal \in \mathbb{R}^{p\times p} \tag{97}$$

then we can rewrite the equation as

$$q_1 = \left(\hat{m}^2 + \hat{q}_1\right)\frac{\mathrm{tr}\left[\hat{\boldsymbol{F}}\hat{\boldsymbol{F}}'\right]}{p} = \frac{\hat{m}^2 + \hat{q}_1}{\gamma}\left(\frac{\mathrm{tr}\,\hat{\boldsymbol{F}}}{p}\right)^2 = \left(1 + \frac{\hat{q}_1}{\hat{m}^2}\right)m^2, \tag{98}$$

where in the second equality we used the fact that $\hat{\boldsymbol{F}}$ and $\hat{\boldsymbol{F}}'$ are asymptotically free.

Ridge regression – Let us consider the simple case of ridge regression with $f_0(x) = x$. We will give here the channel equations that are obtained straightforwardly as

$$\begin{aligned} \hat{v} &= \frac{\alpha}{1+v} & \hat{q}_0 &= \alpha\frac{\rho - 2m + q_0}{(1+v)^2}, \\ \hat{m} &= \frac{1}{1+v}\frac{\alpha}{\sqrt{\gamma}}, & \hat{q}_1 &= \alpha\frac{\rho - 2m + q_1}{(1+v)^2}. \end{aligned} \tag{99}$$

The kernel limit is also obtained straightforwardly taking the $\alpha \to 0$ limit and rescaling $\hat{q}_0 \mapsto \alpha\hat{q}_0$, $\hat{q}_1 \mapsto \alpha\hat{q}_1$, $\hat{m} \mapsto \sqrt{\alpha}\hat{m}$, $\hat{v} \mapsto \alpha\hat{v}$.

$$\begin{aligned} v &= \frac{(1-\delta)\kappa_1^2 + \sqrt{(1-\delta)^2\kappa_1^4 + 2(\kappa_*^2+\lambda)(1+\delta)\kappa_1^2 + (\kappa_*^2+\lambda)^2} + \kappa_*^2 - \lambda}{2\lambda} \\ m &= \frac{1}{1 + \lambda\frac{v+1}{\delta\kappa_1^2}} \\ q_0 = q_1 &= \frac{\delta - 2m + \rho}{\left(1 + 2\lambda\frac{v+1}{\delta\kappa_1^2}\right)^2 - 1} \equiv q. \end{aligned} \tag{100}$$

Binary classification problem We consider now the case $f_0(x) = \mathrm{sign}(x)$, corresponding to a binary classification problem, and we write down the channel equations for this problem in the case of logistic and hinge loss. In this case we have that

$$\begin{aligned} \mathcal{Z}^0(y, \omega_0, \sigma_0) &= \frac{\delta(y-1) + \delta(y+1)}{2}\left(1 + \mathrm{erf}\left(\frac{y\omega_0}{\sqrt{2\sigma_0}}\right)\right), \\ \partial_\mu\mathcal{Z}^0(y, \omega_0, \sigma_0) &= \left(\delta(y-1) - \delta(y+1)\right)\frac{\mathrm{e}^{-\frac{\omega_0^2}{2\sigma_0}}}{\sqrt{2\pi\sigma_0}}. \end{aligned} \tag{101}$$

If we pick a logistic loss in the form $\hat{\ell}(y, \boldsymbol{\mu}) = \sum_k \ln(1 + \mathrm{e}^{-y\mu_k})$, then the proximal h solves the equation

$$h = \omega + \frac{yv}{1 + \mathrm{e}^{yh}}, \tag{102}$$

in such a way that $f = \frac{\eta-\omega}{v}$ satisfies

$$\partial_\omega f = -\left(v + 2\cosh\left(y\frac{vf+\omega}{2}\right)\right)^{-1}. \tag{103}$$

If we use instead a hing loss $\hat{\ell}(y, \boldsymbol{\mu}) = \sum_k \max(0, 1 - y\mu_k)$, the proximal is such that

$$f = \begin{cases} y & \text{if } 1-v > \omega y, \\ \frac{y-\omega}{v} & \text{if } 1-v < \omega y < 1, \\ 0 & \text{otherwise,} \end{cases} \qquad \partial_\omega f = \begin{cases} -\frac{1}{v} & \text{if } 1-v < \omega y < 1, \\ 0 & \text{otherwise.} \end{cases} \tag{104}$$

In the case of the hinge loss, the simple form of the proximal allows for a more explicit expression of the channel equations. Introducing

$$\hat{\sigma} = \rho - \frac{2m^2}{q_0+q_1}, \qquad \mathcal{N}(\zeta,\zeta';\eta) = \frac{\exp\left(-\frac{\zeta^2+\zeta'^2-2\eta\zeta\zeta'}{2(1-\eta^2)}\right)}{2\pi\sqrt{1-\eta^2}} \tag{105}$$

we obtain

$$\begin{aligned}
\hat{v} &= \frac{\alpha}{v}\int_{\frac{1-v}{\sqrt{q_0}}}^{1/\sqrt{q_0}} \frac{\mathrm{e}^{-\frac{\zeta^2}{2}}\mathrm{d}\zeta}{\sqrt{2\pi}}\left(1+\mathrm{erf}\left(\frac{m\zeta}{\sqrt{2q_0\sigma}}\right)\right), \\
\hat{m} &= \frac{\alpha}{\sqrt{\rho\gamma}v}\left[\frac{v+\mathrm{erf}\left(\sqrt{\frac{\rho}{2q_0\sigma}}\right)-(1-v)\mathrm{erf}\left((1-v)\sqrt{\frac{\rho}{2q_0\sigma}}\right)}{\sqrt{2\pi}}+\sqrt{\frac{q_0\sigma}{\rho}}\frac{\exp\left(-\frac{\rho}{2q_0\sigma}\right)-\exp\left(\frac{\rho(1-v)^2}{2q_0\sigma}\right)}{2\pi}\right] \\
\hat{q}_0 &= \alpha\left[\int_{-\infty}^{\frac{1-v}{\sqrt{q_0}}}\frac{\mathrm{e}^{-\frac{\zeta^2}{2}}\mathrm{d}\zeta}{\sqrt{2\pi}}\left(1+\mathrm{erf}\left(\frac{m\zeta}{\sqrt{2q_0\sigma}}\right)\right)+\int_{\frac{1-v}{\sqrt{q_0}}}^{1/\sqrt{q_0}}\frac{\mathrm{e}^{-\frac{\zeta^2}{2}}\mathrm{d}\zeta}{\sqrt{2\pi}}\left(1+\mathrm{erf}\left(\frac{m\zeta}{\sqrt{2q_0\sigma}}\right)\right)\left(\frac{1-\sqrt{q_0}\zeta}{v}\right)^2\right], \\
\hat{q}_1 &= \alpha\iint_{-\infty}^{\frac{1-v}{\sqrt{q_0}}}\mathcal{N}(\zeta,\zeta';{}^{q_1}/_{q_0})\mathrm{d}\zeta\mathrm{d}\zeta'\left(1+\mathrm{erf}\left(\frac{m}{\sqrt{2q_0\hat{\sigma}}}\frac{\zeta+\zeta'}{1+{}^{q_1}/_{q_0}}\right)\right) \\
&\quad +2\alpha\int_{\frac{1-v}{\sqrt{q_0}}}^{1/\sqrt{q_0}}\mathrm{d}\zeta\int_{-\infty}^{\frac{1-v}{\sqrt{q_0}}}\mathrm{d}\zeta'\mathcal{N}(\zeta,\zeta';{}^{q_1}/_{q_0})\left(1+\mathrm{erf}\left(\frac{m}{\sqrt{2q_0\hat{\sigma}}}\frac{\zeta+\zeta'}{1+{}^{q_1}/_{q_0}}\right)\right)\left(\frac{1-\sqrt{q_0}\zeta}{v}\right) \\
&\quad +\alpha\iint_{\frac{1-v}{\sqrt{q_0}}}^{1/\sqrt{q_0}}\mathrm{d}\zeta\mathrm{d}\zeta'\mathcal{N}(\zeta,\zeta';{}^{q_1}/_{q_0})\left(1+\mathrm{erf}\left(\frac{m}{\sqrt{2q_0\hat{\sigma}}}\frac{\zeta+\zeta'}{1+{}^{q_1}/_{q_0}}\right)\right)\left(\frac{1-\sqrt{q_0}\zeta}{v}\right)\left(\frac{1-\sqrt{q_0}\zeta'}{v}\right).
\end{aligned} \tag{106}$$

Let us now make the change of variables $\zeta \mapsto \frac{\sqrt{q_0+q_1}z+\sqrt{q_0-q_1}z'}{\sqrt{2q_0}}$ and $\zeta' \mapsto \frac{\sqrt{q_0+q_1}z-\sqrt{q_0-q_1}z'}{\sqrt{2q_0}}$. This allows us to rewrite the expression for q_1 as

$$\begin{aligned}
\hat{q}_1 &= \alpha\int_{-\infty}^{\sqrt{2}(1-v)}\mathcal{N}(z;0,q_0+q_1)\left(1+\mathrm{erf}\left(\frac{mz}{\sqrt{\hat{\sigma}}(q_0+q_1)}\right)\right)\mathrm{erf}\left(\frac{\sqrt{2}(1-v)-z}{\sqrt{q_0-q_1}}\right) \\
&+2\alpha\int_{-\infty}^{\sqrt{2}(1-v/2)}\mathcal{N}(z;0,q_0+q_1)\left(1+\mathrm{erf}\left(\frac{mz}{\sqrt{\hat{\sigma}}(q_0+q_1)}\right)\right)\left(\frac{1}{v}-\frac{1}{v}\left[\frac{z}{2}\mathrm{erf}\left(\frac{x}{\sqrt{q_0-q_1}}\right)+\sqrt{\frac{q_0-q_1}{2\pi}}\mathrm{e}^{-\frac{x^2}{2}}\right]_{|\sqrt{2}(1-v)-z|}^{\sqrt{2}-z}\right)\mathrm{d}z \\
&+\alpha\iint_{\frac{1-v}{\sqrt{q_0}}}^{1/\sqrt{q_0}}\mathrm{d}\zeta\mathrm{d}\zeta'\mathcal{N}(z;0,q_0+q_1)\mathcal{N}(z':0,1)\left(1+\mathrm{erf}\left(\frac{mz}{\sqrt{\hat{\sigma}}(q_0+q_1)}\right)\right)\left(\frac{1-\sqrt{q_0}\zeta}{v}\right)\left(\frac{1-\sqrt{q_0}\zeta'}{v}\right).
\end{aligned} \tag{107}$$

B. Proof of the main theorem

In this section we prove Theorem 4.1, from which all other analytical results in the paper can be deduced. We start by reminding the learning problem defining the ensemble of estimators with a few auxiliary notations, so that this part is

self contained. The exact match with the replica prediction will be given at the end of the proof. The sketch of proof is one pioneered in (Bayati & Montanari, 2011b; Donoho & Montanari, 2016) and is the following: the estimator $\mathbf{W}^*$ is expressed as the limit of a carefully chosen sequence, an *approximate message-passing iteration* (Bayati & Montanari, 2011a; Zdeborová & Krzakala, 2016), whose iterates can be exactly characterized using an auxiliary, closed form iteration, the *state evolution equations*. We then show that converging trajectories of such an AMP iteration can be systematically found.

B.1. The learning problem

We start by reminding the definition of the problem. Consider the following generative model

$$\mathbf{y} = f_0(\frac{1}{\sqrt{d}}\mathbf{X}_0\mathbf{w}_0, \boldsymbol{\epsilon}_0) \tag{108}$$

where $\mathbf{y} \in \mathbb{R}^n, \boldsymbol{X}_0 \sim \mathcal{N}(0, \boldsymbol{\Sigma}_{00}) \in \mathbb{R}^{n\times d}, \mathbf{w}_0 \in \mathbb{R}^d$, $\boldsymbol{\epsilon}_0 \in \mathbb{R}^d$ is a noise vector and $\boldsymbol{\Sigma}_{\mathbf{00}} \in \mathbb{R}^{d\times d}$ is a positive definite matrix. The goal is to learn this generative model using an ensemble of predictors $\mathbf{W} = \left[\mathbf{w}_1|\mathbf{w}_2|...|\mathbf{w}_K\right] \in \mathbb{R}^{p\times K}$ where each predictor $\mathbf{w}_k \in \mathbb{R}^p, k \in [1, K]$ is learned using a sample dataset $\mathbf{X}_k \in \mathbb{R}^{n\times p}$, where, for any $i \in [1, n]$ and $k \in [0, K]$, we have:

$$\mathbb{E}\left[\mathbf{x}_i^k(\mathbf{x}_i^{k'})^\top\right] = \boldsymbol{\Sigma}_{kk'} \tag{109}$$

where each sample is Gaussian and we denote :

$$\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{00} & \boldsymbol{\Sigma}_{01} & ... & \boldsymbol{\Sigma}_{0K} \\ \boldsymbol{\Sigma}_{10} & \boldsymbol{\Sigma}_{11} & ... & \boldsymbol{\Sigma}_{1K} \\ ... & & & \\ \boldsymbol{\Sigma}_{K0} & \boldsymbol{\Sigma}_{K1} & ... & \boldsymbol{\Sigma}_{KK} \end{bmatrix} \in \mathbb{R}^{(Kp+d)\times(Kp+d)}. \tag{110}$$

The predictors interact with each sample dataset in a linear way, i.e. we will consider a generalized linear model acting on the ensemble of products $\{\mathbf{X}_k\mathbf{w}_k\}_{k=1}^K$:

$$\mathbf{W}^* \in \underset{\mathbf{W}\in\mathbb{R}^{p\times K}}{\arg\min}\, \mathcal{L}\left(\mathbf{y}, \left\{\frac{1}{\sqrt{p}}\mathbf{X}_k\mathbf{w}_k\right\}_{k=1}^K\right) + r_0(\mathbf{W}) \tag{111}$$

where $\mathcal{L}, r_0$ are convex functions. We wish to determine the asymptotic properties of the estimator $\mathbf{W}^*$ in the limit where $n, p, d \to \infty$ with fixed ratios $\alpha = n/p, \gamma = d/p$. We now list the necessary assumptions for our main theorem to hold.

Assumptions –

- the functions $\mathcal{L}, r_0$ are proper, closed, lower-semicontinuous, convex functions. The loss function $\mathcal{L}$ is pseudo-lipschitz of order 2 in both its arguments and the regularisation r_0 is pseudo-Lipschitz of order 2. The cost function $\mathcal{L}(\mathbf{X}.) + r(.)$ is coercive.
- for any $1 \leqslant k \leqslant K$, the matrix $\boldsymbol{\Sigma}_k \in \mathbb{R}^{p\times p}$ is symmetric and there exist strictly positive constants κ_0, κ_1 such that $\kappa_0 \leqslant \lambda_{min}(\boldsymbol{\Sigma}_k) \leqslant \lambda_{max}(\boldsymbol{\Sigma}_k) \leqslant \kappa_1$. We also assume that the matrix $\boldsymbol{\Sigma}$ is positive definite.
- their exists a positive constant C_{f_0} such that $\left\|f_0(\frac{1}{\sqrt{d}}\mathbf{X}_0\mathbf{w}_0, \boldsymbol{\epsilon}_0)\right\|_2 \leqslant C_{f_0}\left(\left\|\frac{1}{\sqrt{d}}\mathbf{X}_0\mathbf{w}_0\right\|_2 + \|\boldsymbol{\epsilon}_0\|_2\right)$
- the dimensions n, p, d grow linearly with finite ratios $\alpha = n/p$ and $\gamma = d/p$.
- the ground truth vector $\mathbf{w}_0 \in \mathbb{R}^d$ and noise vector $\boldsymbol{\epsilon}_0 \in \mathbb{R}^n$ are sampled from subgaussian probability distributions independent from each other and from all other random quantities of the learning problem.

The proof method we will employ involves expressing the estimator $\mathbf{W}^*$ as the limit of a carefully chosen sequence. In the case of non-strictly convex problems, the estimator may not be unique, making it unclear what estimator is reached by the sequence (at best we know it belongs to the set of zeroes of the subgradient of the cost function). We thus start with the following problem

$$\mathbf{W}^* \in \underset{\mathbf{W}\in\mathbb{R}^{p\times K}}{\arg\min}\, \mathcal{L}(\mathbf{y}, \{\mathbf{X}_k\mathbf{w}_k\}_{k=1}^K) + r_{\lambda_2}(\mathbf{W}) \tag{112}$$

$$\text{where, for any } \mathbf{W} \in \mathbb{R}^{p\times K},\ r_{\lambda_2}(\mathbf{W}) = r_0(\mathbf{W}) + \frac{\lambda_2}{2}\|\mathbf{W}\|_F^2 \tag{113}$$

i.e. we add a ridge regularisation to the initial problem to make it strongly convex. We will relax this additional strong convexity constraint later on.

B.2. Asymptotics for the strongly convex problem

We now reformulate the minimization problem Eq.(112) to make it amenable to an approximate message-passing iteration (AMP). The key feature of this ensembling problem, outside of the convexity which will be crucial to control the trajectories of the AMP iteration, is the fact that each predictor only interacts linearly with each design sample, along with the correlation structure of the overall dataset. We are effectively sampling n vectors of size $(Kp+d)$ from the Gaussian distribution with covariance $\mathbf{\Sigma}$, i.e. $\left[\mathbf{x}_0|\mathbf{x}_1|...|\mathbf{x}_K\right] \sim \mathcal{N}(0,\mathbf{\Sigma})$. We then write $\{\mathbf{X}_k\mathbf{w}_k\}_{k=0}^K = \left[\mathbf{X}_0\mathbf{w}_0|...|\mathbf{X}_K\mathbf{w}_K\right] \in \mathbb{R}^{n\times(K+1)}$, such that

$$\left[\mathbf{X}_0\mathbf{w}_0|...|\mathbf{X}_K\mathbf{w}_K\right] = \left[\mathbf{X}_0|...|\mathbf{X}_K\right]\mathbf{W} = \mathbf{Z}\mathbf{\Sigma}^{1/2}\begin{bmatrix}\mathbf{w}_0 & 0\\ 0 & \tilde{\mathbf{W}}\end{bmatrix} \tag{114}$$

$$\text{where}\quad \tilde{\mathbf{W}} = \begin{bmatrix}\mathbf{w}_1 & 0 & ... & 0\\ 0 & \mathbf{w}_2 & ... & 0\\ & & ... & \\ 0 & 0 & ... & \mathbf{w}_K\end{bmatrix} \in \mathbb{R}^{Kp\times K} \tag{115}$$

and $\mathbf{Z} \in \mathbb{R}^{n\times(Kp+d)}$ is a random matrix with i.i.d. $\mathcal{N}(0,1)$ elements. Then, any sample $\left[\mathbf{x}_0|\mathbf{x}_1|...|\mathbf{x}_K\right]$ may be rewritten as

$$\mathbf{x}_0 = \Psi^{1/2}\mathbf{a} \quad\text{and}\quad \left[\mathbf{x}_1|...|\mathbf{x}_K\right] = \Phi^\top\Psi^{-1/2}\mathbf{a} + \left(\Omega - \Phi^\top\Psi^{-1}\Phi\right)^{1/2}\mathbf{b} \tag{116}$$

$$\mathbf{X}_0 = \mathbf{A}\Psi^{1/2} \quad\text{and}\quad \left[\mathbf{X}_1|...|\mathbf{X}_K\right] = \mathbf{A}\Psi^{-1/2}\Phi + \mathbf{B}\left(\Omega - \Phi^\top\Psi^{-1}\Phi\right)^{1/2} \tag{117}$$

where $\mathbf{a} \in \mathbb{R}^d, \mathbf{b} \in \mathbb{R}^{Kp}$ are vectors with i.i.d. standard normal components, $\mathbf{A} \in \mathbb{R}^{n\times d}, \mathbf{B} \in \mathbb{R}^{n\times Kp}$ are the corresponding design matrices, and the covariance matrices are given by $\mathbf{\Psi} = \mathbf{\Sigma}_{00} \in \mathbb{R}^{d\times d}, \mathbf{\Phi} = \left[\mathbf{\Sigma}_{11}|\mathbf{\Sigma}_{12}|\mathbf{\Sigma}_{13}...|\mathbf{\Sigma}_{1K}\right] \in \mathbb{R}^{d\times Kp}$ and

$$\mathbf{\Omega} = \begin{bmatrix}\mathbf{\Sigma}_{11} & \mathbf{\Sigma}_{12} & ... & \mathbf{\Sigma}_{1K}\\ \mathbf{\Sigma}_{21} & \mathbf{\Sigma}_{22} & ... & \mathbf{\Sigma}_{2K}\\ ... & & & \\ \mathbf{\Sigma}_{K1} & \mathbf{\Sigma}_{K2} & ... & \mathbf{\Sigma}_{KK}\end{bmatrix} \in \mathbb{R}^{Kp\times Kp} \tag{118}$$

The optimization problem may then be written, introducing the appropriate scalings

$$\tilde{\mathbf{W}}^* \in \underset{\tilde{\mathbf{W}}\in\mathbb{R}^{Kp\times K}}{\arg\min}\ \mathcal{L}\left(f_0(\frac{1}{\sqrt{d}}\mathbf{A}\tilde{\mathbf{w}}_0), \frac{1}{\sqrt{p}}\left(\mathbf{A}\Psi^{-1/2}\Phi + \mathbf{B}\left(\Omega - \Phi^\top\Psi^{-1}\Phi\right)^{1/2}\right)\tilde{\mathbf{W}}\right) + r(\tilde{\mathbf{W}}) \tag{119}$$

where we let $\tilde{\mathbf{w}}_0 = \Psi^{1/2}\mathbf{w}_0$, its scaled norm $\rho_{\tilde{\mathbf{w}}_0} = \frac{1}{d}\|\tilde{\mathbf{w}}_0\|_2^2$ and we introduced the function

$$r : \mathbb{R}^{Kp\times K} \to \mathbb{R} \tag{120}$$

$$\tilde{\mathbf{W}} \to r_{\lambda_2}(\mathbf{W}) \tag{121}$$

. In order to isolate the contribution correlated with the teacher, we condition the design matrix $\mathbf{A}$ on the teacher distribution $\mathbf{y}$, we can write

$$\mathbf{A} = \mathbb{E}\left[\mathbf{A}|\mathbf{y}\right] + \mathbf{A} - \mathbb{E}\left[\mathbf{A}|\mathbf{y}\right] \tag{122}$$

$$= \mathbb{E}\left[\mathbf{A}|\mathbf{A}\tilde{\mathbf{w}}_0\right] + \mathbf{A} - \mathbb{E}\left[\mathbf{A}|\mathbf{A}\tilde{\mathbf{w}}_0\right] \tag{123}$$

$$= \mathbf{A}\mathbf{P}_{\tilde{\mathbf{w}}_0} + \tilde{\mathbf{A}}\mathbf{P}^\perp_{\tilde{\mathbf{w}}_0} \tag{124}$$

where $\tilde{\mathbf{A}}$ is an independent copy of $\mathbf{A}$, see (Bayati & Montanari, 2011a) Lemma 11. The cost function then becomes

$$\mathcal{L}\left(f_0\left(\sqrt{\rho_{\tilde{\mathbf{w}}_0}}\mathbf{s}\right), \frac{1}{\sqrt{p}}\left(\mathbf{s}\frac{(\Phi^\top\mathbf{w}_0)^\top}{\sqrt{d\rho_{\tilde{\mathbf{w}}_0}}} + \tilde{\mathbf{A}}\mathbf{P}^\perp_{\tilde{\mathbf{w}}_0}\Psi^{-1/2}\Phi + \mathbf{B}\left(\Omega - \Phi^\top\Psi^{-1}\Phi\right)^{1/2}\right)\tilde{\mathbf{W}}\right) + r(\tilde{\mathbf{W}}) \tag{125}$$

where $\mathbf{s} = \mathbf{A}\frac{\tilde{\mathbf{w}}_0}{\|\tilde{\mathbf{w}}_0\|_2} \in \mathbb{R}^n$ is an i.i.d. standard normal vector. The term $\tilde{\mathbf{A}}\mathbf{P}^{\perp}_{\tilde{\mathbf{w}}_0}\Psi^{-1/2}\Phi + \mathbf{B}\left(\Omega - \Phi^\top\Psi^{-1}\Phi\right)^{1/2}$ can then be represented as a $\mathbb{R}^{n\times Kp}$ Gaussian matrix with covariance

$$\Phi^\top\Psi^{-1/2}\mathbf{P}^{\perp}_{\tilde{\mathbf{w}}_0}\Psi^{-1/2}\Phi + \Omega - \Phi^\top\Psi^{-1}\Phi = \Omega - \Phi^\top\Psi^{-1/2}\mathbf{P}_{\tilde{\mathbf{w}}_0}\Psi^{-1/2}\Phi \tag{126}$$

$$= \Omega - \Phi^\top\Psi^{-1/2}\frac{\tilde{\mathbf{w}}_0\tilde{\mathbf{w}}_0^\top}{\|\tilde{\mathbf{w}}_0\|_2^2}\Psi^{-1/2}\Phi = \Omega - \frac{\mathbf{c}\mathbf{c}^\top}{\|\tilde{\mathbf{w}}_0\|_2^2} \tag{127}$$

where we introduced $\mathbf{c} = \Phi^\top\mathbf{w}_0 \in \mathbb{R}^{Kp}$ and $\rho_{\mathbf{c}} = \frac{1}{p}\|\mathbf{c}\|_2^2$, reaching the cost function

$$\mathcal{L}\left(f_0\left(\sqrt{\rho_{\tilde{\mathbf{w}}_0}}\mathbf{s}\right), \frac{1}{\sqrt{p}}\left(\mathbf{s}\frac{\mathbf{c}^\top}{\sqrt{d\rho_{\tilde{\mathbf{w}}_0}}} + \mathbf{Z}\left(\Omega - \frac{\mathbf{c}\mathbf{c}^\top}{\|\tilde{\mathbf{w}}_0\|_2^2}\right)^{1/2}\right)\tilde{\mathbf{W}}\right) + r(\tilde{\mathbf{W}}) \tag{128}$$

Introducing $\mathbf{m} = \frac{1}{\sqrt{dp}}\tilde{\mathbf{W}}^\top\mathbf{c} \in \mathbb{R}^K, \mathbf{C} = \Omega - \frac{\mathbf{c}\mathbf{c}^\top}{\|\tilde{\mathbf{w}}_0\|_2^2} \in \mathbb{R}^{Kp\times Kp}$, and the Lagrange multiplier $\boldsymbol{\nu}$ associated to $\mathbf{m}$, the optimization problem can equivalently be written

$$\inf_{\mathbf{m}\in\mathbb{R}^K, \tilde{\mathbf{W}}\in\mathbb{R}^{Kp\times K}} \sup_{\boldsymbol{\nu}\in\mathbb{R}^K} \mathcal{L}\left(f_0\left(\sqrt{\rho_{\tilde{\mathbf{w}}_0}}\mathbf{s}\right), \mathbf{s}\frac{\mathbf{m}^\top}{\sqrt{\rho_{\tilde{\mathbf{w}}_0}}} + \frac{1}{\sqrt{p}}\mathbf{Z}\mathbf{C}^{1/2}\tilde{\mathbf{W}}\right) + r(\tilde{\mathbf{W}}) - \boldsymbol{\nu}^\top\left(\tilde{\mathbf{W}}^\top\mathbf{c} - \sqrt{dp}\mathbf{m}\right) \tag{129}$$

We now look for an explicit expression of the matrix square root $\mathbf{C}^{1/2}$

$$\mathbf{C} = \Omega^{1/2}\left(Id - \frac{\Omega^{-1/2}\mathbf{c}(\Omega^{-1/2}\mathbf{c})^\top}{\|\tilde{\mathbf{w}}_0\|_2^2}\right)\Omega^{1/2} \quad \text{let} \quad \tilde{\mathbf{c}} = \Omega^{-1/2}\mathbf{c} \tag{130}$$

$$= \Omega^{1/2}\left(\mathbf{P}^{\perp}_{\tilde{\mathbf{c}}} + \kappa\mathbf{P}_{\tilde{\mathbf{c}}}\right)\Omega^{1/2} \quad \text{where} \quad \kappa = 1 - \frac{\|\tilde{\mathbf{c}}\|_2^2}{\|\tilde{\mathbf{w}}_0\|_2^2} \tag{131}$$

$$= \Omega^{1/2}\left(\mathbf{P}^{\perp}_{\tilde{\mathbf{c}}} + \sqrt{\kappa}\mathbf{P}_{\tilde{\mathbf{c}}}\right)\left(\mathbf{P}^{\perp}_{\tilde{\mathbf{c}}} + \sqrt{\kappa}\mathbf{P}_{\tilde{\mathbf{c}}}\right)\Omega^{1/2} \tag{132}$$

where the positivity of κ is ensured by the positive-definiteness of $\boldsymbol{\Sigma}$. The problem then becomes

$$\inf_{\mathbf{m},\tilde{\mathbf{W}}} \sup_{\boldsymbol{\nu}} \mathcal{L}\left(f_0\left(\sqrt{\rho_{\tilde{\mathbf{w}}_0}}\mathbf{s}\right), \mathbf{s}\frac{\mathbf{m}^\top}{\sqrt{\rho_{\tilde{\mathbf{w}}_0}}} + \frac{\sqrt{\kappa}}{\sqrt{p}}\mathbf{Z}\mathbf{P}_{\tilde{\mathbf{c}}}\Omega^{1/2}\tilde{\mathbf{W}} + \frac{1}{\sqrt{p}}\tilde{\mathbf{Z}}\mathbf{P}^{\perp}_{\tilde{\mathbf{c}}}\Omega^{1/2}\tilde{\mathbf{W}}\right) + r(\tilde{\mathbf{W}}) - \boldsymbol{\nu}^\top\left(\tilde{\mathbf{W}}^\top\mathbf{c} - \sqrt{dp}\mathbf{m}\right) \tag{133}$$

where $\tilde{\mathbf{Z}}$ is an independent copy of $\mathbf{Z}$, see (Bayati & Montanari, 2011a) Lemma 11. Then

$$\frac{\sqrt{\kappa}}{\sqrt{p}}\mathbf{Z}\mathbf{P}_{\tilde{\mathbf{c}}}\Omega^{1/2}\tilde{\mathbf{W}} = \frac{\sqrt{\kappa}}{\sqrt{p}}\tilde{\mathbf{s}}\frac{\mathbf{c}^\top\tilde{\mathbf{W}}}{\|\tilde{\mathbf{c}}\|_2} \tag{134}$$

$$= \sqrt{\kappa}\tilde{\mathbf{s}}\frac{\mathbf{c}^\top\tilde{\mathbf{W}}}{p\sqrt{\rho_{\tilde{\mathbf{c}}}}} \tag{135}$$

$$= \tilde{\mathbf{s}}\frac{\sqrt{\gamma\kappa}\mathbf{m}^\top}{\sqrt{\rho_{\tilde{\mathbf{c}}}}} \tag{136}$$

where $\tilde{\mathbf{s}} = \mathbf{Z}\frac{\tilde{\mathbf{c}}}{\|\tilde{\mathbf{c}}\|_2}$ is an i.i.d. standard normal vector and $\rho_{\tilde{\mathbf{c}}} = \frac{1}{p}\|\tilde{\mathbf{c}}\|_2^2$ such that the optimization problem becomes

$$\inf_{\mathbf{m},\tilde{\mathbf{W}}} \sup_{\boldsymbol{\nu}} \mathcal{L}\left(f_0\left(\sqrt{\rho_{\tilde{\mathbf{w}}_0}}\mathbf{s}\right), \mathbf{s}\frac{\mathbf{m}^\top}{\sqrt{\rho_{\tilde{\mathbf{w}}_0}}} + \tilde{\mathbf{s}}\frac{\sqrt{\gamma\kappa}\mathbf{m}^\top}{\sqrt{\rho_{\tilde{\mathbf{c}}}}} + \frac{1}{\sqrt{p}}\tilde{\mathbf{Z}}\mathbf{P}^{\perp}_{\tilde{\mathbf{c}}}\Omega^{1/2}\tilde{\mathbf{W}}\right) + r(\tilde{\mathbf{W}}) - \boldsymbol{\nu}^\top\left(\tilde{\mathbf{W}}^\top\mathbf{c} - \sqrt{dp}\mathbf{m}\right) \tag{137}$$

Now let $\mathbf{U} = \mathbf{P}^{\perp}_{\tilde{\mathbf{c}}}\Omega^{1/2}\tilde{\mathbf{W}}$, such that $\tilde{\mathbf{W}} = \Omega^{-1/2}\left(\frac{\sqrt{\gamma}\tilde{\mathbf{c}}\mathbf{m}^\top}{\rho_{\tilde{\mathbf{c}}}} + \mathbf{U}\right)$. The equivalent problem in $\mathbf{U}$ reads

$$\inf_{\mathbf{m},\mathbf{U}} \sup_{\boldsymbol{\nu}} \mathcal{L}\left(f_0\left(\sqrt{\rho_{\tilde{\mathbf{w}}_0}}\mathbf{s}\right), \mathbf{s}\frac{\mathbf{m}^\top}{\sqrt{\rho_{\tilde{\mathbf{w}}_0}}} + \tilde{\mathbf{s}}\frac{\sqrt{\gamma\kappa}\mathbf{m}^\top}{\sqrt{\rho_{\tilde{\mathbf{c}}}}} + \frac{1}{\sqrt{p}}\tilde{\mathbf{Z}}\mathbf{U}\right) + r(\Omega^{-1/2}\left(\frac{\sqrt{\gamma}\tilde{\mathbf{c}}\mathbf{m}^\top}{\rho_{\tilde{\mathbf{c}}}} + \mathbf{U}\right)) - \boldsymbol{\nu}^\top\mathbf{U}^\top\tilde{\mathbf{c}} \tag{138}$$

Note that the constraint defining $\mathbf{m}$ automatically enforces the orthogonality constraint on $\mathbf{U}$ w.r.t. $\tilde{\mathbf{c}}$. The following lemma characterizes properties of the feasibility sets of $\mathbf{U}, \mathbf{m}, \boldsymbol{\nu}$.

Lemma B.1. *Consider the optimization problem Eq.(138). Then there exist constants $C_{\mathbf{U}}, C_{\mathbf{m}}, C_{\boldsymbol{\nu}}$ such that*

$$\frac{1}{\sqrt{p}}\|\mathbf{U}\|_F \leqslant C_{\mathbf{U}}, \quad \|\mathbf{m}\|_2 \leqslant C_{\mathbf{m}}, \quad \|\boldsymbol{\nu}\|_2 \leqslant C_{\boldsymbol{\nu}} \tag{139}$$

with high probability as $n, p, d \to \infty$.

Proof. Consider the optimization problem defining $\tilde{\mathbf{W}}^*$

$$\tilde{\mathbf{W}}^* \in \underset{\tilde{\mathbf{W}} \in \mathbb{R}^{Kp \times K}}{\arg\min}\ \mathcal{L}(\mathbf{y}, \mathbf{X}\tilde{\mathbf{W}}) + \tilde{r}_0(\mathbf{W}) + \frac{\lambda_2}{2}\left\|\tilde{\mathbf{W}}\right\|_F^2 \tag{140}$$

which, owing to the convexity of the cost function, verifies

$$\frac{1}{p}\left(\mathcal{L}(\mathbf{y}, \mathbf{X}\tilde{\mathbf{W}}^*) + \tilde{r}_0(\tilde{\mathbf{W}}^*) + \frac{\lambda_2}{2}\left\|\tilde{\mathbf{W}}^*\right\|_F^2\right) \leqslant \frac{1}{p}\left(\mathcal{L}(\mathbf{y}, 0) + \tilde{r}_0(0)\right) \tag{141}$$

The functions $\mathcal{L}$ and $\tilde{r}_0$ are assumed to be proper, thus their sum is bounded below for any value of their arguments and we may write

$$\frac{1}{p}\frac{\lambda_2}{2}\left\|\tilde{\mathbf{W}}^*\right\|_F^2 \leqslant \frac{1}{p}\left(\mathcal{L}(\mathbf{y}, 0) + \tilde{r}_0(0)\right) \tag{142}$$

The pseudo-Lipschitz assumption on $\mathcal{L}$ and $\tilde{r}_0$ then implies that there exist positive constants $C_{\mathcal{L}}$ and $C_{\tilde{r}_0}$ such that

$$\frac{1}{p}\frac{\lambda_2}{2}\left\|\tilde{\mathbf{W}}^*\right\|_F^2 \leqslant \frac{1}{p}\left(C_{\mathcal{L}}\left(1 + \|\mathbf{y}\|_2^2\right)\right) + C_{\tilde{r}_0} \tag{143}$$

$$\leqslant \frac{1}{p}\left(C_{\mathcal{L}}\left(1 + C_{f_0}\left\|\frac{1}{\sqrt{d}}\mathbf{X}_0\mathbf{w}_0\right\|_2^2 + C_{f_0}\left\|\boldsymbol{\epsilon}_0^2\right\|\right)\right) + C_{\tilde{r}_0} \tag{144}$$

where the second line follows from the scaling assumption on the teacher function f_0. Hence

$$\frac{1}{p}\frac{\lambda_2}{2}\left\|\tilde{\mathbf{W}}^*\right\|_F^2 \leqslant C_{\mathcal{L}}\left(1 + C_{f_0}\left\|\frac{1}{\sqrt{d}}\mathbf{A}\right\|_{op}^2\left\|\Psi^{1/2}\right\|_{op}^2\frac{\gamma}{d}\|\mathbf{w}\|_0^2 + C_{f_0}\frac{\alpha}{n}\|\boldsymbol{\epsilon}_0\|_2^2\right) + C_{\tilde{r}_0} \tag{145}$$

where $\|\bullet\|_{op}$ denotes the operator norm of a given matrix, and we remind that $\mathbf{A}$ has i.i.d. $\mathcal{N}(0,1)$ elements. By assumption the maximum singular value of $\Psi^{1/2}$ is bounded. The maximum singular value of a random matrix with i.i.d. $\mathcal{N}(0, \frac{1}{d})$ elements is bounded with high probability as $n, p, d \to \infty$, see e.g., (Vershynin, 2010). Finally, $\mathbf{w}_0$ and $\boldsymbol{\epsilon}_0$ are sampled from subgaussian probability distributions, thus their scaled norms are bounded with high probability as $n, p, d \to \infty$ according to Bernstein's inequality, see e.g., (Vershynin, 2018). An application of the union bound then leads to the following statement: there exists a constant $C_{\tilde{\mathbf{W}}}$ such that $\frac{1}{p}\left\|\tilde{\mathbf{W}}\right\|_2^2 \leqslant C_{\tilde{\mathbf{W}}}$, with high probability as $n, p, d \to \infty$. Now using the definition of $\mathbf{U}$

$$\frac{1}{p}\|\mathbf{U}\|_F^2 = \frac{1}{p}\left\|\mathbf{P}_{\tilde{\mathbf{c}}}^{\perp}\Omega^{1/2}\tilde{\mathbf{W}}\right\|_F^2 \tag{146}$$

$$\leqslant \left\|\mathbf{P}_{\tilde{\mathbf{c}}}^{\perp}\right\|_{op}^2\left\|\Omega^{1/2}\right\|_{op}^2\frac{1}{p}\left\|\tilde{\mathbf{W}}\right\|_F \tag{147}$$

where the singular values of $\mathbf{P}_{\tilde{\mathbf{c}}}^{\perp}$ and $\Omega^{1/2}$ are bounded with probability one. Therefore there exists a constant $C_{\mathbf{U}}$ such that $\frac{1}{\sqrt{p}}\|\mathbf{U}\| \leqslant C_{\mathbf{U}}$ with high probability as $n, p, d \to \infty$. Then, by definition of $\mathbf{m}$ and the Cauchy-Schwarz inequality

$$\|\mathbf{m}\|_2^2 \leqslant \frac{1}{d}\|\mathbf{c}\|_2^2\frac{1}{p}\left\|\tilde{\mathbf{W}}\right\|_F^2 \tag{148}$$

$$\leqslant \|\Phi\|_{op}^2\frac{1}{d}\|\mathbf{w}_0\|_2^2\frac{1}{p}\left\|\tilde{\mathbf{W}}\right\|_F^2 \tag{149}$$

combining the results previously established on $\tilde{\mathbf{W}}$ and $\mathbf{w}_0$ with the fact that the maximum singular value of Φ is bounded, there exists a positive constant $C_{\mathbf{m}}$ such that $\|\mathbf{m}\|_2 \leqslant C_{\mathbf{m}}$ with high probability as $n, p, d \to \infty$. We finally turn to $\boldsymbol{\nu}$. The optimality condition for $\mathbf{m}$ in problem Eq.(129) gives

$$\boldsymbol{\nu} = -\frac{1}{\sqrt{dp}} \frac{\mathbf{s}^\top}{\sqrt{\rho_{\tilde{\mathbf{w}}_0}}} \partial \mathcal{L}\left(\mathbf{y}, \frac{\mathbf{s}\mathbf{m}^\top}{\sqrt{\rho_{\tilde{\mathbf{w}}_0}}} + \frac{1}{\sqrt{p}} \mathbf{Z}\mathbf{C}^{1/2}\tilde{\mathbf{W}}^*\right) \tag{150}$$

The pseudo-Lipschtiz assumption on $\mathcal{L}$ implies that we can find a constant $\mathcal{C}_{\partial\mathcal{L}}$ such that

$$\|\boldsymbol{\nu}\|_2^2 = \frac{1}{dp} \frac{\|\mathbf{s}\|_2^2}{\rho_{\tilde{\mathbf{w}}_0}} C_{\mathcal{L}} \left(1 + \|\mathbf{y}\|_2^2 + \left\| \frac{\mathbf{s}\mathbf{m}^\top}{\sqrt{\rho_{\tilde{\mathbf{w}}_0}}} + \frac{1}{\sqrt{p}} \mathbf{Z}\mathbf{C}^{1/2}\tilde{\mathbf{W}}^* \right\|_2^2\right) \tag{151}$$

the last bound then follows from similar arguments as those employed above. □

The optimization problem Eq.(138) is convex and feasible. Furthermore, we may reduce the feasibility sets of $\mathbf{m}, \boldsymbol{\nu}$ to compact spaces, and the function of $\mathbf{U}$ is coercive and thus has bounded lower level sets. Strong duality then implies we can invert the order of minimization to obtain the equivalent problem

$$\inf_{\mathbf{m}} \sup_{\boldsymbol{\nu}} \inf_{\mathbf{U}} \mathcal{L}\left(f_0\left(\sqrt{\rho_{\tilde{\mathbf{w}}_0}}\mathbf{s}\right), \mathbf{s}\frac{\mathbf{m}^\top}{\sqrt{\rho_{\tilde{\mathbf{w}}_0}}} + \tilde{\mathbf{s}}\frac{\sqrt{\gamma\kappa}\mathbf{m}^\top}{\sqrt{\rho_{\tilde{\mathbf{c}}}}} + \frac{1}{\sqrt{p}}\tilde{\mathbf{Z}}\mathbf{U}\right) + r(\Omega^{-1/2}\left(\frac{\sqrt{\gamma}\tilde{\mathbf{c}}\mathbf{m}^\top}{\rho_{\tilde{\mathbf{c}}}} + \mathbf{U}\right)) - \boldsymbol{\nu}^\top\mathbf{U}^\top\tilde{\mathbf{c}} \tag{152}$$

and study the optimization problem in $\mathbf{U}$ at fixed $\mathbf{m}, \boldsymbol{\nu}$:

$$\inf_{\mathbf{U}\in\mathbb{R}^{Kp\times K}} \tilde{\mathcal{L}}(\frac{1}{\sqrt{p}}\tilde{\mathbf{Z}}\mathbf{U}) + \tilde{r}(\mathbf{U}) \tag{153}$$

where we defined the functions

$$\tilde{\mathcal{L}} : \mathbb{R}^{n\times K} \to \mathbb{R} \tag{154}$$

$$\frac{1}{\sqrt{p}}\tilde{\mathbf{Z}}\mathbf{U} \to \mathcal{L}\left(f_0\left(\sqrt{\rho_{\tilde{\mathbf{w}}_0}}\mathbf{s}\right), \mathbf{s}\frac{\mathbf{m}^\top}{\sqrt{\rho_{\tilde{\mathbf{w}}_0}}} + \tilde{\mathbf{s}}\frac{\sqrt{\gamma\kappa}\mathbf{m}^\top}{\sqrt{\rho_{\tilde{\mathbf{c}}}}} + \frac{1}{\sqrt{p}}\tilde{\mathbf{Z}}\mathbf{U}\right) \tag{155}$$

$$\tilde{r} : \mathbb{R}^{Kp\times K} \to \mathbb{R} \tag{156}$$

$$\mathbf{U} \to r(\Omega^{-1/2}\left(\frac{\sqrt{\gamma}\tilde{\mathbf{c}}\mathbf{m}^\top}{\rho_{\tilde{\mathbf{c}}}} + \mathbf{U}\right)) - \boldsymbol{\nu}^\top\mathbf{U}^\top\tilde{\mathbf{c}} \tag{157}$$

and the random matrix $\tilde{\mathbf{Z}}$ with i.i.d. $\mathcal{N}(0,1)$ elements is independent from all other random quantities in the problem. The asymptotic properties of the unique solution to this optimization problem can now be studied with a non-separable, matrix-valued approximate message passing iteration. The AMP iteration solving problem Eq.(153) is given in the following lemma

Lemma B.2. *Consider the following AMP iteration*

$$\boldsymbol{u}^{t+1} = \tilde{\boldsymbol{Z}}^\top \boldsymbol{h}_t(\boldsymbol{v}^t) - \mathbf{e}_t(\boldsymbol{u}^t)\langle \boldsymbol{h}_t' \rangle^\top \tag{158}$$

$$\boldsymbol{v}^t = \tilde{\boldsymbol{Z}}\mathbf{e}_t(\boldsymbol{u}^t) - \boldsymbol{h}_{t-1}(\boldsymbol{v}^{t-1})\langle \mathbf{e}_t' \rangle^\top \tag{159}$$

where for any $t \in \mathbb{N}$

$$\boldsymbol{h}_t(\mathbf{v}^t) = \left(\boldsymbol{R}_{\mathcal{L}(\mathbf{y},.),\mathbf{S}^t}(\mathbf{s}\frac{\mathbf{m}^\top}{\sqrt{\rho_{\tilde{\mathbf{w}}_0}}} + \tilde{\mathbf{s}}\frac{\sqrt{\gamma\kappa}\mathbf{m}^\top}{\sqrt{\rho_{\tilde{\mathbf{c}}}}} + \mathbf{v}^t) - \left(\mathbf{s}\frac{\mathbf{m}^\top}{\sqrt{\rho_{\tilde{\mathbf{w}}_0}}} + \tilde{\mathbf{s}}\frac{\sqrt{\gamma\kappa}\mathbf{m}^\top}{\sqrt{\rho_{\tilde{\mathbf{c}}}}} + \mathbf{v}^t\right)\right)(\mathbf{S}^t)^{-1} \tag{160}$$

$$\mathbf{e}_t(\mathbf{u}^t) = \boldsymbol{R}_{r(\Omega^{-1/2}.),\hat{\mathbf{S}}^t}\left(\mathbf{u}^t\hat{\mathbf{S}}^t + \Omega^{-1/2}\mathbf{c}\boldsymbol{\nu}^\top\hat{\mathbf{S}}^t + \frac{\sqrt{\gamma}\tilde{\mathbf{c}}\mathbf{m}^\top}{\rho_{\tilde{\mathbf{c}}}}\right) - \frac{\sqrt{\gamma}\tilde{\mathbf{c}}\mathbf{m}^\top}{\rho_{\tilde{\mathbf{c}}}} \tag{161}$$

$$\textit{and} \quad \mathbf{S}^t = \langle(\mathbf{e}^t)'\rangle^\top, \quad \hat{\mathbf{S}}^t = -\left(\langle(\mathbf{h}^t)'\rangle^\top\right)^{-1} \tag{162}$$

Then the fixed point $(\mathbf{u}^\infty, \mathbf{v}^\infty)$ *of this iteration verifies*

$$\boldsymbol{R}_{r(\Omega^{-1/2}.),\hat{\mathbf{S}}^\infty}\left(\mathbf{u}^\infty\hat{\mathbf{S}}^\infty + \Omega^{-1/2}\mathbf{c}\boldsymbol{\nu}^\top\hat{\mathbf{S}}^\infty + \frac{\sqrt{\gamma}\tilde{\mathbf{c}}\mathbf{m}^\top}{\rho_{\tilde{\mathbf{c}}}}\right) - \frac{\sqrt{\gamma}\tilde{\mathbf{c}}\mathbf{m}^\top}{\rho_{\tilde{\mathbf{c}}}} = \mathbf{U}^* \tag{163}$$

$$\boldsymbol{R}_{\mathcal{L}(\mathbf{y},.),\mathbf{S}^\infty}(\mathbf{s}\frac{\mathbf{m}^\top}{\sqrt{\rho_{\tilde{\mathbf{w}}_0}}} + \tilde{\mathbf{s}}\frac{\sqrt{\gamma\kappa}\mathbf{m}^\top}{\sqrt{\rho_{\tilde{\mathbf{c}}}}} + \mathbf{v}^\infty) - \mathbf{s}\frac{\mathbf{m}^\top}{\sqrt{\rho_{\tilde{\mathbf{w}}_0}}} + \tilde{\mathbf{s}}\frac{\sqrt{\gamma\kappa}\mathbf{m}^\top}{\sqrt{\rho_{\tilde{\mathbf{c}}}}} = \tilde{\mathbf{Z}}\mathbf{U}^* \tag{164}$$

where $\mathbf{U}^*$ *is the unique solution to the optimization problem Eq.*(153).

Proof. To find the correct form of the non-linearities in the AMP iteration, we match the optimality condition of problem Eq.(153) with the generic form of the fixed point of the AMP iteration Eq.(234). In the subsequent derivation, we absorb the scaling $\frac{1}{\sqrt{d}}$ in the matrix $\tilde{\boldsymbol{Z}}$, such that its elements are i.i.d. $\mathcal{N}(0, 1/d)$, and omit time indices for simplicity. Going back to problem Eq. (153), its optimality condition reads :

$$\tilde{\boldsymbol{Z}}^\top\partial\tilde{\mathcal{L}}(\tilde{\boldsymbol{Z}}\boldsymbol{U}) + \partial\tilde{r}(\boldsymbol{U}) = 0 \tag{165}$$

For any pair of $K \times K$ symmetric positive definite matrices $\mathbf{S}, \hat{\mathbf{S}}$, this optimality condition is equivalent to

$$\tilde{\boldsymbol{Z}}^\top\left(\partial\tilde{\mathcal{L}}(\tilde{\boldsymbol{Z}}\boldsymbol{U})\mathbf{S} + \tilde{\mathbf{Z}}\mathbf{U}\right)\mathbf{S}^{-1} + \left(\partial\tilde{r}(\boldsymbol{U})\hat{\mathbf{S}} + \mathbf{U}\right)\hat{\mathbf{S}}^{-1} = \tilde{\mathbf{Z}}^\top\tilde{\mathbf{Z}}\mathbf{U}\mathbf{S}^{-1} + \mathbf{U}\hat{\mathbf{S}}^{-1} \tag{166}$$

where we added the same quantity on both sides of the equality. For the loss function, we can then introduce the resolvent, formally D-resolvent:

$$\hat{\boldsymbol{v}} = \partial\tilde{\mathcal{L}}(\tilde{\boldsymbol{Z}}\boldsymbol{U})\boldsymbol{S} + \boldsymbol{Z}\boldsymbol{U} \iff \tilde{\boldsymbol{Z}}\boldsymbol{U} = \boldsymbol{R}_{\tilde{\mathcal{L}},\boldsymbol{S}}(\hat{\boldsymbol{v}}) \tag{167}$$

such that

$$\boldsymbol{R}_{\tilde{\mathcal{L}},\boldsymbol{S}}(\hat{\boldsymbol{v}}) = (\mathrm{Id} + \partial\tilde{\mathcal{L}}(\bullet)\boldsymbol{S})^{-1}(\hat{\boldsymbol{v}}) = \underset{\boldsymbol{T}\in\mathbb{R}^{n\times K}}{\mathrm{argmin}}\left\{\tilde{\mathcal{L}}(\boldsymbol{T}) + \frac{1}{2}\mathrm{tr}\big((\boldsymbol{T}-\hat{\boldsymbol{v}})\boldsymbol{S}^{-1}(\boldsymbol{T}-\hat{\boldsymbol{v}})^\top\big)\right\} \tag{168}$$

Similarly for the regularisation, introduce

$$\hat{\mathbf{u}} \equiv \left(\mathrm{Id} + \partial\tilde{r}(\bullet)\hat{\mathbf{S}}\right)(\boldsymbol{U}) \qquad \boldsymbol{U} = \boldsymbol{R}_{\tilde{r},\hat{\mathbf{S}}}(\hat{\mathbf{u}}) \tag{169}$$

where $\mathbf{S} \in \mathbb{R}^{K\times K}$ is a positive definite matrix, and

$$\boldsymbol{R}_{\tilde{r},\hat{\mathbf{S}}}(\hat{\mathbf{v}}) = \left(\mathrm{Id} + \partial\tilde{r}(\bullet)\hat{\mathbf{S}}\right)^{-1}(\hat{\mathbf{v}}) = \underset{\boldsymbol{T}\in\mathbb{R}^{Kp\times K}}{\arg\min}\left\{\tilde{r}(\boldsymbol{T}) + \frac{1}{2}\mathrm{tr}\left((\boldsymbol{T}-\hat{\mathbf{v}})\hat{\mathbf{S}}^{-1}(\boldsymbol{T}-\hat{\mathbf{v}})^\top\right)\right\} \tag{170}$$

where $\hat{\mathbf{S}} \in \mathbb{R}^{K\times K}$ is a positive definite matrix, and $\hat{\mathbf{v}} \in \mathbb{R}^{d\times K}$. The optimality condition Eq.(166) may then be rewritten as:

$$\tilde{\boldsymbol{Z}}^\top\left(\boldsymbol{R}_{\tilde{\mathcal{L}},\boldsymbol{S}}(\hat{\mathbf{v}}) - \hat{\mathbf{v}}\right)\boldsymbol{S}^{-1} = (\hat{\mathbf{u}} - \boldsymbol{R}_{\tilde{r},\hat{\mathbf{S}}}(\hat{\mathbf{u}}))\hat{\mathbf{S}}^{-1} \tag{171}$$

$$\tilde{\boldsymbol{Z}}\boldsymbol{R}_{\tilde{r},\hat{\mathbf{S}}}(\hat{\mathbf{u}}) = \boldsymbol{R}_{\tilde{\mathcal{L}},\boldsymbol{S}}(\hat{\mathbf{v}}) \tag{172}$$

where both equations should be satisfied. We can now define update functions based on the previously obtained block decomposition. The fixed point of the matrix-valued AMP Eq.(234), omitting the time indices for simplicity, reads:

$$\boldsymbol{u} + \mathbf{e}(\boldsymbol{u})\langle\boldsymbol{h}'\rangle^\top = \tilde{\boldsymbol{Z}}^\top\boldsymbol{h}(\boldsymbol{v}) \tag{173}$$

$$\boldsymbol{v} + \boldsymbol{h}(\boldsymbol{v})\langle\mathbf{e}'\rangle^\top = \tilde{\boldsymbol{Z}}\mathbf{e}(\boldsymbol{u}) \tag{174}$$

Matching this fixed point with the optimality condition Eq.(171) suggests the following mapping:

$$\begin{aligned} \boldsymbol{h}(\mathbf{v}) &= \left(\boldsymbol{R}_{\tilde{\mathcal{L}},\boldsymbol{S}}(\mathbf{v}) - \mathbf{v}\right)\boldsymbol{S}^{-1}, & \boldsymbol{S} &= \langle\mathbf{e}'\rangle^\top, \\ \mathbf{e}(\mathbf{u}) &= \boldsymbol{R}_{\tilde{r},\hat{\mathbf{S}}}(\mathbf{u}\hat{\mathbf{S}}), & \hat{\mathbf{S}} &= -(\langle\boldsymbol{h}'\rangle^\top)^{-1}, \end{aligned} \tag{175}$$

where we redefined $\hat{\mathbf{u}} \equiv \hat{\mathbf{u}}\hat{\mathbf{S}}$ in (169). We are now left with the task of evaluating the resolvents of $\tilde{\mathcal{L}}, \tilde{r}$ as expressions of the original functions $\mathcal{L}, r$. Starting with the loss function, we get

$$\boldsymbol{R}_{\tilde{\mathcal{L}},\boldsymbol{S}}(\mathbf{v}) = \underset{\mathbf{x}\in\mathbb{R}^{n\times K}}{\arg\min}\left\{\mathcal{L}\left(f_0\left(\sqrt{\rho_{\tilde{\mathbf{w}}_0}}\mathbf{s}\right), \mathbf{s}\frac{\mathbf{m}^\top}{\sqrt{\rho_{\tilde{\mathbf{w}}_0}}} + \tilde{\mathbf{s}}\frac{\sqrt{\gamma\kappa}\mathbf{m}^\top}{\sqrt{\rho_{\tilde{\mathbf{c}}}}} + \mathbf{x}\right) + \frac{1}{2}\mathrm{tr}\left((\mathbf{x}-\mathbf{v})\mathbf{S}^{-1}(\mathbf{x}-\mathbf{v})\right)^\top\right\} \tag{176}$$

letting $\tilde{\mathbf{x}} = \mathbf{s}\frac{\mathbf{m}^\top}{\sqrt{\rho_{\tilde{\mathbf{w}}_0}}} + \tilde{\mathbf{s}}\frac{\sqrt{\gamma\kappa}\mathbf{m}^\top}{\sqrt{\rho_{\tilde{\mathbf{c}}}}} + \mathbf{x}$, the problem is equivalent to

$$\begin{aligned}
\boldsymbol{R}_{\tilde{\mathcal{L}},\boldsymbol{S}}(\mathbf{v}) &= \underset{\tilde{\mathbf{x}}\in\mathbb{R}^{n\times K}}{\arg\min}\Bigg\{\mathcal{L}\left(f_0\left(\sqrt{\rho_{\tilde{\mathbf{w}}_0}}\mathbf{s}\right), \tilde{\mathbf{x}}\right) \\
&+ \frac{1}{2}\mathrm{tr}\left((\tilde{\mathbf{x}} - (\mathbf{s}\frac{\mathbf{m}^\top}{\sqrt{\rho_{\tilde{\mathbf{w}}_0}}} + \tilde{\mathbf{s}}\frac{\sqrt{\gamma\kappa}\mathbf{m}^\top}{\sqrt{\rho_{\tilde{\mathbf{c}}}}} + \mathbf{v}))\mathbf{S}^{-1}(\tilde{\mathbf{x}} - (\mathbf{s}\frac{\mathbf{m}^\top}{\sqrt{\rho_{\tilde{\mathbf{w}}_0}}} + \tilde{\mathbf{s}}\frac{\sqrt{\gamma\kappa}\mathbf{m}^\top}{\sqrt{\rho_{\tilde{\mathbf{c}}}}} + \mathbf{v}))^\top\right)\Bigg\} - \mathbf{s}\frac{\mathbf{m}^\top}{\sqrt{\rho_{\tilde{\mathbf{w}}_0}}} - \tilde{\mathbf{s}}\frac{\sqrt{\gamma\kappa}\mathbf{m}^\top}{\sqrt{\rho_{\tilde{\mathbf{c}}}}} && (177)\\
&= \boldsymbol{R}_{\mathcal{L}(\mathbf{y},.),\mathbf{S}}(\mathbf{s}\frac{\mathbf{m}^\top}{\sqrt{\rho_{\tilde{\mathbf{w}}_0}}} + \tilde{\mathbf{s}}\frac{\sqrt{\gamma\kappa}\mathbf{m}^\top}{\sqrt{\rho_{\tilde{\mathbf{c}}}}} + \mathbf{v}) - \mathbf{s}\frac{\mathbf{m}^\top}{\sqrt{\rho_{\tilde{\mathbf{w}}_0}}} - \tilde{\mathbf{s}}\frac{\sqrt{\gamma\kappa}\mathbf{m}^\top}{\sqrt{\rho_{\tilde{\mathbf{c}}}}} && (178)
\end{aligned}$$

and the corresponding non-linearity will then be

$$\boldsymbol{h}(\mathbf{v}) = \left(R_{\mathcal{L}(\mathbf{y},.),\mathbf{S}}(\mathbf{s}\frac{\mathbf{m}^\top}{\sqrt{\rho_{\tilde{\mathbf{w}}_0}}} + \tilde{\mathbf{s}}\frac{\sqrt{\gamma\kappa}\mathbf{m}^\top}{\sqrt{\rho_{\tilde{\mathbf{c}}}}} + \mathbf{v}) - \left(\mathbf{s}\frac{\mathbf{m}^\top}{\sqrt{\rho_{\tilde{\mathbf{w}}_0}}} + \tilde{\mathbf{s}}\frac{\sqrt{\gamma\kappa}\mathbf{m}^\top}{\sqrt{\rho_{\tilde{\mathbf{c}}}}} + \mathbf{v}\right)\right)\mathbf{S}^{-1} \tag{179}$$

Moving to the regularization, the resolvent reads

$$\boldsymbol{R}_{\tilde{r},\hat{\mathbf{S}}}(\mathbf{u}) = \underset{\mathbf{x}\in\mathbb{R}^{Kp\times K}}{\arg\min}\left\{r\left(\Omega^{-1/2}\left(\frac{\sqrt{\gamma}\tilde{\mathbf{c}}\mathbf{m}^\top}{\rho_{\tilde{\mathbf{c}}}} + \mathbf{x}\right)\right) - \boldsymbol{\nu}^\top\mathbf{x}^\top\Omega^{-1/2}\mathbf{c} + \frac{1}{2}\mathrm{tr}\left((\mathbf{x}-\mathbf{u})\hat{\mathbf{S}}^{-1}(\mathbf{x}-\mathbf{u})^\top\right)\right\} \tag{180}$$

letting $\tilde{\mathbf{x}} = \frac{\sqrt{\gamma}\tilde{\mathbf{c}}\mathbf{m}^\top}{\rho_{\tilde{\mathbf{c}}}} + \mathbf{x}$, we obtain

$$\begin{aligned}
\boldsymbol{R}_{\tilde{r},\hat{\mathbf{S}}}(\mathbf{u}) = \underset{\tilde{\mathbf{x}}\in\mathbb{R}^{Kp\times K}}{\arg\min}&\Bigg\{r\left(\Omega^{-1/2}\tilde{\mathbf{x}}\right) - \boldsymbol{\nu}^\top\tilde{\mathbf{x}}^\top\Omega^{-1/2}\mathbf{c} && (181)\\
&+ \frac{1}{2}\mathrm{tr}\left((\tilde{\mathbf{x}} - \left(\mathbf{u} + \frac{\sqrt{\gamma}\tilde{\mathbf{c}}\mathbf{m}^\top}{\rho_{\tilde{\mathbf{c}}}}\right))\hat{\mathbf{S}}^{-1}(\tilde{\mathbf{x}} - \left(\mathbf{u} + \frac{\sqrt{\gamma}\tilde{\mathbf{c}}\mathbf{m}^\top}{\rho_{\tilde{\mathbf{c}}}}\right))^\top\right)\Bigg\} - \frac{\sqrt{\gamma}\tilde{\mathbf{c}}\mathbf{m}^\top}{\rho_{\tilde{\mathbf{c}}}} && (182)\\
= \underset{\tilde{\mathbf{x}}\in\mathbb{R}^{Kp\times K}}{\arg\min}&\Bigg\{r\left(\Omega^{-1/2}\tilde{\mathbf{x}}\right) && (183)\\
&+ \frac{1}{2}\mathrm{tr}\left((\tilde{\mathbf{x}} - \left(\mathbf{u} + \Omega^{-1/2}\mathbf{c}\boldsymbol{\nu}^\top\hat{\mathbf{S}} + \frac{\sqrt{\gamma}\tilde{\mathbf{c}}\mathbf{m}^\top}{\rho_{\tilde{\mathbf{c}}}}\right))\hat{\mathbf{S}}^{-1}(\tilde{\mathbf{x}} - \left(\mathbf{u} + \Omega^{-1/2}\mathbf{c}\boldsymbol{\nu}^\top\hat{\mathbf{S}} + \frac{\sqrt{\gamma}\tilde{\mathbf{c}}\mathbf{m}^\top}{\rho_{\tilde{\mathbf{c}}}}\right))^\top\right)\Bigg\} && (184)\\
&- \frac{\sqrt{\gamma}\tilde{\mathbf{c}}\mathbf{m}^\top}{\rho_{\tilde{\mathbf{c}}}} && (185)\\
\boldsymbol{R}_{r(\Omega^{-1/2}.),\hat{\mathbf{S}}}&\left(\mathbf{u} + \Omega^{-1/2}\mathbf{c}\boldsymbol{\nu}^\top\hat{\mathbf{S}} + \frac{\sqrt{\gamma}\tilde{\mathbf{c}}\mathbf{m}^\top}{\rho_{\tilde{\mathbf{c}}}}\right) - \frac{\sqrt{\gamma}\tilde{\mathbf{c}}\mathbf{m}^\top}{\rho_{\tilde{\mathbf{c}}}} && (186)
\end{aligned}$$

Which gives the following non-linearity for the AMP iteration

$$e(\mathbf{u}) = \boldsymbol{R}_{r(\Omega^{-1/2}.),\hat{\mathbf{S}}}\left(\mathbf{u}\hat{\mathbf{S}} + \Omega^{-1/2}\mathbf{c}\boldsymbol{\nu}^\top\mathbf{V} + \frac{\sqrt{\gamma}\tilde{\mathbf{c}}\mathbf{m}^\top}{\rho_{\tilde{\mathbf{c}}}}\right) - \frac{\sqrt{\gamma}\tilde{\mathbf{c}}\mathbf{m}^\top}{\rho_{\tilde{\mathbf{c}}}} \tag{187}$$

□

The following lemma then gives the exact asymptotics at each time step of the AMP iteration solving problem Eq.(153) : its *state evolution equations*.

Lemma B.3. *Consider the AMP iteration Eq.(158-162). Assume it is initialized with* $\mathbf{u}^0$ *such that* $\lim_{d\to\infty}\frac{1}{d}\left\|\mathbf{e}_0(\boldsymbol{u}^0)^\top\mathbf{e}_0(\boldsymbol{u}^0)\right\|_{\mathrm{F}}$ *exists, a positive definite matrix* $\hat{\mathbf{S}}_0$*, and* $\boldsymbol{h}_{-1} \equiv 0$*. Then for any* $t \in \mathbb{N}$*, and any pair of seqeunces of uniformly pseudo-Lipschitz functions* $\phi_{1,n} : \mathbb{R}^{Kp\times K}$ *and* $\phi_{2,n} : \mathbb{R}^{n\times K}$*, the following holds*

$$\phi_{1,n}\left(\mathbf{u}^t\right) \overset{P}{\simeq} \mathbb{E}\left[\phi_{1,n}\left(\mathbf{G}(\hat{\mathbf{Q}}^t)^{1/2}\right)\right] \tag{188}$$

$$\phi_{2,n}\left(\mathbf{v}^t\right) \overset{P}{\simeq} \mathbb{E}\left[\phi_{2,n}\left(\mathbf{H}(\mathbf{Q}^t)^{1/2}\right)\right] \tag{189}$$

where $\mathbf{G} \in \mathbb{R}^{Kp\times K}$ *and* $\mathbf{H} \in \mathbb{R}^{n\times K}$ *are independent random matrices with i.i.d. standard normal elements, and* $\mathbf{Q}^t, \hat{\mathbf{Q}}^t, \mathbf{V}^t, \hat{\mathbf{V}}^t$ *are given by the equations*

$$\mathbf{Q}^t = \frac{1}{p}\mathbb{E}\Bigg[\left(\boldsymbol{R}_{r(\Omega^{-1/2}.),(\hat{\mathbf{V}}^t)^{-1}}\left(\mathbf{G}(\hat{\mathbf{Q}}^t)^{1/2}(\hat{\mathbf{V}}^t)^{-1} + \Omega^{-1/2}\mathbf{c}\boldsymbol{\nu}^\top(\hat{\mathbf{V}}^t)^{-1} + \frac{\sqrt{\gamma}\tilde{\mathbf{c}}\mathbf{m}^\top}{\rho_{\tilde{\mathbf{c}}}}\right) - \frac{\sqrt{\gamma}\tilde{\mathbf{c}}\mathbf{m}^\top}{\rho_{\tilde{\mathbf{c}}}}\right)^\top$$
$$\left(\boldsymbol{R}_{r(\Omega^{-1/2}.),(\hat{\mathbf{V}}^t)^{-1}}\left(\mathbf{G}(\hat{\mathbf{Q}}^t)^{1/2}(\hat{\mathbf{V}}^t)^{-1} + \Omega^{-1/2}\mathbf{c}\boldsymbol{\nu}^\top(\hat{\mathbf{V}}^t)^{-1} + \frac{\sqrt{\gamma}\tilde{\mathbf{c}}\mathbf{m}^\top}{\rho_{\tilde{\mathbf{c}}}}\right) - \frac{\sqrt{\gamma}\tilde{\mathbf{c}}\mathbf{m}^\top}{\rho_{\tilde{\mathbf{c}}}}\right)\Bigg] \tag{190}$$

$$\hat{\mathbf{Q}}^t = \frac{1}{p}\mathbb{E}\Bigg[\left(\left(\boldsymbol{R}_{\mathcal{L}(\mathbf{y},.),\mathbf{V}^{t-1}}(.) - Id\right)\left(\mathbf{s}\frac{\mathbf{m}^\top}{\sqrt{\rho_{\tilde{\mathbf{w}}_0}}} + \tilde{\mathbf{s}}\frac{\sqrt{\gamma\kappa}\mathbf{m}^\top}{\sqrt{\rho_{\tilde{\mathbf{c}}}}} + \mathbf{H}(\mathbf{Q}^{t-1})^{1/2}\right)(\mathbf{V}^{t-1})^{-1}\right)^\top \tag{191}$$
$$\left(\boldsymbol{R}_{\mathcal{L}(\mathbf{y},.),\mathbf{V}^{t-1}}(.) - Id\right)\left(\mathbf{s}\frac{\mathbf{m}^\top}{\sqrt{\rho_{\tilde{\mathbf{w}}_0}}} + \tilde{\mathbf{s}}\frac{\sqrt{\gamma\kappa}\mathbf{m}^\top}{\sqrt{\rho_{\tilde{\mathbf{c}}}}} + \mathbf{H}(\mathbf{Q}^{t-1})^{1/2}\right)(\mathbf{V}^{t-1})^{-1}\Bigg] \tag{192}$$

$$\mathbf{V}^t = \frac{1}{p}\mathbb{E}\left[(\hat{\mathbf{Q}}^t)^{-1/2}\mathbf{G}^\top R_{r(\Omega^{-1/2}.),(\hat{\mathbf{V}}^t)^{-1}}\left(\mathbf{G}(\hat{\mathbf{Q}}^t)^{1/2}(\hat{\mathbf{V}}^t)^{-1} + \Omega^{-1/2}\mathbf{c}\boldsymbol{\nu}^\top(\hat{\mathbf{V}}^t)^{-1} + \frac{\sqrt{\gamma}\tilde{\mathbf{c}}\mathbf{m}^\top}{\rho_{\tilde{\mathbf{c}}}}\right)\right] \tag{193}$$

$$\hat{\mathbf{V}}^t = -\frac{1}{p}\mathbb{E}\left[(\mathbf{Q}^{t-1})^{-1/2}\mathbf{H}^\top\left(\left(\boldsymbol{R}_{\mathcal{L}(\mathbf{y},.),\mathbf{V}^{t-1}}(.) - Id\right)\left(\mathbf{s}\frac{\mathbf{m}^\top}{\sqrt{\rho_{\tilde{\mathbf{w}}_0}}} + \tilde{\mathbf{s}}\frac{\sqrt{\gamma\kappa}\mathbf{m}^\top}{\sqrt{\rho_{\tilde{\mathbf{c}}}}} + \mathbf{H}(\mathbf{Q}^{t-1})^{1/2}\right)\right)(\mathbf{V}^{t-1})^{-1}\right] \tag{194}$$

Proof. Owing to the properties of Bregman proximity operators (Bauschke et al., 2003; 2006), the update functions in the AMP iteration Eq.(158-162) are Lipschitz continuous. Thus under the assumptions made on the initialization, the assumptions of Theorem B.7 are verified, which gives the desired result. □

Lemma B.4. *Consider iteration Eq.(158-162), where the parameters* $\boldsymbol{Q}, \hat{\boldsymbol{Q}}, \boldsymbol{V}, \hat{\boldsymbol{V}}$ *are initialized at any fixed point of the state evolution equations of Lemma B.3. For any sequence initialized with* $\hat{\boldsymbol{V}}_0 = \hat{\boldsymbol{V}}$ *and* $\boldsymbol{u}_0$ *such that*

$$\lim_{d\to\infty}\frac{1}{d}\mathbf{e}_0(\boldsymbol{u}_0)^\top\mathbf{e}_0(\boldsymbol{u}_0) = \boldsymbol{Q} \tag{195}$$

the following holds

$$\lim_{t\to\infty}\lim_{p\to\infty}\frac{1}{\sqrt{p}}\left\|\boldsymbol{u}^t - \boldsymbol{u}^\star\right\|_{\mathrm{F}} = 0 \quad \lim_{t\to\infty}\lim_{d\to\infty}\frac{1}{\sqrt{p}}\left\|\boldsymbol{v}^t - \boldsymbol{v}^\star\right\|_{\mathrm{F}} = 0 \tag{196}$$

Proof. The proof of this lemma is identical to that of Lemma 7 from (Loureiro et al., 2021b). □

Combining these results, we obtain the following asymptotic characterization of $\mathbf{U}^*$.

Lemma B.5. *For any fixed* $\mathbf{m}$ *and* $\boldsymbol{\nu}$ *in their feasibility sets, let* $\mathbf{U}^*$ *be the unique solution to the optimization problem Eq.(153). Then, for any sequences (in the problem dimension) of pseudo-Lipschitz functions of order 2* $\phi_{1,n} : \mathbb{R}^{n\times K} \to \mathbb{R}$ *and* $\phi_{2,n} : \mathbb{R}^{Kp\times K} \to \mathbb{R}$*, the following holds*

$$\phi_{1,n}\left(\mathbf{U}^*\right) \overset{P}{\simeq} \mathbb{E}\left[\phi_{1,n}\left(\boldsymbol{R}_{r(\Omega^{-1/2}.),\hat{\mathbf{V}}^{-1}}\left(\mathbf{G}\hat{\mathbf{Q}}^{1/2}\hat{\mathbf{V}}^{-1} + \Omega^{-1/2}\mathbf{c}\boldsymbol{\nu}^\top\hat{\mathbf{V}}^{-1} + \frac{\sqrt{\gamma}\tilde{\mathbf{c}}\mathbf{m}^\top}{\rho_{\tilde{\mathbf{c}}}}\right) - \frac{\sqrt{\gamma}\tilde{\mathbf{c}}\mathbf{m}^\top}{\rho_{\tilde{\mathbf{c}}}}\right)\right] \tag{197}$$

$$\phi_{2,n}\left(\frac{1}{\sqrt{p}}\tilde{\mathbf{Z}}\mathbf{U}^*\right) \overset{P}{\simeq} \mathbb{E}\left[\phi_{2,n}\left(\boldsymbol{R}_{\mathcal{L}(\mathbf{y},.),\mathbf{V}}(\mathbf{s}\frac{\mathbf{m}^\top}{\sqrt{\rho_{\tilde{\mathbf{w}}_0}}} + \tilde{\mathbf{s}}\frac{\sqrt{\gamma\kappa}\mathbf{m}^\top}{\sqrt{\rho_{\tilde{\mathbf{c}}}}} + \mathbf{H}\hat{\mathbf{Q}}^{1/2}) - \mathbf{s}\frac{\mathbf{m}^\top}{\sqrt{\rho_{\tilde{\mathbf{w}}_0}}} - \tilde{\mathbf{s}}\frac{\sqrt{\gamma\kappa}\mathbf{m}^\top}{\sqrt{\rho_{\tilde{\mathbf{c}}}}}\right)\right] \tag{198}$$

where $\mathbf{G} \in \mathbb{R}^{Kp\times K}$ *and* $\mathbf{H} \in \mathbb{R}^{n\times K}$ *are independent random matrices with i.i.d. standard normal elements, and* $\mathbf{Q}, \hat{\mathbf{Q}}, \mathbf{V}, \hat{\mathbf{V}}$ *are given by the fixed point (assumed to be unique) of the following set of self consistent equations*

$$\mathbf{Q} = \frac{1}{p}\mathbb{E}\Bigg[\left(\boldsymbol{R}_{r(\Omega^{-1/2}.),\hat{\mathbf{V}}^{-1}}\left(\mathbf{G}\hat{\mathbf{Q}}^{1/2}\hat{\mathbf{V}}^{-1} + \Omega^{-1/2}\mathbf{c}\boldsymbol{\nu}^\top\hat{\mathbf{V}}^{-1} + \frac{\sqrt{\gamma}\tilde{\mathbf{c}}\mathbf{m}^\top}{\rho_{\tilde{\mathbf{c}}}}\right) - \frac{\sqrt{\gamma}\tilde{\mathbf{c}}\mathbf{m}^\top}{\rho_{\tilde{\mathbf{c}}}}\right)^\top \tag{199}$$
$$\left(\boldsymbol{R}_{r(\Omega^{-1/2}.),\hat{\mathbf{V}}^{-1}}\left(\mathbf{G}\hat{\mathbf{Q}}^{1/2}\hat{\mathbf{V}}^{-1} + \Omega^{-1/2}\mathbf{c}\boldsymbol{\nu}^\top\hat{\mathbf{V}}^{-1} + \frac{\sqrt{\gamma}\tilde{\mathbf{c}}\mathbf{m}^\top}{\rho_{\tilde{\mathbf{c}}}}\right) - \frac{\sqrt{\gamma}\tilde{\mathbf{c}}\mathbf{m}^\top}{\rho_{\tilde{\mathbf{c}}}}\right)\Bigg] \tag{200}$$

$$\hat{\mathbf{Q}} = \frac{1}{p}\mathbb{E}\Bigg[\left(\left(\boldsymbol{R}_{\mathcal{L}(\mathbf{y},.),\mathbf{V}}(.) - Id\right)\left(\mathbf{s}\frac{\mathbf{m}^\top}{\sqrt{\rho_{\tilde{\mathbf{w}}_0}}} + \tilde{\mathbf{s}}\frac{\sqrt{\gamma\kappa}\mathbf{m}^\top}{\sqrt{\rho_{\tilde{\mathbf{c}}}}} + \mathbf{H}\mathbf{Q}^{1/2}\right)\mathbf{V}^{-1}\right)^\top \left(\left(\boldsymbol{R}_{\mathcal{L}(\mathbf{y},.),\mathbf{V}}(.) - Id\right)\left(\mathbf{s}\frac{\mathbf{m}^\top}{\sqrt{\rho_{\tilde{\mathbf{w}}_0}}} + \tilde{\mathbf{s}}\frac{\sqrt{\gamma\kappa}\mathbf{m}^\top}{\sqrt{\rho_{\tilde{\mathbf{c}}}}} + \mathbf{H}\mathbf{Q}^{1/2}\right)\mathbf{V}^{-1}\right)\Bigg] \tag{201}$$

$$\mathbf{V} = \frac{1}{p}\mathbb{E}\left[\hat{\mathbf{Q}}^{-1/2}\mathbf{G}^\top \boldsymbol{R}_{r(\Omega^{-1/2}.),\hat{\mathbf{V}}^{-1}}\left(\mathbf{G}\hat{\mathbf{Q}}^{1/2}\hat{\mathbf{V}}^{-1} + \Omega^{-1/2}\mathbf{c}\boldsymbol{\nu}^\top\hat{\mathbf{V}}^{-1} + \frac{\sqrt{\gamma}\tilde{\mathbf{c}}\mathbf{m}^\top}{\rho_{\tilde{\mathbf{c}}}}\right)\right] \tag{202}$$

$$\hat{\mathbf{V}} = -\frac{1}{p}\mathbb{E}\left[\mathbf{Q}^{-1/2}\mathbf{H}^\top\left(\left(\boldsymbol{R}_{\mathcal{L}(\mathbf{y},.),\mathbf{V}}(.) - Id\right)\left(\mathbf{s}\frac{\mathbf{m}^\top}{\sqrt{\rho_{\tilde{\mathbf{w}}_0}}} + \tilde{\mathbf{s}}\frac{\sqrt{\gamma\kappa}\mathbf{m}^\top}{\sqrt{\rho_{\tilde{\mathbf{c}}}}} + \mathbf{H}\mathbf{Q}^{1/2}\right)\mathbf{V}^{-1}\right)\right] \tag{203}$$

Proof. Combining the results of the previous lemmas, this proof is close to that of Theorem 1.5 in (Bayati & Montanari, 2011b). □

Returning to the optimization problem on $\mathbf{m}, \boldsymbol{\nu}$ in Eq.(152), the solution $\mathbf{U}^*$, at any dimension, verifies the zero gradient conditions on $\mathbf{m}, \boldsymbol{\nu}$:

$$\partial\boldsymbol{\nu} = 0 \iff (\mathbf{U}^*)^\top\tilde{\mathbf{c}} = 0 \tag{204}$$

$$\partial\mathbf{m} = 0 \iff \left(\frac{\mathbf{s}}{\sqrt{\rho_{\tilde{\mathbf{w}}_0}}} + \frac{\tilde{\mathbf{s}}\sqrt{\gamma\kappa}}{\rho_{\tilde{\mathbf{c}}}}\right)^\top \mathcal{L}\left(f_0\left(\sqrt{\rho_{\tilde{\mathbf{w}}_0}}\mathbf{s}\right), \mathbf{s}\frac{\mathbf{m}^\top}{\sqrt{\rho_{\tilde{\mathbf{w}}_0}}} + \tilde{\mathbf{s}}\frac{\sqrt{\gamma\kappa}\mathbf{m}^\top}{\sqrt{\rho_{\tilde{\mathbf{c}}}}} + \frac{1}{\sqrt{p}}\tilde{\mathbf{Z}}\mathbf{U}\right) + \frac{\sqrt{\gamma}\tilde{\mathbf{v}}^\top}{\rho_{\tilde{\mathbf{c}}}}\Omega^{-1/2}\partial r(\Omega^{-1/2}\left(\frac{\sqrt{\gamma}\tilde{\mathbf{c}}\mathbf{m}^\top}{\rho_{\tilde{\mathbf{c}}}} + \mathbf{U}\right)) = 0 \tag{205}$$

Using Lemma B.5 while assuming the subgradients of $\mathcal{L}, r$ are pseudo-Lipschitz (we discuss this assumption in subsection B.4), we obtain for $\mathbf{m}$

$$\frac{1}{p}\mathbb{E}\left[\left(\boldsymbol{R}_{r(\Omega^{-1/2}.),\hat{\mathbf{V}}^{-1}}\left(\mathbf{G}\hat{\mathbf{Q}}^{1/2}\hat{\mathbf{V}}^{-1} + \Omega^{-1/2}\mathbf{c}\boldsymbol{\nu}^\top\hat{\mathbf{V}}^{-1} + \frac{\sqrt{\gamma}\tilde{\mathbf{c}}\mathbf{m}^\top}{\rho_{\tilde{\mathbf{c}}}}\right) - \frac{\sqrt{\gamma}\tilde{\mathbf{c}}\mathbf{m}^\top}{\rho_{\tilde{\mathbf{c}}}}\right)^\top\tilde{\mathbf{c}}\right] = 0 \tag{206}$$

$$\iff \mathbf{m} = \frac{1}{\sqrt{dp}}\mathbb{E}\left[\tilde{\mathbf{c}}^\top\boldsymbol{R}_{r(\Omega^{-1/2}.),\hat{\mathbf{V}}^{-1}}\left(\mathbf{G}\hat{\mathbf{Q}}^{1/2}\hat{\mathbf{V}}^{-1} + \Omega^{-1/2}\mathbf{c}\boldsymbol{\nu}^\top\hat{\mathbf{V}}^{-1} + \frac{\sqrt{\gamma}\tilde{\mathbf{c}}\mathbf{m}^\top}{\rho_{\tilde{\mathbf{c}}}}\right)\right] \tag{207}$$

and for $\boldsymbol{\nu}$

$$\frac{1}{p}\mathbb{E}\Bigg[\left(\frac{\mathbf{s}}{\sqrt{\rho_{\tilde{\mathbf{w}}_0}}} + \frac{\tilde{\mathbf{s}}\sqrt{\gamma\kappa}}{\rho_{\tilde{\mathbf{c}}}}\right)^\top \partial\mathcal{L}\left(f_0\left(\sqrt{\rho_{\tilde{\mathbf{w}}_0}}\mathbf{s}\right), \boldsymbol{R}_{\mathcal{L}(\mathbf{y},.),\mathbf{V}}(\mathbf{s}\frac{\mathbf{m}^\top}{\sqrt{\rho_{\tilde{\mathbf{w}}_0}}} + \tilde{\mathbf{s}}\frac{\sqrt{\gamma\kappa}\mathbf{m}^\top}{\sqrt{\rho_{\tilde{\mathbf{c}}}}} + \mathbf{H}\hat{\mathbf{Q}}^{1/2})\right) \tag{208}$$

$$+ \frac{\sqrt{\gamma}\tilde{\mathbf{c}}^\top}{\rho_{\tilde{\mathbf{c}}}}\Omega^{-1/2}\partial r\left(\Omega^{-1/2}\left(\boldsymbol{R}_{r(\Omega^{-1/2}.),\hat{\mathbf{V}}^{-1}}\left(\mathbf{G}\hat{\mathbf{Q}}^{1/2}\hat{\mathbf{V}}^{-1} + \Omega^{-1/2}\mathbf{c}\boldsymbol{\nu}^\top\hat{\mathbf{V}}^{-1} + \frac{\sqrt{\gamma}\tilde{\mathbf{c}}\mathbf{m}^\top}{\rho_{\tilde{\mathbf{c}}}}\right)\right)\right)\Bigg] = 0 \tag{209}$$

Using the definition of D-resolvents, this is equivalent to

$$\frac{1}{p}\mathbb{E}\Bigg[\left(\frac{\mathbf{s}}{\sqrt{\rho_{\tilde{\mathbf{w}}_0}}} + \frac{\tilde{\mathbf{s}}\sqrt{\gamma\kappa}}{\rho_{\tilde{\mathbf{c}}}}\right)^\top \left(Id - \boldsymbol{R}_{\mathcal{L}(\mathbf{y},.),\mathbf{V}}(.)\right)\left(\mathbf{s}\frac{\mathbf{m}^\top}{\sqrt{\rho_{\tilde{\mathbf{w}}_0}}} + \tilde{\mathbf{s}}\frac{\sqrt{\gamma\kappa}\mathbf{m}^\top}{\sqrt{\rho_{\tilde{\mathbf{c}}}}} + \mathbf{H}\hat{\mathbf{Q}}^{1/2}\right)\mathbf{V}^{-1} \tag{210}$$

$$+ \frac{\sqrt{\gamma}\tilde{\mathbf{c}}^\top}{\rho_{\tilde{\mathbf{c}}}}\left(Id - \boldsymbol{R}_{r(\Omega^{-1/2}.),\hat{\mathbf{V}}^{-1}}(.)\right)\left(\mathbf{G}\hat{\mathbf{Q}}^{1/2}\hat{\mathbf{V}}^{-1} + \Omega^{-1/2}\mathbf{c}\boldsymbol{\nu}^\top\hat{\mathbf{V}}^{-1} + \frac{\sqrt{\gamma}\tilde{\mathbf{c}}\mathbf{m}^\top}{\rho_{\tilde{\mathbf{c}}}}\right)\hat{\mathbf{V}}\Bigg] = 0 \tag{211}$$

which simplifies to

$$\boldsymbol{\nu}^\top = -\frac{1}{\sqrt{\gamma}p}\mathbb{E}\Bigg[\left(\frac{\mathbf{s}}{\sqrt{\rho_{\tilde{\mathbf{w}}_0}}} + \frac{\tilde{\mathbf{s}}\sqrt{\gamma\kappa}}{\rho_{\tilde{\mathbf{c}}}}\right)^\top \left(Id - \boldsymbol{R}_{\mathcal{L}(\mathbf{y},.),\mathbf{V}}(.)\right)\left(\mathbf{s}\frac{\mathbf{m}^\top}{\sqrt{\rho_{\tilde{\mathbf{w}}_0}}} + \tilde{\mathbf{s}}\frac{\sqrt{\gamma\kappa}\mathbf{m}^\top}{\sqrt{\rho_{\tilde{\mathbf{c}}}}} + \mathbf{H}\hat{\mathbf{Q}}^{1/2}\right)\mathbf{V}^{-1}\Bigg] \tag{212}$$

which brings us to the following set of six self consistent equations

$$\mathbf{Q} = \frac{1}{p}\mathbb{E}\Bigg[\left(\boldsymbol{R}_{r(\Omega^{-1/2}.),\hat{\mathbf{V}}^{-1}}\left(\mathbf{G}\hat{\mathbf{Q}}^{1/2}\hat{\mathbf{V}}^{-1} + \Omega^{-1/2}\mathbf{c}\boldsymbol{\nu}^\top\hat{\mathbf{V}}^{-1} + \frac{\sqrt{\gamma}\tilde{\mathbf{c}}\mathbf{m}^\top}{\rho_{\tilde{\mathbf{c}}}}\right) - \frac{\sqrt{\gamma}\tilde{\mathbf{c}}\mathbf{m}^\top}{\rho_{\tilde{\mathbf{c}}}}\right)^\top \tag{213}$$

$$\left(\boldsymbol{R}_{r(\Omega^{-1/2}.),\hat{\mathbf{V}}^{-1}}\left(\mathbf{G}\hat{\mathbf{Q}}^{1/2}\hat{\mathbf{V}}^{-1}+\Omega^{-1/2}\mathbf{c}\boldsymbol{\nu}^\top\hat{\mathbf{V}}^{-1}+\frac{\sqrt{\gamma}\tilde{\mathbf{c}}\mathbf{m}^\top}{\rho_{\tilde{\mathbf{c}}}}\right)-\frac{\sqrt{\gamma}\tilde{\mathbf{c}}\mathbf{m}^\top}{\rho_{\tilde{\mathbf{c}}}}\right)\Bigg] \tag{214}$$

$$\hat{\mathbf{Q}}=\frac{1}{p}\mathbb{E}\Bigg[\left(\left(\boldsymbol{R}_{\mathcal{L}(\mathbf{y},.),\mathbf{V}}(.)-Id\right)\left(\mathbf{s}\frac{\mathbf{m}^\top}{\sqrt{\rho_{\tilde{\mathbf{w}}_0}}}+\tilde{\mathbf{s}}\frac{\sqrt{\gamma\kappa}\mathbf{m}^\top}{\sqrt{\rho_{\tilde{\mathbf{c}}}}}+\mathbf{H}\mathbf{Q}^{1/2}\right)\mathbf{V}^{-1}\right)^\top$$
$$\left(\left(\boldsymbol{R}_{\mathcal{L}(\mathbf{y},.),\mathbf{V}}(.)-Id\right)\left(\mathbf{s}\frac{\mathbf{m}^\top}{\sqrt{\rho_{\tilde{\mathbf{w}}_0}}}+\tilde{\mathbf{s}}\frac{\sqrt{\gamma\kappa}\mathbf{m}^\top}{\sqrt{\rho_{\tilde{\mathbf{c}}}}}+\mathbf{H}\mathbf{Q}^{1/2}\right)\mathbf{V}^{-1}\right)\Bigg] \tag{215}$$

$$\mathbf{V}=\frac{1}{p}\mathbb{E}\left[\hat{\mathbf{Q}}^{-1/2}\mathbf{G}^\top\boldsymbol{R}_{r(\Omega^{-1/2}.),\hat{\mathbf{V}}^{-1}}\left(\mathbf{G}\hat{\mathbf{Q}}^{1/2}\hat{\mathbf{V}}^{-1}+\Omega^{-1/2}\mathbf{c}\boldsymbol{\nu}^\top\hat{\mathbf{V}}^{-1}+\frac{\sqrt{\gamma}\tilde{\mathbf{c}}\mathbf{m}^\top}{\rho_{\tilde{\mathbf{c}}}}\right)\right] \tag{216}$$

$$\hat{\mathbf{V}}=-\frac{1}{p}\mathbb{E}\left[\mathbf{Q}^{-1/2}\mathbf{H}^\top\left(\left(\boldsymbol{R}_{\mathcal{L}(\mathbf{y},.),\mathbf{V}}(.)-Id\right)\left(\mathbf{s}\frac{\mathbf{m}^\top}{\sqrt{\rho_{\tilde{\mathbf{w}}_0}}}+\tilde{\mathbf{s}}\frac{\sqrt{\gamma\kappa}\mathbf{m}^\top}{\sqrt{\rho_{\tilde{\mathbf{c}}}}}+\mathbf{H}\mathbf{Q}^{1/2}\right)\mathbf{V}^{-1}\right)\right] \tag{217}$$

$$\mathbf{m}=\frac{1}{\sqrt{dp}}\mathbb{E}\left[\tilde{\mathbf{c}}^\top\boldsymbol{R}_{r(\Omega^{-1/2}.),\hat{\mathbf{V}}^{-1}}\left(\mathbf{G}\hat{\mathbf{Q}}^{1/2}\hat{\mathbf{V}}^{-1}+\Omega^{-1/2}\mathbf{c}\boldsymbol{\nu}^\top\hat{\mathbf{V}}^{-1}+\frac{\sqrt{\gamma}\tilde{\mathbf{c}}\mathbf{m}^\top}{\rho_{\tilde{\mathbf{c}}}}\right)\right] \tag{218}$$

$$\boldsymbol{\nu}^\top=-\frac{1}{\sqrt{\gamma}p}\mathbb{E}\Bigg[\left(\frac{\mathbf{s}}{\sqrt{\rho_{\tilde{\mathbf{w}}_0}}}+\frac{\tilde{\mathbf{s}}\sqrt{\gamma\kappa}}{\rho_{\tilde{\mathbf{c}}}}\right)^\top\left(Id-\boldsymbol{R}_{\mathcal{L}(\mathbf{y},.),\mathbf{V}}(.)\right)\left(\mathbf{s}\frac{\mathbf{m}^\top}{\sqrt{\rho_{\tilde{\mathbf{w}}_0}}}+\tilde{\mathbf{s}}\frac{\sqrt{\gamma\kappa}\mathbf{m}^\top}{\sqrt{\rho_{\tilde{\mathbf{c}}}}}+\mathbf{H}\hat{\mathbf{Q}}^{1/2}\right)\mathbf{V}^{-1}\Bigg] \tag{219}$$

This set of equations then characterizes the asymptotic distribution of the estimator $\mathbf{U}^*$ in the sense of Lemma B.5, with the optimal values of $\mathbf{m}$ and $\boldsymbol{\nu}$. Using the definition of $\mathbf{U}^*$ and $\tilde{\mathbf{Z}}\mathbf{U}^*$, along with the definition of the function r w.r.t. the original regularization function, a tedious but straightforward calculation allows reconstruct the asymptotic properties of $\mathbf{W}^*$ and of the set $\{\mathbf{X}_k\mathbf{w}_k^*\}_{k=1}^K$ given in the main text.

B.3. Relaxing the strong convexity constraint

Assuming the set of self consistent equations (213) have a unique fixed point regardless of the strong convexity assumption, this solution defines a unique set of six order parameters for the $\lambda_2 = 0$ case. Furthermore, using Proposition B.8, the unique estimator $\mathbf{W}^*(\lambda_2)$ solving problem Eq.(112) for strictly positive λ_2 converges to the least-norm solution to the convex (but not strongly) Eq.(111). Thus, for any pseudo-Lipschitz observable of $\mathbf{U}^*(\lambda_2)$, we have, one the one side a continuous function of λ_2 with a unique continuous extension at $\lambda_2 = 0$, and on the other side a function of λ_2 prescribed by the expectation taken w.r.t. the asymptotic Gaussian model parametrised by the state evolution parameters which is defined for all positive values of λ_2. Since both functions match for any strictly positive λ_2, continuity implies they also match for $\lambda_2 = 0$ and we obtain the exact asymptotics of the least ℓ_2 norm solution of problem Eq.(111). Regarding the uniqueness of the solution to the fixed point equations (213), it is shown in (Loureiro et al., 2021a) that a similar set of equations, although for a vector valued variable, i.e. no ensembling, the solution is unique even if the original problem is not strictly convex. This is proven by showing that the fixed point equations are the solution of a strictly convex problem. We expect this to be true here as well, and leave this part for a longer version of this paper.

B.4. A comment on non-pseudo-Lipschitz subgradients

Provided the subgradients in Eq.(204) are pseudo-Lipschitz continuous, the proof goes through. However some convex functions commonly used in machine learning, such as the hinge loss or the ℓ_1 norm for the penalty, have non-pseudo-Lipschitz gradient. To circumvent this issue, one can consider the optimization problem where both loss and regularization are replaced by their Moreau envelopes with strictly positive parameters τ_1, τ_2, as is done in (Celentano et al., 2020) for the LASSO. Moreau envelopes are everywhere differentiable and have Lipschitz gradient for strictly positive values of their parameter (Bauschke et al., 2011), thus the asymptotic characterization holds. One can then take the parameters to zero, using the fact that the limit at zero in the parameters of Moreau envelopes is well defined (Bauschke et al., 2011), recovering the original function. Since proximity operators are defined as strongly convex problems, the sequence of problems defined by the proximal operator of a Moreau envelope with decreasing parameter converges to the proximal operator of the original function when the parameter is taken to zero. Finally, inverting the expectations on random quantities with the limit taking the parameters of the Moreau envelopes to zero can be done by verifying the dominated convergence theorem using the firm-nonexpansiveness of proximity operators and the corresponding bounds on their norms, see (Bauschke et al., 2011) Chapter 4, Section 1. We leave the details of this part to a longer version of this paper.

B.5. Toolbox

In this section, we reproduce part of the appendix of (Loureiro et al., 2021b) for completeness, in order to give an overview of the main concepts and tools on approximate message passing algorithms which will be required for the proof.

Notations — For a given function $\boldsymbol{\phi}\colon \mathbb{R}^{d\times K} \to \mathbb{R}^{d\times K}$, we write :

$$\boldsymbol{\phi}(\boldsymbol{X}) = \begin{bmatrix} \boldsymbol{\phi}^1(\boldsymbol{X}) \\ \vdots \\ \boldsymbol{\phi}^d(\boldsymbol{X}) \end{bmatrix} \in \mathbb{R}^{d\times K} \tag{220}$$

where each $\boldsymbol{\phi}^i\colon \mathbb{R}^{d\times K} \to \mathbb{R}^K$. We then write the $K\times K$ Jacobian

$$\frac{\partial \boldsymbol{\phi}^i}{\partial \boldsymbol{X}_j}(\boldsymbol{X}) = \begin{bmatrix} \frac{\partial \phi^i_1(\boldsymbol{X})}{\partial X_{j1}} & \cdots & \frac{\partial \phi^i_1(\boldsymbol{X})}{\partial X_{jK}} \\ \vdots & \ddots & \vdots \\ \frac{\partial \phi^i_K(\boldsymbol{X})}{\partial X_{j1}} & \cdots & \frac{\partial \phi^i_K(\boldsymbol{X})}{\partial X_{jK}} \end{bmatrix} \in \mathbb{R}^{K\times K} \tag{221}$$

For a given matrix $\boldsymbol{Q}\in\mathbb{R}^{K\times K}$, we write $\boldsymbol{Z}\in\mathbb{R}^{n\times K}\sim\mathcal{N}(\boldsymbol{0},\boldsymbol{Q}\otimes\boldsymbol{I}_n)$ to denote that the lines of $\boldsymbol{Z}$ are sampled i.i.d. from $\mathcal{N}(\boldsymbol{0},\boldsymbol{Q})$. Note that this is equivalent to saying that $\boldsymbol{Z}=\tilde{\boldsymbol{Z}}\boldsymbol{Q}^{1/2}$ where $\tilde{\boldsymbol{Z}}\in\mathbb{R}^{n\times K}$ is an i.i.d. standard normal random matrix. The notation $\overset{\mathrm{P}}{\simeq}$ denotes convergence in probability. We start with some definitions that commonly appear in the approximate message-passing literature, see e.g., (Bayati & Montanari, 2011a; Javanmard & Montanari, 2013; Berthier et al., 2020). The main regularity class of functions we will use is that of pseudo-Lipschitz functions, which roughly amounts to functions with polynomially bounded first derivatives. We include the required scaling w.r.t. the dimensions in the definition for convenience.

Definition B.6 (Pseudo-Lipschitz function)**.** For $k,K\in\mathbb{N}^*$ and any $n,m\in\mathbb{N}^*$, a function $\boldsymbol{\phi}\colon\mathbb{R}^{n\times K}\to\mathbb{R}^{m\times K}$ is called a *pseudo-Lipschitz of order* k if there exists a constant $L(k,K)$ such that for any $\boldsymbol{X},\boldsymbol{Y}\in\mathbb{R}^{n\times K}$,

$$\frac{\|\boldsymbol{\phi}(\boldsymbol{X})-\boldsymbol{\phi}(\boldsymbol{Y})\|_{\mathrm{F}}}{\sqrt{m}} \leqslant L(k,K)\left(1+\left(\frac{\|\boldsymbol{X}\|_{\mathrm{F}}}{\sqrt{n}}\right)^{k-1}+\left(\frac{\|\boldsymbol{Y}\|_{\mathrm{F}}}{\sqrt{n}}\right)^{k-1}\right)\frac{\|\boldsymbol{X}-\boldsymbol{Y}\|_{\mathrm{F}}}{\sqrt{n}} \tag{222}$$

where $\|\bullet\|_{\mathrm{F}}$ denotes the Frobenius norm. Since K will be kept finite, it can be absorbed in any of the constants.

For example, the function $f:\mathbb{R}^{n\times K}\to\mathbb{R}, \boldsymbol{X}\mapsto\frac{1}{n}\|\boldsymbol{X}\|_F^2$ is pseudo-Lipshitz of order 2.

Moreau envelopes and Bregman proximal operators — In our proof, we will also frequently use the notions of Moreau envelopes and proximal operators, see e.g., (Parikh & Boyd, 2014; Bauschke et al., 2011). These elements of convex analysis are often encountered in recent works on high-dimensional asymptotics of convex problems, and more detailed analysis of their properties can be found for example in (Thrampoulidis et al., 2018; Loureiro et al., 2021a). For the sake of brevity, we will only sketch the main properties of such mathematical objects, referring to the cited literature for further details. In this proof, we will mainly use proximal operators acting on sets of real matrices endowed with their canonical scalar product. Furthermore, proximals will be defined with matrix valued parameters in the following way: for a given convex function $f\colon\mathbb{R}^{d\times K}\to\mathbb{R}$, a given matrix $\boldsymbol{X}\in\mathbb{R}^{d\times K}$ and a given symmetric positive definite matrix $\boldsymbol{V}\in\mathbb{R}^{K\times K}$ with bounded spectral norm, we will consider operators of the type

$$\underset{\boldsymbol{T}\in\mathbb{R}^{d\times K}}{\arg\min}\left\{f(\boldsymbol{T})+\frac{1}{2}\mathrm{tr}\left((\boldsymbol{T}-\boldsymbol{X})\boldsymbol{V}^{-1}(\boldsymbol{T}-\boldsymbol{X})^\top\right)\right\} \tag{223}$$

This operator can either be written as a standard proximal operator by factoring the matrix $\boldsymbol{V}^{-1}$ in the arguments of the trace:

$$\mathrm{prox}_{f(\bullet\boldsymbol{V}^{1/2})}(\boldsymbol{X}\boldsymbol{V}^{-1/2})\boldsymbol{V}^{1/2}\in\mathbb{R}^{d\times K} \tag{224}$$

or as a Bregman proximal operator (Bauschke et al., 2003) defined with the Bregman distance induced by the strictly convex, coercive function (for positive definite $\boldsymbol{V}$)

$$\boldsymbol{X}\mapsto\frac{1}{2}\mathrm{tr}(\boldsymbol{X}\boldsymbol{V}^{-1}\boldsymbol{X}^\top) \tag{225}$$

which justifies the use of the Bregman resolvent

$$\underset{\boldsymbol{T}\in\mathbb{R}^{d\times K}}{\arg\min}\left\{f(\boldsymbol{T})+\frac{1}{2}\mathrm{tr}\left((\boldsymbol{T}-\boldsymbol{X})\boldsymbol{V}^{-1}(\boldsymbol{T}-\boldsymbol{X})^{\top}\right)\right\}=\left(\mathrm{Id}+\partial f(\bullet)\boldsymbol{V}\right)^{-1}(\boldsymbol{X}) \tag{226}$$

Many of the usual or similar properties to that of standard proximal operators (i.e. firm non-expansiveness, link with Moreau/Bregman envelopes,...) hold for Bregman proximal operators defined with the function (225), see e.g., (Bauschke et al., 2003; 2006). In particular, we will be using the equivalent notion to firmly nonexpansive operators for Bregman proximity operators, called *D-firm* operators. Consider the Bregman proximal defined with a differentiable, strictly convex, coercive function $g:\mathcal{X}\to\mathbb{R}$, where $\mathcal{X}$ is a given input Hilbert space. Let T be the associated Bregman proximal of a given convex function $f:\mathcal{X}\to\mathbb{R}$, i.e., for any $\mathbf{x}\in\mathcal{X}$

$$T(\mathbf{x})=\underset{\mathbf{y}\in\mathcal{X}}{\arg\min}\left\{f(\mathbf{x})+D_g(\mathbf{x},\mathbf{y})\right\} \tag{227}$$

Then T is *D-firm*, meaning it verifies

$$\langle T\boldsymbol{x}-T\mathbf{y},\nabla g(T\boldsymbol{x})-\nabla g(T\mathbf{y})\rangle\leqslant\langle T\boldsymbol{x}-T\mathbf{y},\nabla g(\boldsymbol{x})-\nabla g(\mathbf{y})\rangle \tag{228}$$

for any $\mathbf{x},\mathbf{y}$ in $\mathcal{X}$.

Gradients of Bregman envelopes Consider, for any $\mathbf{X}\in\mathbb{R}^{d\times K}$ the Bregman envelope

$$\mathcal{M}_{f,\mathbf{V}}(\mathbf{X})=\inf_{\boldsymbol{T}\in\mathbb{R}^{d\times K}}\left\{f(\boldsymbol{T})+\frac{1}{2}\mathrm{tr}\left((\boldsymbol{T}-\boldsymbol{X})\boldsymbol{V}^{-1}(\boldsymbol{T}-\boldsymbol{X})^{\top}\right)\right\} \tag{229}$$

then

$$\nabla_{\mathbf{X}}\mathcal{M}_{f,\mathbf{V}}(\mathbf{X})=\left(\mathbf{X}-\left(\mathrm{Id}+\partial f(\bullet)\boldsymbol{V}\right)^{-1}(\boldsymbol{X})\right)\mathbf{V}^{-1} \tag{230}$$

and

$$\nabla_{\mathbf{V}}\mathcal{M}_{f,\mathbf{V}}(\mathbf{X})=-\frac{1}{2}\left\|\left(\mathbf{X}-\left(\mathrm{Id}+\partial f(\bullet)\boldsymbol{V}\right)^{-1}(\boldsymbol{X})\right)\mathbf{V}^{-1}\right\|_F^2 \tag{231}$$

Gaussian concentration — Gaussian concentration properties are at the root of this proof. Such properties are reviewed in more detail, for example, in (Vershynin, 2018). We refer the interested reader to related works for a more detailed discussion.

Approximate message-passing — Approximate message-passing algorithms are a statistical physics inspired family of iterations which can be used to solve high dimensional inference problems (Zdeborová & Krzakala, 2016). One of the central objects in such algorithms are the so called *state evolution equations*, a low-dimensional recursion equations which allow to exactly compute the high dimensional distribution of the iterates of the sequence. In this proof we will use a specific form of matrix-valued approximate message-passing iteration with non-separable non-linearities. In its full generality, the validity of the state evolution equations in this case is an extension of the works of (Javanmard & Montanari, 2013; Berthier et al., 2020) included in (Gerbelot & Berthier, 2021). Consider a sequence Gaussian matrices $\boldsymbol{A}(n)\in\mathbb{R}^{n\times d}$ with i.i.d. Gaussian entries, $A_{ij}(n)\sim\mathcal{N}(0,1/d)$. For each $n,d\in\mathbb{N}$, consider two sequences of pseudo-Lipschitz functions

$$\{\boldsymbol{h}_t:\mathbb{R}^{n\times K}\to\mathbb{R}^{n\times K}\}_{t\in\mathbb{N}}\qquad\{\mathbf{e}_t:\mathbb{R}^{d\times K}\to\mathbb{R}^{d\times K}\}_{t\in\mathbb{N}} \tag{232}$$

initialized on $\boldsymbol{u}^0\in\mathbb{R}^{d\times K}$ in such a way that the limit

$$\lim_{d\to\infty}\frac{1}{d}\left\|\mathbf{e}_0(\boldsymbol{u}^0)^{\top}\mathbf{e}_0(\boldsymbol{u}^0)\right\|_{\mathrm{F}} \tag{233}$$

exists, and recursively define:

$$\boldsymbol{u}^{t+1}=\boldsymbol{A}^{\top}\boldsymbol{h}_t(\boldsymbol{v}^t)-\mathbf{e}_t(\boldsymbol{u}^t)\langle\boldsymbol{h}_t'\rangle^{\top} \tag{234}$$

$$\boldsymbol{v}^t=\boldsymbol{A}\mathbf{e}_t(\boldsymbol{u}^t)-\boldsymbol{h}_{t-1}(\boldsymbol{v}^{t-1})\langle\mathbf{e}_t'\rangle^{\top} \tag{235}$$

where the dimension of the iterates are $\boldsymbol{u}^t \in \mathbb{R}^{d\times K}$ and $\boldsymbol{v}^t \in \mathbb{R}^{n\times K}$. The terms in brackets are defined as:

$$\langle \boldsymbol{h}_t' \rangle = \frac{1}{d}\sum_{i=1}^{n} \frac{\partial \boldsymbol{h}_t^i}{\partial \boldsymbol{v}_i}(\boldsymbol{v}^t) \in \mathbb{R}^{K\times K} \quad \langle \mathbf{e}_t' \rangle = \frac{1}{d}\sum_{i=1}^{d} \frac{\partial \mathbf{e}_t^i}{\partial \boldsymbol{u}_i}(\boldsymbol{u}^t) \in \mathbb{R}^{K\times K} \tag{236}$$

We define now the *state evolution recursion* on two sequences of matrices $\{\boldsymbol{Q}_{r,s}\}_{s,r\geqslant 0}$ and $\{\hat{\boldsymbol{Q}}_{r,s}\}_{s,r\geqslant 1}$ initialized with $\boldsymbol{Q}_{0,0} = \lim_{d\to\infty} \frac{1}{d}\mathbf{e}_0(\boldsymbol{u}^0)^\top \mathbf{e}_0(\boldsymbol{u}^0)$:

$$\boldsymbol{Q}_{t+1,s} = \boldsymbol{Q}_{s,t+1} = \lim_{d\to\infty} \frac{1}{d}\mathbb{E}\left[\mathbf{e}_s(\hat{\boldsymbol{Z}}^s)^\top \mathbf{e}_{t+1}(\hat{\boldsymbol{Z}}^{t+1})\right] \in \mathbb{R}^{K\times K} \tag{237}$$

$$\hat{\boldsymbol{Q}}_{t+1,s+1} = \hat{\boldsymbol{Q}}_{s+1,t+1} = \lim_{d\to\infty} \frac{1}{d}\mathbb{E}\left[\boldsymbol{h}_s(\boldsymbol{Z}^s)^\top \boldsymbol{h}_t(\boldsymbol{Z}^t)\right] \in \mathbb{R}^{K\times K} \tag{238}$$

where $(\boldsymbol{Z}^0, \dots, \boldsymbol{Z}^{t-1}) \sim \mathcal{N}(\mathbf{0}, \{\boldsymbol{Q}_{r,s}\}_{0\leqslant r,s\leqslant t-1} \otimes \boldsymbol{I}_n), (\hat{\boldsymbol{Z}}^1, \dots, \hat{\boldsymbol{Z}}^t) \sim \mathcal{N}(\mathbf{0}, \{\hat{\boldsymbol{Q}}_{r,s}\}_{1\leqslant r,s\leqslant t} \otimes \boldsymbol{I}_d)$. Then the following holds

Theorem B.7. *In the setting of the previous paragraph, for any sequence of pseudo-Lipschitz functions* $\phi_n : (\mathbb{R}^{n\times K} \times \mathbb{R}^{d\times K})^t \to \mathbb{R}$*, for* $n, d \to +\infty$*:*

$$\phi_n(\boldsymbol{u}^0, \boldsymbol{v}^0, \boldsymbol{u}^1, \boldsymbol{v}^1, \dots, \boldsymbol{v}^{t-1}, \boldsymbol{u}^t) \overset{\mathrm{P}}{\simeq} \mathbb{E}\left[\phi_n\left(\boldsymbol{u}^0, \boldsymbol{Z}^0, \hat{\boldsymbol{Z}}^1, \boldsymbol{Z}^1, \dots, \boldsymbol{Z}^{t-1}, \hat{\boldsymbol{Z}}^t\right)\right] \tag{239}$$

where $(\boldsymbol{Z}^0, \dots, \boldsymbol{Z}^{t-1}) \sim \mathcal{N}(\mathbf{0}, \{\boldsymbol{Q}_{r,s}\}_{0\leqslant r,s\leqslant t-1} \otimes \boldsymbol{I}_n), (\hat{\boldsymbol{Z}}^1, \dots, \hat{\boldsymbol{Z}}^t) \sim \mathcal{N}(\mathbf{0}, \{\hat{\boldsymbol{Q}}_{r,s}\}_{1\leqslant r,s\leqslant t} \otimes \boldsymbol{I}_n)$.

A useful result from convex analysis Here we remind a result from (Bauschke et al., 2011) describing the limiting behavior of regularized estimators for vanishing regularization.

Proposition B.8. *(Theorem 26.20 from (Bauschke et al., 2011)) Let f and h be proper, lower semi-continuous, convex functions. Suppose that* $\arg\min f \cap dom(g) \neq \emptyset$ *and that h is coercive and strictly convex. Then g admits a unique minimizer* $\mathbf{x}_0$ *over* $\arg\min f$ *and , for every* $\epsilon \in]0, 1[$*, the regularized problem*

$$\arg\min_{\mathbf{x}} f(\mathbf{x}) + \epsilon h(\mathbf{x}) \tag{240}$$

admits a unique solution $\mathbf{x}_\epsilon$*. If we assume further that h is uniformly convex on any closed ball of the input space, then* $\lim_{\epsilon\to 0} \mathbf{x}_\epsilon = \mathbf{x}_0$.