

PAPER • OPEN ACCESS

Learning curves for the multi-class teacher–student perceptron

To cite this article: Elisabetta Cornacchia *et al* 2023 *Mach. Learn.: Sci. Technol.* **4** 015019

View the [article online](#) for updates and enhancements.

You may also like

- [Theoretical characterization of uncertainty in high-dimensional linear classification](#)
Lucas Clarté, Bruno Loureiro, Florent Krzakala et al.
- [Methodology to evaluate the uncertainty associated with nanoparticle dimensional measurements by SEM](#)
L Cruzier, A Delvallée, A Allard et al.
- [Disclosure of enterprise risk management in ASEAN 5: Sustainable development for green economy](#)
Chairani and Sylvia Veronica Siregar



PAPER

Learning curves for the multi-class teacher–student perceptron

OPEN ACCESS

RECEIVED

21 September 2022

ACCEPTED FOR PUBLICATION

11 January 2023



PUBLISHED

14 February 2023

Original Content from this work may be used under the terms of the [Creative Commons Attribution 4.0 licence](#).

Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI.



Elisabetta Cornacchia^{1,10}, Francesca Mignacco^{2,3,4,10,*} , Rodrigo Veiga^{5,6,10} , Cédric Gerbelot⁷, Bruno Loureiro^{5,8} and Lenka Zdeborová⁹

¹ Ecole Polytechnique Fédérale de Lausanne (EPFL), Mathematical Data Science (MDS) lab CH-1015 Lausanne, Switzerland

² Université Paris-Saclay, CNRS, CEA, Institut de physique théorique, 91191 Gif-sur-Yvette, France

³ Joseph Henry Laboratories of Physics, Princeton University, Princeton, NJ 08544, United States of America

⁴ Initiative for the Theoretical Sciences, Graduate Center, City University of New York, New York, NY 10016, United States of America

⁵ Ecole Polytechnique Fédérale de Lausanne (EPFL), Information, Learning and Physics (IdePHICS) lab CH-1015 Lausanne, Switzerland

⁶ Universidade de São Paulo, Instituto de Física São Paulo, Brazil

⁷ Courant Institute of Mathematical Sciences, New York University, New York, NY 10012, United States of America

⁸ Département d'Informatique, École Normale Supérieure—PSL & CNRS, 45 rue d'Ulm, F-75230 Paris cedex 05, France

⁹ Ecole Polytechnique Fédérale de Lausanne (EPFL) Statistical Physics of Computation (SPOC) lab CH-1015 Lausanne, Switzerland

¹⁰ These authors contributed equally.

* Author to whom any correspondence should be addressed.

E-mail: fmignacco@princeton.edu

Keywords: multi-class classification, empirical risk minimization, high-dimensional statistics

Abstract

One of the most classical results in high-dimensional learning theory provides a closed-form expression for the generalisation error of binary classification with a single-layer teacher–student perceptron on i.i.d. Gaussian inputs. Both Bayes-optimal (BO) estimation and empirical risk minimisation (ERM) were extensively analysed in this setting. At the same time, a considerable part of modern machine learning practice concerns multi-class classification. Yet, an analogous analysis for the multi-class teacher–student perceptron was missing. In this manuscript we fill this gap by deriving and evaluating asymptotic expressions for the BO and ERM generalisation errors in the high-dimensional regime. For Gaussian teacher, we investigate the performance of ERM with both cross-entropy and square losses, and explore the role of ridge regularisation in approaching Bayes-optimality. In particular, we observe that regularised cross-entropy minimisation yields close-to-optimal accuracy. Instead, for Rademacher teacher we show that a first-order phase transition arises in the BO performance.

1. Introduction

Starting with the seminal work of Gardner and Derrida [1] the teacher–student perceptron is a broadly adopted and studied model for high-dimensional supervised binary (i.e. two classes) classification. In this model the input data are Gaussian independent identically distributed (i.i.d.) and a single-layer teacher neural network with randomly chosen i.i.d. weights from some distribution generates the labels. A student neural network then uses the input data and labels to *learn* the teacher function. The corresponding generalisation error as a function of the number of samples per dimension $\alpha = n/d$ was first derived using the replica method from statistical physics in the limit $n, d \rightarrow \infty$ for a range of teacher weights distributions (Gaussian and Rademacher being the most commonly considered) and for a range of estimators, e.g. Bayes-optimal (BO) or empirical risk minimisation (ERM) with common losses, see reviews [2–4] and references therein. Notably, the phase transition in the optimal generalisation error of the teacher–student perceptron with Rademacher teacher weights [5, 6] is possibly one of the earliest examples of the so-called *statistical-to-computational trade-offs* that are currently broadly studied in high-dimensional statistics and inference. More recently, these works on the teacher–student perceptron have been put on rigorous ground in [7] for the BO estimation, and in [8] for ERM with convex losses.

Modern machine learning classification tasks most often involve more than two classes, e.g. 10 for classification on MNIST or CIFAR10, or even 1000 in ImageNet. Multi-class classification is hence more commonly considered in practice. To the best of our knowledge, the analysis of the high-dimensional teacher–student perceptron has not been generalised to the multi-class setting yet. Closely related settings, such as multi-task Gaussian process regression [9] and high-dimensional multi-class classification with Gaussian mixture data [10–13] were recently reported, extending to the multi-class case previous works on binary classification, see, e.g. [14–17]. However, as far as we know, the teacher–student setting is still missing. In this paper we fill this gap. We define the multi-class teacher–student perceptron model and provide the following main contributions¹¹:

- C1 We derive, prove and evaluate an asymptotic closed-form expression for the generalisation error of the BO estimator in high-dimensions. In the case of Rademacher teacher weights we unveil a first-order phase transition in the learning curve.
- C2 Similarly, we derive, prove and evaluate an asymptotic closed-form expression for the generalisation performance of ridge-regularised ERM with convex losses. In particular, we discuss and compare two widely used loss functions: the square and cross-entropy losses.
- C3 We compare optimally regularised cross-entropy classification to the Bayesian classifier, and conclude that for three classes the two are extremely close, in analogy with what was observed for two classes [8].

The expressions in C1 and C2 depend on few scalar order parameters that can be efficiently obtained by solving numerically a self-consistent system of equations. The main technical difficulty of analysing the teacher–student perceptron with $k > 2$ classes is that the corresponding closed-form formulas are given in terms of a set of coupled self-consistent equations on $(k - 1) \times (k - 1)$ dimensional *matrix variables* (a.k.a. *order parameters*), involving $(k - 1)$ -dimensional integrals. This poses some challenges in both the mathematical proof and the numerical evaluation of their solution. In this work, we overcome these difficulties by building on recent works with similar matrix structure, notably the committee machine [18, 19] and the supervised k -cluster Gaussian mixture classification [20]. The heuristic replica method allows to derive a generic set of equations covering both the BO case and the ERM cases. The rigorous proof for the BO case is given in [18, 19] based on an interpolation argument. The ERM case, proven in this paper, adds the difficulty of non-Bayes optimality to the matrix valued problem. This prevents the use of both interpolation methods as in [18] or convex Gaussian comparison inequalities, see e.g. [21]. Here we handle those difficulties by employing a similar proof strategy as in [20, 22], which leverages on the rigorous analysis of matrix-valued approximate message passing iterations. Although the planted model considered here is more elaborate than in [20, 22], we show that this problem is also amenable to a matrix-valued AMP iteration by decomposing the data matrix into two parts aligned and orthogonal to the subspace spanned by the columns of the teacher weights.

In this work, we focus on the teacher–student model as a theoretical playground where important practical questions can be quantitatively studied, e.g. the impact of regularisation on the quality of learning or how to optimally tune hyper-parameters. However, the assumption of i.i.d. Gaussian data is too restrictive to capture the structure of real datasets. Hence, it is relevant to study extensions of this synthetic model that can capture the complexity of realistic settings. For instance, adapting the realistic data models presented in [20, 23–26] to multi-class classification beyond the square loss is a feasible future extension of the present work.

1.1. The data model

We consider a multi-class classification problem where the training data $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_n)^\top \in \mathbb{R}^{n \times d}$ are composed of n d -dimensional i.i.d. standard Gaussian samples, where $x_{\mu i} \sim \mathcal{N}(x_{\mu i} | 0, 1)$, $\forall i \in \{1, \dots, d\}$, $\forall \mu \in \{1, \dots, n\}$. The corresponding labels are $\mathbf{Y} = (\mathbf{y}_1, \dots, \mathbf{y}_n)^\top \in \{0, 1\}^{n \times k}$, each representing the one-hot encoding of one of k possible classes. In particular, we assume the labels are generated by a *teacher* matrix $\mathbf{W}^* = (\mathbf{w}_1^*, \dots, \mathbf{w}_k^*) \in \mathbb{R}^{d \times k}$ as

$$y_{\mu l} = \begin{cases} 1 & \text{if } l = \operatorname{argmax}_{h \in \{1, \dots, k\}} (\mathbf{w}_h^{*\top} \mathbf{x}_\mu) \\ 0 & \text{otherwise} \end{cases}, \quad \forall \mu \in \{1, \dots, n\}. \quad (1)$$

In the following, we will denote the output channel as $\phi_{\text{out}}(\mathbf{v}) := \mathbf{e}_{\operatorname{argmax}_{i \in [k]} \{v_i\}} \in \{0, 1\}^k$, where \mathbf{e}_h is the standard one-hot vector with h th site equal to 1 and all other entries equal to zero. The teacher matrix \mathbf{W}^* is drawn with i.i.d. entries either from a standard Gaussian $w_{il}^* \sim \mathcal{N}(w_{il}^* | 0, 1)$ or a Rademacher distribution

¹¹ Code repository: https://github.com/rodsveiga/mc_perceptron.

$w_{il}^* = \pm 1$ with equal probability. Note that for $k = 2$ this problem corresponds to the well-studied perceptron problem with binary labels [1, 4].

In what follows, we will be interested in the problem of *learning* the teacher target function in the high-dimensional setting, where $n, d \rightarrow \infty$ at a fixed rate, or *sample complexity*, $\alpha = n/d$, under two estimation procedures: empirical risk minimisation (ERM) and Bayes-optimal (BO) estimation.

1.2. Empirical risk minimisation (ERM)

In the first case, the statistician (or student) is given only the training data (\mathbf{X}, \mathbf{Y}) , and has to learn the teacher weights \mathbf{W}^* with a multi-class perceptron model $\hat{\mathbf{y}}(\mathbf{x}) = \phi_{\text{out}}(\mathbf{W}^\top \mathbf{x})$ by minimising a regularised empirical risk over the training set:

$$\hat{\mathbf{W}} = \operatorname{argmin}_{\mathbf{W} \in \mathbb{R}^{d \times k}} [\mathcal{L}(\mathbf{W}; \mathbf{X}, \mathbf{Y}) + r_\lambda(\mathbf{W})], \tag{2}$$

with $\mathcal{L}(\mathbf{W}; \mathbf{X}, \mathbf{Y}) = \sum_{\mu=1}^n \ell(\mathbf{W}^\top \mathbf{x}_\mu, \mathbf{y}_\mu)$. The loss ℓ accounts for the performance of the weights \mathbf{W} over a single training point. Two widely-used loss functions for multi-class classification are the cross-entropy $\ell(\mathbf{z}, \mathbf{y}) = -\sum_{l=1}^k y_l \ln \left(e^{z_l} / \sum_{l=1}^k e^{z_l} \right)$ and the square loss $\ell(\mathbf{z}, \mathbf{y}) = (\mathbf{z} - \mathbf{y})^\top (\mathbf{z} - \mathbf{y}) / 2$. We focus on ridge regularisation $r_\lambda(\mathbf{W}) = \lambda \|\mathbf{W}\|_F^2 / 2$, where $\|\cdot\|_F$ is the Frobenius norm.

1.3. Bayes-optimal estimator

In the second case, known as *Bayes-optimal* setting, the student has access not only to the training data but also to prior knowledge on the teacher weights distribution P_w^* and on the model generating the inputs and labels (1). In the teacher–student setting under consideration, where labels are generated by a noiseless channel, the BO estimator for the label \mathbf{y}_{new} of a previously unseen point \mathbf{x}_{new} can be computed directly from the BO estimator $\hat{\mathbf{W}}_{\text{BO}}$ of the teacher weights as $\hat{\mathbf{y}}_{\text{new}} = \phi_{\text{out}}(\hat{\mathbf{W}}_{\text{BO}}^\top \mathbf{x}_{\text{new}})$. The matrix $\hat{\mathbf{W}}_{\text{BO}}$ is the minimiser of the mean-squared error with respect to the ground-truth \mathbf{W}^* , i.e.

$$\hat{\mathbf{W}}_{\text{BO}} = \operatorname{argmin}_{\mathbf{W}} \mathbb{E}_{\mathbf{W}^* | (\mathbf{X}, \mathbf{Y})} \|\mathbf{W} - \mathbf{W}^*\|_F^2 = \mathbb{E}_{\mathbf{W}^* | (\mathbf{X}, \mathbf{Y})} [\mathbf{W}^*]. \tag{3}$$

Note that computing explicitly the Bayesian estimator requires computing the posterior distribution, which in general is unfeasible in high-dimensions. However, as we shall see, its performance can be characterised exactly in such limit. A key quantity in our derivation is the *free entropy density*:

$$\Phi = \lim_{d \rightarrow \infty} \frac{1}{d} \mathbb{E}_{\mathbf{X}, \mathbf{W}^*} \ln Z_d, \tag{4}$$

where the *partition function* Z_d is the normalisation of the posterior distribution over the weights

$$P(\mathbf{W} | \mathbf{X}, \mathbf{Y}) = \frac{1}{Z_d} \prod_{l=1}^k P_w^*(\mathbf{W}_l) \prod_{\mu=1}^n \delta(\mathbf{y}_\mu - \phi_{\text{out}}(\mathbf{W}^\top \mathbf{x}_\mu)). \tag{5}$$

In the BO setting, the free entropy density is closely related to the mutual information density between the labels and the weights, see [7] for an explicit discussion of this connection.

1.4. Generalisation error

The performance of different optimisation strategies is measured through the average *generalisation error*, i.e. the expected error on a fresh sample, also referred to as ‘problem average error’ in the machine learning literature [27, 28]. As it is commonly done for classification, in this work we will be interested in the misclassification rate (a.k.a. 0/1 error):

$$\varepsilon_{\text{gen}}(\alpha) = \mathbb{E}_{\mathbf{x}_{\text{new}}, \mathbf{X}, \mathbf{W}^*} \mathbb{1}[\hat{\mathbf{y}}(\hat{\mathbf{W}}(\alpha)) \neq \mathbf{y}_{\text{new}}], \tag{6}$$

where \mathbf{x}_{new} is a previously unseen data point and \mathbf{y}_{new} the corresponding label, generated by the teacher as in equation (1). Similarly, the estimator $\hat{\mathbf{y}}$ is generated by the weight matrix $\hat{\mathbf{W}}$, which in turn depends on the training set. We compare the performance obtained via ERM to the one of the BO estimator from equation (3). Note that equation (6) for the BO error can be written as

$$\begin{aligned} \varepsilon_{\text{gen}}^{\text{Bayes}} &= \frac{1}{2} \mathbb{E}_{\mathbf{X}, \mathbf{x}, \mathbf{W}^*} \|\phi_{\text{out}}(\mathbf{W}^{*\top} \mathbf{x}) - \phi_{\text{out}}(\langle \mathbf{W}^\top \mathbf{x} \rangle)\|_2^2 \\ &= 1 - \mathbb{E}_{\mathbf{X}, \mathbf{x}, \mathbf{W}^*} \left[\phi_{\text{out}}(\mathbf{W}^{*\top} \mathbf{x})^\top \phi_{\text{out}}(\hat{\mathbf{W}}_{\text{BO}}^\top \mathbf{x}) \right], \end{aligned} \tag{7}$$

where for brevity $\mathbf{x} = \mathbf{x}_{\text{new}}$ and $\langle \cdot \rangle = \mathbb{E}_{\mathbf{W} | (\mathbf{X}, \mathbf{W}^*)}$, and we have used that $\|\phi_{\text{out}}(\cdot)\|_2^2 \equiv 1$. Since the distribution of \mathbf{x}_{new} is rotationally invariant, the averaged quantity $\mathbb{E}_{\mathbf{X}, \mathbf{x}_{\text{new}}, \mathbf{W}^*} \left[\phi_{\text{out}}(\mathbf{W}^{*\top} \mathbf{x}_{\text{new}})^\top \cdot \phi_{\text{out}}(\hat{\mathbf{W}}_{\text{BO}}^\top \mathbf{x}_{\text{new}}) \right]$ only depends on the correlation between \mathbf{W}^* and $\hat{\mathbf{W}}_{\text{BO}}$, which as we will show later concentrates to the maximiser of the free entropy (4) in the high-dimensional limit.

2. Main theoretical results

From the definition of the generalisation error in (6) and of the teacher model (1), it is easy to see that crucially the generalisation error only depends on the statistics of the k -dimensional quantities $(\mathbf{W}^{*\top} \mathbf{x}_{\text{new}}, \hat{\mathbf{W}}^\top \mathbf{x}_{\text{new}}) \in \mathbb{R}^k \times \mathbb{R}^k$ (a.k.a. *local fields*) – both for BO estimation and ERM. Therefore, characterising the sufficient statistics of the local fields is equivalent to characterising the generalisation error. Our key theoretical result is that in the high-dimensional limit considered here the local fields are jointly Gaussian, and therefore the generalisation error only depends on the correlation $\bar{\mathbf{m}}_d$ between the teacher \mathbf{W}^* and the estimator $\hat{\mathbf{W}}$, and the covariances \mathbf{Q}_d^* and $\bar{\mathbf{q}}_d$ of the teacher and the estimator respectively (a.k.a. the *overlaps*): $\bar{\mathbf{m}}_d \equiv (\frac{1}{d}) \hat{\mathbf{W}}^\top \mathbf{W}^*$, $\bar{\mathbf{q}}_d \equiv (\frac{1}{d}) \hat{\mathbf{W}}^\top \hat{\mathbf{W}}$ and $\mathbf{Q}_d^* \equiv (\frac{1}{d}) \mathbf{W}^{*\top} \mathbf{W}^*$. Note that we keep the subscript d to emphasise that these definitions are still in finite dimension and to distinguish them from the corresponding overlaps in the high-dimensional limit. As we will show next, these low-dimensional sufficient statistics can be computed explicitly by solving a set of coupled $(k - 1) \times (k - 1)$ self-consistent equations.

2.1. Performance of empirical risk minimization

Our result holds under the following assumptions, in addition to the Gaussian hypothesis on the design matrix \mathbf{X} . **Assumptions:**

- (A1) the functions \mathcal{L}, r_λ are proper, closed, lower-semicontinuous, convex functions. The loss \mathcal{L} is differentiable and pseudo-Lipschitz of order 2 in both its arguments. We assume additionally that the regularisation r_λ is strongly convex, differentiable and pseudo-Lipschitz of order 2;
- (A2) the dimensions n, d grow linearly according to the finite ratio $\alpha = n/d$;
- (A3) the lines of the ground truth matrix $\mathbf{W}^* \in \mathbb{R}^{d \times k}$ are sampled i.i.d. from a sub-Gaussian probability distribution in \mathbb{R}^k .

Theorem 2.1. *Let $\xi \sim \mathcal{N}_k(\mathbf{0}, \mathbf{I}_k)$. Under (A1)–(A3), for any pair of pseudo-Lipschitz functions $\psi_1 : \mathbb{R}^{d \times k} \rightarrow \mathbb{R}, \psi_2 : \mathbb{R}^{n \times k} \rightarrow \mathbb{R}$ of order 2, the estimator $\hat{\mathbf{W}}$ and $\hat{\mathbf{Z}} = \frac{1}{\sqrt{d}} \mathbf{X} \hat{\mathbf{W}}$ satisfy:*

$$\psi_1(\hat{\mathbf{W}}) \xrightarrow{P} \mathbb{E}_\xi \left[\mathcal{Z}_w^*(\hat{\mathbf{m}} \hat{\mathbf{q}}^{-1/2} \xi, \hat{\mathbf{m}}^T \hat{\mathbf{q}}^{-1} \hat{\mathbf{m}}) \psi_1 \left(\mathbf{f}_w(\hat{\mathbf{q}}^{1/2} \xi, \hat{\mathbf{V}}) \right) \right], \tag{8}$$

$$\psi_2(\hat{\mathbf{Z}}) \xrightarrow{P} \mathbb{E}_\xi \left[\int_{\mathbb{R}^k} d\mathbf{y} \mathcal{Z}_{\text{out}}^*(\mathbf{y}, \mathbf{m} \mathbf{q}^{-1/2} \xi, \mathbf{Q}^* - \mathbf{m}^\top \mathbf{q}^{-1} \mathbf{m}) \psi_2 \left(\mathbf{f}_{\text{out}}(\mathbf{y}, \mathbf{q}^{1/2} \xi, \mathbf{V}) \right) \right], \tag{9}$$

where \xrightarrow{P} denotes convergence in probability as $n, d \rightarrow \infty$ and the parameters $(\mathbf{m}, \mathbf{q}, \mathbf{V})$ are the solution (assumed to be unique) of the following set of self-consistent equations (where we introduced the auxiliary parameters $(\hat{\mathbf{m}}, \hat{\mathbf{q}}, \hat{\mathbf{V}})$):

$$\left\{ \begin{array}{l} \mathbf{m} = \mathbb{E}_\xi \left[\mathcal{Z}_w^*(\hat{\mathbf{m}} \hat{\mathbf{q}}^{-1/2} \xi, \hat{\mathbf{m}}^T \hat{\mathbf{q}}^{-1} \hat{\mathbf{m}}) \mathbf{f}_w^*(\hat{\mathbf{m}} \hat{\mathbf{q}}^{-1/2} \xi, \hat{\mathbf{m}}^T \hat{\mathbf{q}}^{-1} \hat{\mathbf{m}}) \mathbf{f}_w(\hat{\mathbf{q}}^{1/2} \xi, \hat{\mathbf{V}})^\top \right], \\ \mathbf{q} = \mathbb{E}_\xi \left[\mathcal{Z}_w^*(\hat{\mathbf{m}} \hat{\mathbf{q}}^{-1/2} \xi, \hat{\mathbf{m}}^T \hat{\mathbf{q}}^{-1} \hat{\mathbf{m}}) \mathbf{f}_w(\hat{\mathbf{q}}^{1/2} \xi, \hat{\mathbf{V}}) \mathbf{f}_w(\hat{\mathbf{q}}^{1/2} \xi, \hat{\mathbf{V}})^\top \right], \\ \mathbf{V} = \mathbb{E}_\xi \left[\mathcal{Z}_w^*(\hat{\mathbf{m}} \hat{\mathbf{q}}^{-1/2} \xi, \hat{\mathbf{m}}^T \hat{\mathbf{q}}^{-1} \hat{\mathbf{m}}) \partial_\gamma \mathbf{f}_w(\hat{\mathbf{q}}^{1/2} \xi, \hat{\mathbf{V}}) \right], \\ \mathbf{m} = \alpha \mathbb{E}_\xi \left[\int_{\mathbb{R}^k} d\mathbf{y} \mathcal{Z}_{\text{out}}^*(\mathbf{y}, \mathbf{m} \mathbf{q}^{-1/2} \xi, \mathbf{Q}^* - \mathbf{m}^\top \mathbf{q}^{-1} \mathbf{m}) \right. \\ \quad \left. \mathbf{f}_{\text{out}}^*(\mathbf{y}, \mathbf{m} \mathbf{q}^{-1/2} \xi, \mathbf{Q}^* - \mathbf{m}^\top \mathbf{q}^{-1} \mathbf{m}) \mathbf{f}_{\text{out}}(\mathbf{y}, \mathbf{q}^{1/2} \xi, \mathbf{V})^\top \right], \\ \mathbf{q} = \alpha \mathbb{E}_\xi \left[\int_{\mathbb{R}^k} d\mathbf{y} \mathcal{Z}_{\text{out}}^*(\mathbf{y}, \mathbf{m} \mathbf{q}^{-1/2} \xi, \mathbf{Q}^* - \mathbf{m}^\top \mathbf{q}^{-1} \mathbf{m}) \right. \\ \quad \left. \mathbf{f}_{\text{out}}(\mathbf{y}, \mathbf{q}^{1/2} \xi, \mathbf{V}) \mathbf{f}_{\text{out}}(\mathbf{y}, \mathbf{q}^{1/2} \xi, \mathbf{V})^\top \right], \\ \hat{\mathbf{V}} = -\alpha \mathbb{E}_\xi \left[\int_{\mathbb{R}^k} d\mathbf{y} \mathcal{Z}_{\text{out}}^*(\mathbf{y}, \mathbf{m} \mathbf{q}^{-1/2} \xi, \mathbf{Q}^* - \mathbf{m}^\top \mathbf{q}^{-1} \mathbf{m}) \partial_\omega \mathbf{f}_{\text{out}}(\mathbf{y}, \mathbf{q}^{1/2} \xi, \mathbf{V}) \right]. \end{array} \right. \tag{10}$$

We have made use of the following auxiliary functions:

$$\begin{aligned} \mathcal{Z}_w^*(\gamma, \Lambda) &= \mathbb{E}_{w^*} e^{-\frac{1}{2} w^{*\top} \Lambda w^* + \gamma^\top w^*}, & \mathcal{Z}_{\text{out}}^*(\mathbf{y}, \omega, \mathbf{V}) &= \mathbb{E}_{\mathcal{Z}^*} \delta\left(\mathbf{y} - \phi_{\text{out}}\left(\mathbf{V}^{1/2} \mathcal{Z}^* + \omega\right)\right), \\ f_w^*(\gamma, \Lambda) &= \partial_\gamma \ln \mathcal{Z}_w^*(\gamma, \Lambda), & f_{\text{out}}^*(\mathbf{y}, \omega, \mathbf{V}) &= \partial_\omega \ln \mathcal{Z}_{\text{out}}^*(\mathbf{y}, \omega, \mathbf{V}), \\ f_w(\gamma, \Lambda) &= \text{prox}_{\Lambda^{-1} r_\lambda}(\Lambda^{-1} \gamma), & f_{\text{out}}(\mathbf{y}, \omega, \mathbf{V}) &= \text{prox}_{\mathcal{V}\ell(\cdot, \mathbf{y})}(\omega). \end{aligned} \quad (11)$$

where $w^* \sim P_w^*$, that is the distribution (in \mathbb{R}^k) of the teacher weights, $\mathcal{Z}^* \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_k)$. For brevity, in the definitions we have indicated by γ, Λ, ω the generic arguments of the auxiliary functions. \mathbf{V} is the covariance matrix of the normally distributed local fields $\mathcal{Z}^*, \mathcal{Z}$. For any function f

$$\text{prox}_{\tau f}(\mathbf{x}) = \underset{\mathcal{Z} \in \mathbb{R}^k}{\text{argmin}} \left[\frac{1}{2} (\mathcal{Z} - \mathbf{x})^\top \tau^{-1} (\mathcal{Z} - \mathbf{x}) + f(\mathcal{Z}) \right] \quad (12)$$

is a proximal operator (here defined with matrix parameters, see appendix D.3 for more detail). The simplified expressions of the auxiliary functions are provided in appendix C.1 and depend on the choices of the teacher weights distribution, the regularisation and the loss function.

2.1.1. Proof outline

We now provide a short outline of the proof for the asymptotic performance of the estimator obtained with convex ERM. The idea, pioneered in [29] for a vector valued LASSO problem, is to express the estimator $\hat{\mathbf{W}}$ as the limit of a carefully chosen sequence whose iterates have an exact, rigorous asymptotic characterisation. Such a sequence can be built using an approximate message-passing algorithm, which offers the possibility of treating low-rank matrix-valued iterates, with *state evolution* equations characterizing their high-dimensional statistics. In order to determine the AMP sequence with the correct fixed point, we decompose the optimisation problem defining the multi-class estimator, isolating components that are aligned and orthogonal to the subspace spanned by the teacher weights in order to separate random quantities correlated and independent with the labels \mathbf{Y} . We then design the AMP sequence whose fixed point matches the optimality condition of the ERM problem, and rigorously obtain its state evolution equations using [30, 31]. Using the strong convexity of the problem, we show that converging trajectories of the AMP sequence can be systematically found, ultimately characterising the unique minimiser of the ERM problem with the fixed point of the state evolution equations which match those of the replica prediction.

2.2. Bayes-optimal performance

The sufficient statistics describing the performance of the BO estimator (3) can also be derived in the high-dimensional limit, and are closely related to the free entropy density. Indeed, in appendix B we show that the BO estimator can be fully characterised by only one overlap matrix $\mathbf{q} \in \mathbb{R}^{k \times k}$ which is given by the solution of following extremisation problem:

$$\Phi = \text{extr}_{\mathbf{q}, \hat{\mathbf{q}}} \left\{ -\frac{1}{2} \text{Tr}[\mathbf{q}\hat{\mathbf{q}}] + \Psi_w^*(\hat{\mathbf{q}}) + \alpha \Psi_{\text{out}}^*(\mathbf{q}) \right\}, \quad (13a)$$

$$\Psi_w^*(\hat{\mathbf{q}}) = \mathbb{E}_\xi \left[\mathcal{Z}_w^*(\hat{\mathbf{q}}^{1/2} \xi, \hat{\mathbf{q}}) \ln \mathcal{Z}_w^*(\hat{\mathbf{q}}^{1/2} \xi, \hat{\mathbf{q}}) \right], \quad (13b)$$

$$\Psi_{\text{out}}^*(\mathbf{q}) = \mathbb{E}_\xi \left[\int_{\mathbb{R}^k} d\mathbf{y} \mathcal{Z}_{\text{out}}^*(\mathbf{y}; \mathbf{q}^{1/2} \xi, \mathbf{Q}^* - \mathbf{q}) \ln \mathcal{Z}_{\text{out}}^*(\mathbf{y}; \mathbf{q}^{1/2} \xi, \mathbf{Q}^* - \mathbf{q}) \right], \quad (13c)$$

where $(\mathcal{Z}_w^*, \mathcal{Z}_{\text{out}}^*)$ are the auxiliary functions defined in equations (11), and Φ the free entropy density of equation (4). Extremising the equation above leads to the following set of self-consistent equations:

$$\begin{cases} \mathbf{q} = \mathbb{E}_\xi \left[\mathcal{Z}_w^*(\hat{\mathbf{q}}^{1/2} \xi, \hat{\mathbf{q}}) f_w^*(\hat{\mathbf{q}}^{1/2} \xi, \hat{\mathbf{q}}) f_w^*(\hat{\mathbf{q}}^{1/2} \xi, \hat{\mathbf{q}})^\top \right], \\ \hat{\mathbf{q}} = \alpha \mathbb{E}_\xi \int_{\mathbb{R}^k} d\mathbf{y} \mathcal{Z}_{\text{out}}^*(\mathbf{y}; \mathbf{q}^{1/2} \xi, \mathbf{Q}^* - \mathbf{q}) f_{\text{out}}^*(\mathbf{y}; \mathbf{q}^{1/2} \xi, \mathbf{Q}^* - \mathbf{q}) f_{\text{out}}^*(\mathbf{y}; \mathbf{q}^{1/2} \xi, \mathbf{Q}^* - \mathbf{q})^\top \end{cases} \quad (14)$$

where $(f_w^*, f_{\text{out}}^*)$ are defined in equations (11). Note the similarity between equations (14) above and equations (10) for the sufficient statistics of ERM. Indeed, the equations above can be obtained from those of ERM via the following mapping, known in the context of statistical physics as *Nishimori conditions* [32]:

$$f_w \rightarrow f_w^*, f_{\text{out}} \rightarrow f_{\text{out}}^*; \quad \mathbf{m} \rightarrow \mathbf{q}, \hat{\mathbf{m}} \rightarrow \hat{\mathbf{q}}; \quad \mathbf{V} \rightarrow \mathbf{Q}^* - \mathbf{q}, \hat{\mathbf{V}} \rightarrow \hat{\mathbf{Q}}^* + \hat{\mathbf{q}}. \quad (15)$$

Intuitively, the student's additional knowledge of the data generating process is translated by choosing the same set of auxiliary functions as the teacher. These conditions imply, on average, no statistical difference between the ground truth configuration and a configuration sampled uniformly at random from the posterior distribution. Therefore, in the BO setting there is no distinction between the teacher–student overlap and the student self-overlap. This connection is further discussed in appendix B, where we show that both sets of equations can be derived from a common framework.

Despite the close similarity between the two sets of self-consistent equations, note one key difference: the set of extrema in equations (14) is not necessarily a single point. This means that, differently from equations (10), the fixed-point of the self-consistent equations (14) might not be unique. In this case, it is important to stress that the overlap \mathbf{q} corresponding to the BO estimator (3) is, by definition, the one with highest free entropy density. Therefore, the BO generalisation error is evaluated by finding the fixed point of equations (14) that maximises the free entropy (13).

A proof of this claim and of equations (14) and (13) for the BO case was done in ([18], Theorem 3.1, and [19]) for the committee machine, by an interpolation method that shows the correctness of the replica prediction for the free-entropy of the system. Their proof applies to teacher–student committee machines with bounded output channel, prior distribution with finite second moment and Gaussian i.i.d. inputs. Therefore, it applies to the multi-class perceptron of our setting, with both priors.

2.3. Generalisation error

The characterisation of the error in the high-dimensional limit is a direct consequence of Theorem 2.1.

Corollary 1. *In the high-dimensional limit the asymptotic generalisation error associated to the ERM estimator (2) can be expressed only as a function of the parameters (\mathbf{m}, \mathbf{q}) obtained by solving the self-consistent equations (10):*

$$\varepsilon_{\text{gen}} = \mathbb{P}_{(\boldsymbol{\nu}, \boldsymbol{\mu}) \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})} (\phi_{\text{out}}(\boldsymbol{\mu}) \neq \phi_{\text{out}}(\boldsymbol{\nu})), \quad \text{where} \quad \boldsymbol{\Sigma} = \begin{bmatrix} \mathbf{Q}^* & \mathbf{m} \\ \mathbf{m} & \mathbf{q} \end{bmatrix}. \quad (16)$$

The proof of Corollary 1 is straightforward and follows by noticing that for any \mathbf{v} , the function $\psi(\boldsymbol{z}) = \mathbb{1}(\phi_{\text{out}}(\boldsymbol{z}) \neq \phi_{\text{out}}(\mathbf{v}))$ is pseudo-Lipschitz.

As one can expect from the discussion in section 2.2, the BO error is obtained by a similar, but simpler expression depending only on the overlap \mathbf{q} , obtained by extremising (13):

$$\varepsilon_{\text{gen}} = \mathbb{P}_{\boldsymbol{\xi} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_k)} \left(\phi_{\text{out}}(\mathbf{q}^{1/2} \boldsymbol{\xi}) \neq \phi_{\text{out}}(\mathbf{Q}^{*1/2} \boldsymbol{\xi}) \right). \quad (17)$$

3. Approximate message-passing algorithm

In order to illustrate our theoretical results for the performance of the Bayesian (3) and ERM (2) estimators, we would like to compare our asymptotic expressions for the generalisation error with finite instance simulations. On one hand, the regularised empirical risk defined in (2) is strongly convex, and therefore it can be readily minimised with any descent-based algorithm such as gradient descent or stochastic gradient descent. Indeed, in the ERM simulations that follow we employ out-of-the-box multi-class solvers from Scikit learn [33] to assess our theoretical result from Theorem 2.1. On the other hand, explicitly computing the Bayesian estimator (3) requires sampling from the posterior, an operation which is prohibitively costly in high-dimensions. Instead, in this manuscript we employ an *Approximate Message Passing* (AMP) algorithm to efficiently approximate the posterior marginals. AMP has several interesting properties which make it a popular tool in the study of random problems. First, it is proven to be optimal among a class of random estimation problems [34], and for this reason it is widely used as a benchmark to assess algorithmic complexity. Second, it admits a set of scalar *state evolution equations* allowing to track its performance in high-dimensions [30].

For the BO estimation problem considered here, AMP is summarised by the pseudo-code in Algorithm 1, which can be found in appendix E. It follows the well-known AMP algorithm for generalised linear estimation [35, 36], which takes advantage of the high-dimensional limit $d \rightarrow \infty$ by approximating the posterior distribution (5) by a multivariate Gaussian through a belief propagation procedure expanded in powers of d^{-1} . The difference is that the estimators $\hat{\mathbf{w}}_j$ are k -dimensional vectors and their variances $\hat{\mathbf{C}}_j$ are $k \times k$ dimensional matrices, $j = 1, \dots, d$. The update functions \mathbf{f}_{out} and $\mathbf{f}_{\mathbf{w}}$ are defined in appendix C.1. For a detailed derivation of the algorithm, see [37].

Several versions of this k -fold AMP and the associated state evolution appeared in previous works, e.g. [18]. It can be shown that the state evolution equations associated to Algorithm 1 for BO estimation coincide exactly with the self-consistent equations (14) presented in section 2.2 starting from an *uninformed*

initialisation $\mathbf{q}_0 \approx \mathbf{0}$ [18]. This interesting property implies that when the extremisation problem in equation (13) has only one extremiser, AMP provides an exact approximation to the BO estimator in the high-dimensional limit. Instead, when there are more than one maxima in equation (13), AMP will converge to an estimator with overlap \mathbf{q} closest to the uninformed initial condition. If this is not the global maximum, this corresponds to a situation where AMP differs from the BO estimator. Since AMP provides a bound on the performance of first-order algorithms, this situation is an example of an *algorithmic hard phase*, where it is conjectured that the statistical optimal performance cannot be achieved by algorithms running in time $\sim O(d^2)$.

We have implemented Algorithm 1 for $k = 3$ classes using the mapping presented in appendix A, which makes the estimators $(k - 1)$ -dimensional vectors and their variances $(k - 1) \times (k - 1)$ dimensional matrices. The detailed expressions for the computation of the denoising functions, as well as the integrals to be numerically evaluated are presented in appendix F.

4. Results for $k = 3$ classes

In this section we apply our theoretical results to the case of $k = 3$ classes and compare them with numerical simulations. The prior reduction discussed in appendix A allows us to implement easily and efficiently the numerical experiments in this case. Although our theory is valid for any finite k , the experiments at k larger than 3 are computationally more demanding and we leave their implementation to future work.

We investigate the dependence of the learning curves on the sample complexity α . First, we consider the case of Rademacher teacher prior and show that a first-order phase transition arises in the BO performance. Then, we turn to Gaussian teacher prior and explore the role of the regularisation strength λ in approaching the BO performance with ERM. For an overview on the empirical and theoretical literature on learning curves, see [28].

4.1. Bayes-optimal performance for Rademacher teacher

The main difference between Gaussian and Rademacher teacher is that in the second case perfect generalisation is achievable at finite sample complexity, in line with the results known for the two-classes case of [2, 5, 6]. To compute the optimal information-theoretical performance, we have evaluated the global extremum of the replica free entropy. To this end, we have run the replica saddle point iterations equations (14) with both uninformed and informed initialisations and computed the free entropy (13) of the fixed points (if distinct) reached by the two initialisations. In figure 1 we report the generalisation error corresponding to the fixed points reached by the two initialisations, along with their corresponding free entropy in the inset. We found that indeed, for Rademacher teacher weights, the generalisation error decreases continuously for $\alpha \leq \alpha_{\text{IT}}^{(k=3)} \approx 2.45$, and then jumps to zero for all $\alpha > \alpha_{\text{IT}}^{(k=3)}$. From a statistical physics perspective, this discontinuous transition in the error corresponds to a *first-order phase transition* associated to the discontinuous appearance of a second extremum associated to perfect learning in the free energy potential. As we have previously discussed, the state evolution of the AMP Algorithm 1 is equivalent to gradient descent on the free energy potential (13) starting from an uninformed random initialisation. Therefore, the appearance of a second extremum away from zero implies that AMP is not able to achieve the BO statistical performance. Since AMP is conjectured to be optimal among first-order methods [34], this result is an example of a fundamental *statistical-to-algorithmic gap* in this problem. For $\alpha > \alpha_{\text{algo}}^{(k=3)} \approx 2.89$, we observe that the uninformed minimum disappears, and we can check that this coincides with the sample complexity at which AMP is able to achieve zero generalisation error from random initialisation. This marks the algorithmic threshold, i.e. the sample complexity beyond which perfect generalisation is reachable algorithmically efficiently. Our findings thus suggest the existence of an algorithmic *hard phase* for $\alpha_{\text{IT}}^{(k=3)} < \alpha < \alpha_{\text{algo}}^{(k=3)}$, where the theoretically optimal performance is not reachable by efficient algorithms.

We note here the comparison with the canonical perceptron with Rademacher teacher weights and two classes, where the same thresholds are well known to be $\alpha_{\text{IT}}^{(k=2)} = 1.249$, $\alpha_{\text{algo}}^{(k=2)} = 1.493$ [5–7]. Naturally, these values are roughly twice smaller than the ones for $k = 3$ since for k classes the teacher has $k - 1$ independent d -dimensional binary elements that need to be recovered in order to reach perfect generalisation. Comparing more precisely the values for $k = 3$ and also their difference, all are slightly smaller than the double of the values for $k = 2$.

4.2. Bayes-optimal performance for Gaussian teacher

Figures 2–4 summarise our results for the case of Gaussian teacher weights. The BO error, computed from equation (7), is depicted by the dashed black line in both figures and is a smooth, monotonically-decreasing function of the sample complexity α . Interestingly, for Gaussian teacher weights, the Bayes-optimal AMP algorithm—described in section 3 and marked by the green symbols in figure 2—achieves the BO

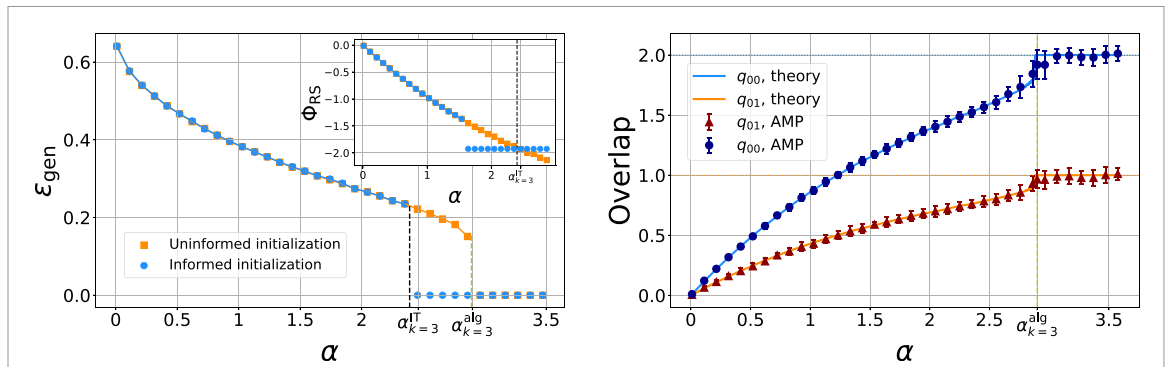


Figure 1. AMP for Rademacher teacher prior with $k = 3$ classes. *Left:* Generalisation error ϵ_{gen} as a function of α evaluated via equations (17). The orange points mark the error asymptotically reached by the randomly initialised AMP. The blue points mark the BO error. The inset depicts the corresponding free entropies, their crossing locating the information-theoretic transition to perfect generalisation at $\alpha_{k=3}^{\text{IT}} \approx 2.45$. AMP reaches perfect generalisation starting from $\alpha_{k=3}^{\text{alg}} \approx 2.89$. *Right:* Diagonal (q_{00}) and anti-diagonal (q_{01}) entries of the self-overlap matrix as a function of α in the BO setting. The full lines mark the fixed points of equations (14), the symbols represent the result obtained by the AMP algorithm averaged over 20 runs.

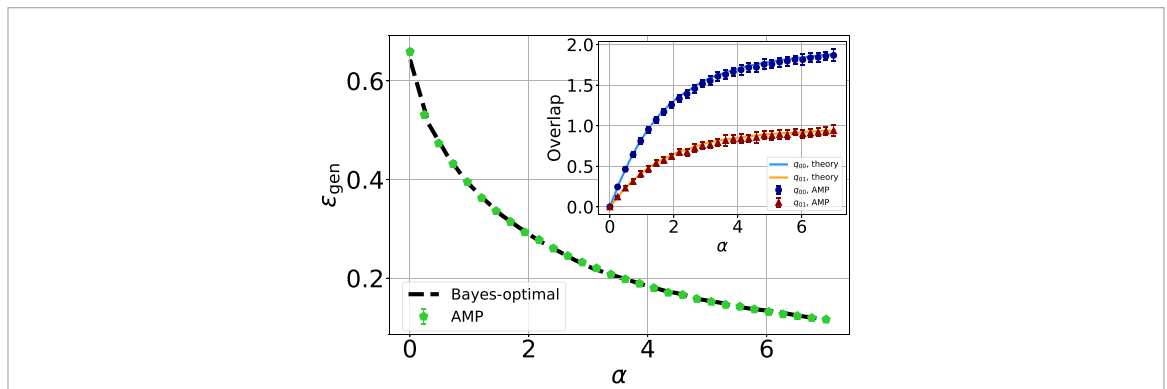


Figure 2. AMP for Gaussian teacher prior: Generalisation error ϵ_{gen} as a function of α . The green symbols mark the performance of AMP (averaged over 20 runs). The dashed black line marks the BO error. The inset displays the diagonal (q_{00}) and anti-diagonal (q_{01}) entries of the overlap matrix in the BO setting. The full lines mark the fixed points of equations (14), while the symbols represent the result obtained from the AMP algorithm.

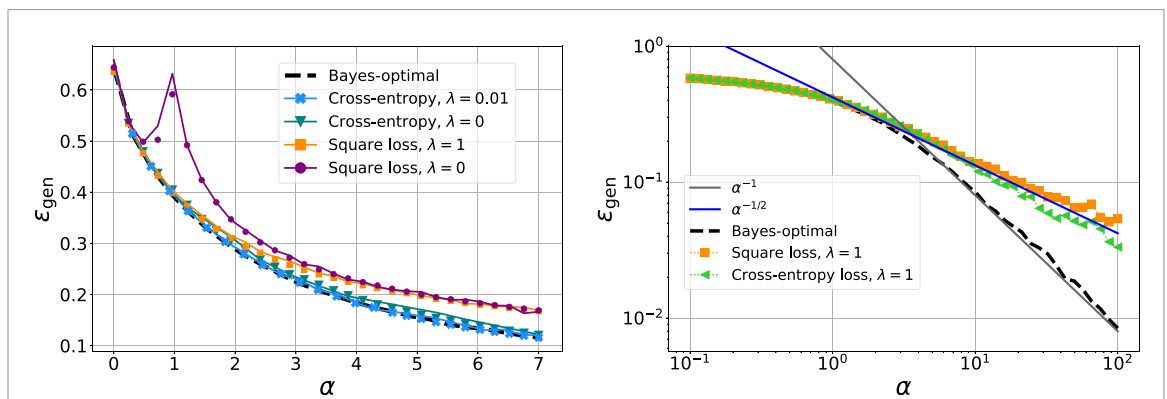
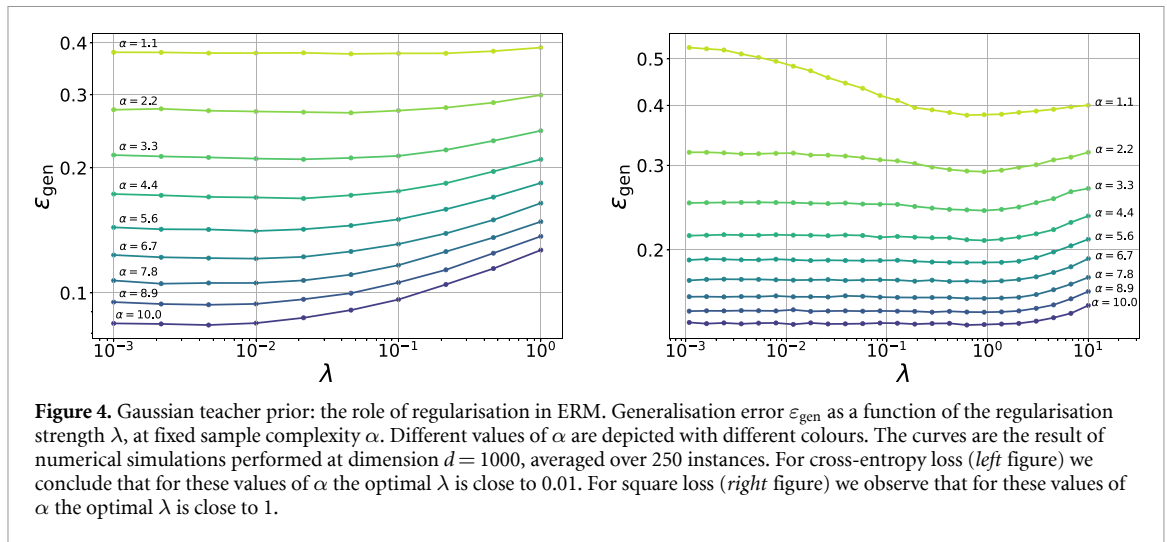


Figure 3. Bayes-optimal and ERM performances for Gaussian teacher weights. *Left:* Generalisation error ϵ_{gen} as a function of the sample complexity α . The dashed line marks the BO error. Full lines mark the performance of ERM with cross-entropy (blue) and square loss (orange), both at optimised ridge regularisation ($\lambda = 0.01$ and $\lambda = 1$ respectively, see figure 4) from the fixed points of equation (10). The symbols mark the results from numerical simulations at $d = 1000$, averaged over 250 seeds. We also plot the performance of simulations at zero regularisation and theory at $\lambda \rightarrow 0^+$, for both cross-entropy (teal) and square loss (purple). *Right:* Large- α behaviour of the error. The dashed line marks the BO error, the symbols mark ERM at fixed $\lambda = 1$.

performance. This is highly non-trivial: computing the Bayesian estimator usually requires sampling from the posterior distribution of the weights given the data, and therefore can be prohibitively costly in the high-dimensional regime considered here. For Gaussian weights AMP provides an exact approximation of the posterior marginals in quadratic time in the input dimension.



4.3. Approaching Bayes-optimality with ERM

Instead, how does ERM compare to the Bayesian estimator? Note that the empirical risk in equation (2) is convex, and therefore, at variance with the posterior estimation, this problem can be readily simulated using descent-based algorithms such as stochastic gradient descent. The generalisation error obtained by ERM is plotted in figure 3 as a function of the sample complexity. The full lines depict our theoretical predictions for the learning curves while the symbols mark the results from numerical simulations performed at finite dimension $d = 1000$ (more details on the numerics are provided in appendix G). We find excellent agreement between the two. For both cross-entropy and square losses, we show the performance achieved without regularisation ($\lambda = 0$) and with optimal λ , obtained by cross-validation on a fixed grid, in figure 4. Interestingly, we find that the optimally-regularised cross-entropy loss achieves a close-to-optimal performance, while the square loss maintains a finite gap with respect to the BO error even at fine-tuned regularisation strength. Similar results were obtained for the two-classes teacher student perceptron [8]. The fact that regularised cross-entropy minimisation is so close to optimal also in multi-class classification is remarkable and the generality of this finding is worth further investigation.

4.4. Large- α behaviour

Figure 3 (right) considers again a Gaussian teacher prior and explores the behaviour of the generalisation error at large sample complexity. The BO performance is depicted in black and decays as $1/\alpha$ in the large- α regime. On the other hand, the performance obtained by ERM at fixed λ displays a slower decay $\alpha^{-1/2}$. This is again compatible with the behaviour observed in the two-classes case [8]. It remains to be analysed whether for $k > 2$ the optimally regularised ERM achieves the $1/\alpha$ rate as it does for the two classes.

4.5. The role of regularisation

Figure 4 further illustrates the role played by ridge regularisation. We plot the generalisation error as a function of the regularisation strength λ at fixed sample complexity α for the cross-entropy (left) and the square loss (right). Different curves represent different values of sample complexity. We observe that the optimal regularisation depends only very mildly on the sample complexity α for this range of values of α .

Data availability statement

No new data were created or analysed in this study.

Acknowledgment

We acknowledge funding from the ERC under the European Union's Horizon 2020 Research and Innovation Programme Grant Agreement 714608-SMiLe. R V was partially financed by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior—Brasil (CAPES) - Finance Code 001. R V is grateful to EPFL and IdePHICS lab for their generous hospitality during the realization of this project. This work started as a part of the doctoral course Statistical Physics For Optimization and Learning taught at EPFL in spring 2021.

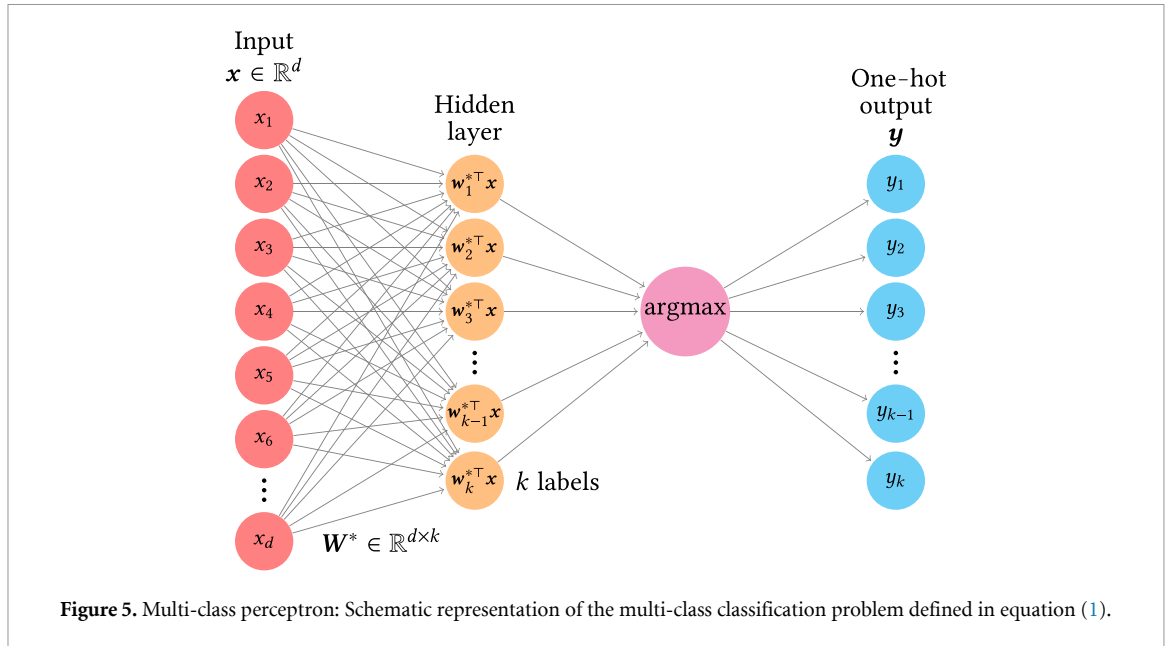


Figure 5. Multi-class perceptron: Schematic representation of the multi-class classification problem defined in equation (1).

Appendix A. Prior reduction

In this section we explain the mapping from k to $k - 1$ dimensions that we apply to evaluate our theoretical results from equation (10) as well as to implement Algorithm 1. Figure 5 displays schematically the multi-class perceptron. The intuition for the prior reduction is exactly the same of the binary perceptron: the knowledge of $k - 1$ components of the one-hot label representation \mathbf{y} is enough to determine the remaining component. Nevertheless for $k > 2$, shifting the weights in order to reproduce this structure introduces additional correlations that must be taken into account.

We recall that \mathbf{W}^* is a $d \times k$ matrix, and denote by \mathbf{w}_l^* , $1 \leq l \leq k$, its columns, each corresponding to a different class. Notice that the label $\mathbf{y} = \mathbf{e}_{\text{argmax}_l(\{\mathbf{w}_l^{*\top} \mathbf{x}\}_{l \in [k]})}$ given by equation (1) of a data point \mathbf{x} can be equivalently expressed by taking the k th-component, i.e. $\mathbf{w}_k^{*\top} \mathbf{x}$, as a reference for comparison and setting

$$\tilde{\mathbf{w}}_h^* \leftarrow \mathbf{w}_h^* - \mathbf{w}_k^* \quad \text{for all } 1 \leq h \leq k, \tag{A.1}$$

so that $\tilde{\mathbf{w}}_k^* = \mathbf{0}$, and the problem is reduced to $k - 1$ dimensions. We then replace \mathbf{W}^* by $\tilde{\mathbf{W}}^* \in \mathbb{R}^{d \times (k-1)}$. Denoting $\mathbf{1}_k$ as the k -dimensional vector with all entries equal to 1, we present schematically in figure 6 the prior reduction.

Note that this mapping introduces correlations along the columns of $\tilde{\mathbf{W}}^*$, but not along the rows, i.e. the d components of each vector $\tilde{\mathbf{w}}_i^*$ remain i.i.d. Therefore, the prior over the weights is still factorizable along the extensive dimension d .

A.1. Gaussian prior

In the Gaussian prior setting,

$$P_w(\mathbf{w}^*) = \mathcal{N}(\mathbf{W}^* | \mathbf{0}, \mathbf{I}_k) = \frac{1}{(2\pi)^{k/2}} \exp\left(-\frac{1}{2} \mathbf{w}^{*\top} \mathbf{w}^*\right), \tag{A.2}$$

where $\mathbf{w}^* \in \mathbb{R}^k$ is a column of the matrix \mathbf{W}^* , the transformation imposes a new covariance matrix with elements

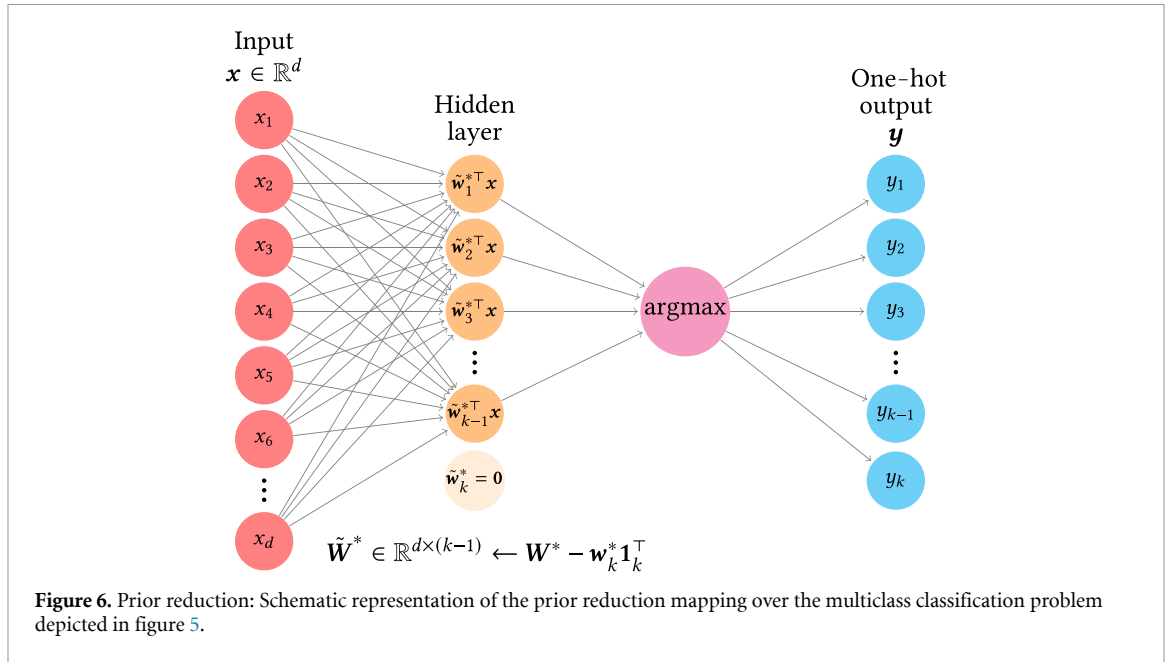
$$\Sigma_{jl} \equiv \text{Cov}\left(\mathbf{w}_j^* - \mathbf{w}_k^*, \mathbf{w}_l^* - \mathbf{w}_k^*\right). \tag{A.3}$$

By making use of the identity

$$\text{Cov}(\alpha \mathbf{a} + \beta \mathbf{b}, \gamma \mathbf{c} + \delta \mathbf{d}) = \alpha \gamma \text{Cov}(\mathbf{a}, \mathbf{c}) + \alpha \delta \text{Cov}(\mathbf{a}, \mathbf{d}) + \beta \gamma \text{Cov}(\mathbf{b}, \mathbf{c}) + \beta \delta \text{Cov}(\mathbf{b}, \mathbf{d}), \tag{A.4}$$

one can write

$$\mathcal{N}(\mathbf{W}^* | \mathbf{w}_k^*, \Sigma) \propto \exp\left[-\frac{1}{2} (\mathbf{w}^* - \mathbf{w}_k^*)^\top \Sigma^{-1} (\mathbf{w}^* - \mathbf{w}_k^*)\right] \leftarrow \mathcal{N}(\mathbf{W}^* | \mathbf{0}, \mathbf{I}_k), \tag{A.5}$$



with

$$\Sigma = \mathbf{I}_k - \mathbf{e}_k \mathbf{e}_k^\top + \mathbf{e}_k \sum_{l \neq k} \mathbf{e}_l^\top + \left(\sum_{l \neq k} \mathbf{e}_l \right) \mathbf{e}_k^\top, \tag{A.6}$$

Thus, since all contributions related to the k th degree of freedom become zero, the transformation (A.1) allows us to write the mapping

$$\mathbf{I}_{k-1} \leftarrow \mathbf{I}_k - \mathbf{e}_k \mathbf{e}_k^\top, \tag{A.7a}$$

$$\mathbf{1}_{k-1} \mathbf{1}_{k-1}^\top \leftarrow \mathbf{e}_k^\top + \mathbf{e}_k \sum_{l \neq k} \mathbf{e}_l^\top + \left(\sum_{l \neq k} \mathbf{e}_l \right) \mathbf{e}_k^\top; \tag{A.7b}$$

and finally for $\tilde{\mathbf{w}}^* \in \mathbb{R}^{k-1}$:

$$\mathcal{N}(\tilde{\mathbf{w}}^* | \mathbf{0}, \tilde{\Sigma}) \propto \exp \left(-\frac{1}{2} \tilde{\mathbf{w}}^{*\top} \tilde{\Sigma}^{-1} \tilde{\mathbf{w}}^* \right), \tag{A.8a}$$

with covariance $\tilde{\Sigma} \in \mathbb{R}^{(k-1) \times (k-1)}$ given by

$$\tilde{\Sigma} = \mathbf{I}_{k-1} + \mathbf{1}_{k-1} \mathbf{1}_{k-1}^\top. \tag{A.8b}$$

Therefore each row of the reduced matrix $\tilde{\mathbf{W}}^*$ follows a Gaussian distribution with $\mathbf{0}$ mean and covariance matrix given by equation (A.8b).

A.2. Rademacher prior

In the Rademacher setting, $\mathbf{w}^* \in \mathbb{R}^k$,

$$P_w(\mathbf{W}^*) = \frac{1}{2^k} \prod_{l=1}^k [\delta(w_l^* + 1) + \delta(w_l^* - 1)], \tag{A.9}$$

we can write

$$\begin{aligned} P_w(\mathbf{W}^*) &= \frac{1}{2^k} \prod_{l=1}^k [\delta(w_l^* - w_k^* + w_k^* + 1) + \delta(w_l^* - w_k^* + w_k^* - 1)] \\ &= \frac{1}{2^k} [\delta(w_k^* + 1) + \delta(w_k^* - 1)] \prod_{l=1}^{k-1} [\delta((w_l^* - w_k^*) + w_k^* + 1) + \delta((w_l^* - w_k^*) + w_k^* - 1)], \end{aligned} \tag{A.10}$$

leading to the reduced prior for $\tilde{\mathbf{w}}^* \in \mathbb{R}^{k-1}$:

$$P_{\tilde{\mathbf{w}}^*}(\tilde{\mathbf{w}}^* | \mathbf{w}_k^*) = \frac{1}{2^k} [\delta(\mathbf{w}_k^* + 1) + \delta(\mathbf{w}_k^* - 1)] \prod_{l=1}^{k-1} [\delta(\tilde{\mathbf{w}}_l^* + \mathbf{w}_k^* + 1) + \delta(\tilde{\mathbf{w}}_l^* + \mathbf{w}_k^* - 1)]. \quad (\text{A.11})$$

The dimensional reduction simplifies our analysis when it comes to the numerical evaluation both of the Gaussian integrals in equation (10) and of the prior and channel updates of AMP, as discussed in appendix F. The same reformulation applies straightforwardly to the parameters \mathbf{m} and \mathbf{q} .

Appendix B. Replica calculation

In this section, we carry out the (heuristic) replica computation leading to the system of equations (10) in the main text. We consider a general setting where the student has access to a prior distribution P_w over the teacher weights and a model distribution P_{out} , which can be the true ones or not. This formulation encompasses both the Bayes-optimal and non Bayes-optimal settings. As we shall see in the following, ERM can be seen as a special case of the latter. The posterior distribution of the student weights is given by

$$P(\{\mathbf{w}_l\}_{l=1}^k | \mathbf{X}, \mathbf{Y}) = \frac{1}{Z_d} \prod_{l=1}^k P_w(\mathbf{W}_l) \prod_{\mu=1}^n P_{\text{out}}(\mathbf{y}_\mu | \{h_{\mu l}\}_{l=1}^k) \quad (\text{B.1})$$

where we have defined $h_{\mu l} = \mathbf{w}_l^\top \mathbf{x}_\mu / \sqrt{d}$. The partition function is then

$$Z_d = \int_{\mathbb{R}^{d \times k}} d\mathbf{w} \prod_{l=1}^k P_w(\mathbf{W}_l) \prod_{\mu=1}^n P_{\text{out}}(\mathbf{y}_\mu | \{h_{\mu l}\}_{l=1}^k). \quad (\text{B.2})$$

By using the *replica trick*, we can compute the free entropy in the high-dimensional limit as

$$\Phi := \lim_{d \rightarrow \infty} \Phi_d := \lim_{d \rightarrow \infty} \frac{1}{d} \mathbb{E}_{\mathbf{X}, \mathbf{W}^*} \ln Z_d \approx \lim_{d \rightarrow \infty} \lim_{p \rightarrow 0^+} \frac{1}{d} \frac{\partial}{\partial p} \mathbb{E}_{\mathbf{X}, \mathbf{W}^*} Z_d^p. \quad (\text{B.3})$$

We can then rewrite the average in equation (B.3) as

$$\mathbb{E}_{\mathbf{X}, \mathbf{W}^*} Z_d^p = \mathbb{E}_{\mathbf{X}, \mathbf{W}^*} \left[\int_{\mathbb{R}^{d \times k}} d\mathbf{w} \prod_{l=1}^k P_w(\mathbf{W}_l) \prod_{\mu=1}^n P_{\text{out}}(\mathbf{y}_\mu | \{h_{\mu l}\}_{l=1}^k) \right]^p \quad (\text{B.4})$$

$$= \mathbb{E}_{\mathbf{X}, \mathbf{W}^*} \left[\prod_{a=1}^p \int_{\mathbb{R}^{d \times k}} d\mathbf{w}^a \prod_{l=1}^k P_w^a(\mathbf{W}_l^a) \prod_{\mu=1}^n P_{\text{out}}(\mathbf{y}_\mu | \{h_{\mu l}^a\}_{l=1}^k) \right] \quad (\text{B.5})$$

$$= \mathbb{E}_{\mathbf{X}} \int_{\mathbb{R}^{n \times k}} d\mathbf{Y} \prod_{a=0}^p \left[\int_{\mathbb{R}^{d \times k}} d\mathbf{w}^a \prod_{l=1}^k P_w^a(\mathbf{W}_l^a) \prod_{\mu=1}^n P_{\text{out}}^a(\mathbf{y}_\mu | \{h_{\mu l}^a\}_{l=1}^k) \right], \quad (\text{B.6})$$

where above we have renamed $\mathbf{w}^* = \mathbf{w}^0$. In order to account for both the Bayes-optimal and non-Bayes-optimal cases, we keep the distinction between teacher and student distributions by adding an index a to prior and model distributions. In what follows, $P_w^0 = P_w^*$ and $P_{\text{out}}^0 = P_{\text{out}}^*$ refer to the teacher, while $P_w^{a>0} = P_w$ and $P_{\text{out}}^{a>0} = P_{\text{out}}$ to the student. Let us denote the covariance tensor of the $h_{\mu l}^a$ as

$$\mathbb{E}[h_{\mu l}^a h_{\nu l'}^b] = \delta_{\mu\nu} Q_{bl'}^a, \quad (\text{B.7})$$

$$Q_{bl'}^a = \frac{1}{d} \sum_{i=1}^d w_{il}^a w_{il'}^b, \quad (\text{B.8})$$

with $\mathbf{Q}_a^b \in \mathbb{R}^{k \times k}$. We can rewrite the above as

$$\begin{aligned} \mathbb{E}_{\mathbf{X}, \mathbf{W}^*} Z_d^p &= \mathbb{E}_{\mathbf{X}} \int_{\mathbb{R}^{n \times k}} d\mathbf{Y} \prod_{a=0}^p \left[\int_{\mathbb{R}^{d \times k}} d\mathbf{w}^a \prod_{l=1}^k P_w^a(\mathbf{W}_l^a) \prod_{\mu=1}^n P_{\text{out}}^a(\mathbf{y}_\mu | \{h_{\mu l}^a\}_{l=1}^k) \right] \\ &= \prod_{(a,l);(b,l')} \int_{\mathbb{R}} dQ_{bl'}^a I_{\text{prior}}(\{Q_{bl'}^a\}) I_{\text{channel}}(\{Q_{bl'}^a\}), \end{aligned} \quad (\text{B.9})$$

where we have denoted

$$I_{\text{prior}}(\{Q_{bl'}^{al}\}) = \prod_{a=0}^p \int_{\mathbb{R}^{d \times k}} d\mathbf{w}^a \left[\prod_{l=1}^k P_w^a(\mathbf{W}_l^a) \right] \prod_{(a,l);(b,l')} \delta \left(Q_{bl'}^{al} - \frac{1}{d} \sum_{i=1}^d w_{il}^a w_{il'}^b \right), \tag{B.10}$$

$$I_{\text{channel}}(\{Q_{bl'}^{al}\}) = \int_{\mathbb{R}^{n \times k}} d\mathbf{Y} \prod_{a=0}^p \int_{\mathbb{R}^{d \times k}} dh^a \left[\prod_{a=0}^p \prod_{\mu=1}^n P_{\text{out}}^a(\mathbf{y}_\mu | h_\mu^a) \right] \tag{B.11}$$

$$\times \exp \left(-\frac{n}{2} \ln \det \mathbf{Q} - \frac{nk(p+1)}{2} \ln 2\pi - \frac{1}{2} \sum_{\mu=1}^n \sum_{a,b} \sum_{l,l'} h_{\mu l}^a (Q^{-1})_{bl'}^{al} h_{\mu l'}^b \right), \tag{B.12}$$

and we have introduced both the definitions of the overlaps $\{Q_{bl'}^{al}\}$ and the local fields $\{h_{\mu l}^a\}$. We can introduce the Fourier representation of the Dirac δ -functions in the prior term I_{prior} and rewrite

$$\mathbb{E}Z_d^p = \prod_{(a,l);(b,l')} \int_{\mathbb{R}^2} \frac{dQ_{bl'}^{al} d\hat{Q}_{bl'}^{al}}{2\pi} \exp(dH(\mathbf{Q}, \hat{\mathbf{Q}})), \tag{B.13}$$

where we have defined

$$H(\mathbf{Q}, \hat{\mathbf{Q}}) := \frac{1}{2} \sum_{a=0}^p \sum_{l,l'} Q_{al'}^{al} \hat{Q}_{al'}^{al} - \frac{1}{2} \sum_{a \neq b} \sum_{l,l'} Q_{bl'}^{al} \hat{Q}_{bl'}^{al} + \ln I(\{\hat{Q}_{bl'}^{al}\}) + \alpha \ln J(\{Q_{bl'}^{al}\}) \tag{B.14}$$

and the auxiliary functions:

$$I(\{\hat{Q}_{bl'}^{al}\}) = \prod_{a=0}^p \int_{\mathbb{R}^k} d\mathbf{w}^a P_w^a(\mathbf{W}^a) \exp \left(-\frac{1}{2} \sum_{a=0}^p \sum_{l,l'} w_l^a \hat{Q}_{al'}^{al} w_{l'}^a + \frac{1}{2} \sum_{a \neq b} \sum_{l,l'} w_l^a \hat{Q}_{bl'}^{al} w_{l'}^b \right), \tag{B.15}$$

$$J(\{Q_{bl'}^{al}\}) = \int_{\mathbb{R}^k} d\mathbf{y} \prod_{a=0}^p \int_{\mathbb{R}^k} \frac{d\mathbf{h}^a}{(2\pi)^{k(p+1)/2}} \frac{P_{\text{out}}^a(\mathbf{y} | \mathbf{h}^a)}{\sqrt{\det \mathbf{Q}}} \exp \left(-\frac{1}{2} \sum_{a,b} \sum_{l,l'} h_l^a (Q^{-1})_{bl'}^{al} h_{l'}^b \right). \tag{B.16}$$

We observe that, upon exchanging the limits in d and p , the high-dimensional limit of the free entropy can be computed via a saddle-point method:

$$\Phi = \lim_{d \rightarrow \infty} \mathbb{E}_{\mathbf{X}, \mathbf{W}^*} \ln \mathcal{Z}_d = \lim_{p \rightarrow 0^+} \text{extr}_{\mathbf{Q}, \hat{\mathbf{Q}}} [H(\mathbf{Q}, \hat{\mathbf{Q}})]. \tag{B.17}$$

B.1. Replica symmetric ansatz

In order to progress in the calculation, we restrict the extremisation in equation (B.17) to values of $\{\mathbf{Q}, \hat{\mathbf{Q}}\}$ described by a replica symmetric (RS) ansatz [32]. The validity of this ansatz is proven rigorously in appendix D. We distinguish between the Bayes-optimal and non Bayes-optimal cases. Note that in the Bayes-optimal case we can drop the a -index from the prior and model distributions. In the non Bayes-optimal case, we will denote the teacher distributions by P_w^*, P_{out}^* and the student ones simply by P_w, P_{out} .

B.1.1. Bayes-optimal RS ansatz

In the Bayes-optimal setting we make the following ansatz:

$$Q_{al'}^{al} = Q_{ll'}^*, \quad \hat{Q}_{al'}^{al} = \hat{Q}_{ll'}^*, \quad \forall a = 0, \dots, p, \forall l, l' \leq k \tag{B.18}$$

$$Q_{bl'}^{al} = q_{ll'}, \quad \hat{Q}_{bl'}^{al} = \hat{q}_{ll'}, \quad \forall a \neq b, \forall l, l' \leq k. \tag{B.19}$$

The trace term is simplified as follows

$$\frac{1}{2} \sum_{a=0}^p \sum_{l,l'} Q_{al'}^{al} \hat{Q}_{al'}^{al} - \frac{1}{2} \sum_{a \neq b} \sum_{l,l'} Q_{bl'}^{al} \hat{Q}_{bl'}^{al} = \frac{1}{2} (p+1) \sum_{l,l'} \hat{Q}_{ll'}^* Q_{ll'}^* - \frac{1}{2} p(p+1) \sum_{ll'} \hat{q}_{ll'} q_{ll'}. \tag{B.20}$$

The prior and output terms can be simplified by performing a Hubbard–Stratonovich transformation that allows to decouple the replica indices a, b . By indicating the standard Gaussian measure with $\mathcal{D}\xi$: $\xi \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_k)$, we obtain

$$I(\hat{\mathbf{Q}}^*, \hat{\mathbf{q}}) = \int_{\mathbb{R}^k} \mathcal{D}\xi \left[\int_{\mathbb{R}^k} d\mathbf{w} P_w^*(\mathbf{W}) \exp \left(-\frac{1}{2} \mathbf{w}^\top (\hat{\mathbf{Q}}^* + \hat{\mathbf{q}}) \mathbf{w} + \xi^\top \hat{\mathbf{q}}^{\frac{1}{2}} \mathbf{w} \right) \right]^{p+1}, \tag{B.21}$$

$$J(\mathbf{Q}^*, \mathbf{q}) = \int_{\mathbb{R}^k} d\mathbf{y} \int_{\mathbb{R}^k} \mathcal{D}\xi \left[\int_{\mathbb{R}^k} \mathcal{D}\mathbf{h} P_{\text{out}}^*(\mathbf{y} | (\mathbf{Q}^* - \mathbf{q})^{\frac{1}{2}} \mathbf{h} + \mathbf{q}^{\frac{1}{2}} \xi) \right]^{p+1}. \tag{B.22}$$

Since we are interested in the $p \rightarrow 0^+$ limit, it is useful to rewrite

$$\begin{aligned} \ln I(\hat{\mathbf{Q}}^*, \hat{\mathbf{q}}) &= p \int_{\mathbb{R}^k} \mathcal{D}\xi \int_{\mathbb{R}^k} d\mathbf{w} P_w^*(\mathbf{W}) \exp \left(-\frac{1}{2} \mathbf{w}^\top (\hat{\mathbf{Q}}^* + \hat{\mathbf{q}}) \mathbf{w} + \xi^\top \hat{\mathbf{q}}^{\frac{1}{2}} \mathbf{w} \right) \\ &\quad \times \ln \int_{\mathbb{R}^k} d\mathbf{w}' P_w^*(\mathbf{W}') \exp \left(-\frac{1}{2} \mathbf{w}'^\top (\hat{\mathbf{Q}}^* + \hat{\mathbf{q}}) \mathbf{w}' + \xi^\top \hat{\mathbf{q}}^{\frac{1}{2}} \mathbf{w}' \right) + o(p), \end{aligned} \tag{B.23}$$

$$\begin{aligned} \ln J(\mathbf{Q}^*, \mathbf{q}) &= p \int_{\mathbb{R}^k} d\mathbf{y} \int_{\mathbb{R}^k} \mathcal{D}\xi \int_{\mathbb{R}^k} \mathcal{D}\mathbf{h} P_{\text{out}}^*(\mathbf{y} | (\mathbf{Q}^* - \mathbf{q})^{\frac{1}{2}} \mathbf{h} + \mathbf{q}^{\frac{1}{2}} \xi) \\ &\quad \times \ln \int_{\mathbb{R}^k} \mathcal{D}\mathbf{h}' P_{\text{out}}^*(\mathbf{y} | (\mathbf{Q}^* - \mathbf{q})^{\frac{1}{2}} \mathbf{h}' + \mathbf{q}^{\frac{1}{2}} \xi) + o(p). \end{aligned} \tag{B.24}$$

B.1.2. Non-Bayes-optimal RS ansatz

In the non-Bayes-optimal setting we make the following ansatz:

$$Q_{al'}^a = Q_{ll'}^0, \quad \hat{Q}_{al'}^a = \hat{Q}_{ll'}^0, \quad \forall a = 1, \dots, p, \forall l, l' \leq k \tag{B.25}$$

$$Q_{bl'}^a = q_{ll'}, \quad \hat{Q}_{bl'}^a = \hat{q}_{ll'}, \quad \forall a \neq b, a, b = 1, \dots, p, \forall l, l' \leq k \tag{B.26}$$

$$Q_{al'}^{0l} = m_{ll'}, \quad \hat{Q}_{al'}^{0l} = \hat{m}_{ll'}, \quad \forall a = 1, \dots, p, \forall l, l' \leq k \tag{B.27}$$

$$Q_{0l'}^{0l} = Q_{ll}^*, \quad \hat{Q}_{0l'}^{0l} = \hat{Q}_{ll}^*, \quad \forall l, l' \leq k. \tag{B.28}$$

The trace term is simplified as follows

$$\begin{aligned} \frac{1}{2} \sum_{a=0}^p \sum_{l,l'} Q_{al'}^a \hat{Q}_{al'}^a - \frac{1}{2} \sum_{a \neq b} \sum_{l,l'} Q_{bl'}^a \hat{Q}_{bl'}^a &= \frac{1}{2} p \sum_{l,l'} \hat{Q}_{ll}^0 Q_{ll}^0 - \frac{1}{2} p(p-1) \sum_{ll'} \hat{q}_{ll'} q_{ll'} \\ &\quad + \frac{1}{2} \sum_{l,l'} \hat{Q}_{ll}^* Q_{ll}^* - p \sum_{l,l'} m_{ll'} \hat{m}_{ll'}. \end{aligned} \tag{B.29}$$

The prior term is

$$\begin{aligned} I(\hat{\mathbf{Q}}^0, \hat{\mathbf{q}}, \hat{\mathbf{Q}}^*, \hat{\mathbf{m}}) &= \int_{\mathbb{R}^k} \mathcal{D}\xi \int_{\mathbb{R}^k} d\mathbf{w}^* P_w^*(\mathbf{W}^*) \exp \left(-\frac{1}{2} \mathbf{w}^{*\top} \hat{\mathbf{Q}}^* \mathbf{w}^* \right) \\ &\quad \times \left[\int_{\mathbb{R}^k} d\mathbf{w} P_w(\mathbf{W}) \exp \left(-\frac{1}{2} \mathbf{w}^\top (\hat{\mathbf{Q}}^0 + \hat{\mathbf{q}}) \mathbf{w} + \mathbf{w}^{*\top} \hat{\mathbf{m}} \mathbf{w} + \xi^\top \hat{\mathbf{q}}^{\frac{1}{2}} \mathbf{w} \right) \right]^p. \end{aligned} \tag{B.30}$$

In order to compute the output term we need to compute the inverse matrix

$$\mathbf{Q}^{-1} = \begin{bmatrix} \tilde{\mathbf{Q}}^* & \tilde{\mathbf{m}} & \dots & \tilde{\mathbf{m}} \\ \tilde{\mathbf{m}} & \tilde{\mathbf{Q}}^0 & \tilde{\mathbf{q}} & \dots \\ \vdots & \tilde{\mathbf{q}} & \ddots & \tilde{\mathbf{q}} \\ \tilde{\mathbf{m}} & \dots & \tilde{\mathbf{q}} & \tilde{\mathbf{Q}}^0 \end{bmatrix} \in \mathbb{R}^{k(p+1) \times k(p+1)}, \tag{B.31}$$

which has a similar block structure as \mathbf{Q} . The components of the inverse can be computed from the relation $\mathbf{Q}^{-1}\mathbf{Q} = \mathbf{I}_k$ and are given by

$$\begin{aligned} \tilde{\mathbf{Q}}^* &= \left(\mathbf{Q}^* - p\mathbf{m}(\mathbf{Q}^0 + (p-1)\mathbf{q})^{-1}\mathbf{m} \right)^{-1}, \\ \tilde{\mathbf{Q}}^0 &= (\mathbf{Q}^0 - \mathbf{q})^{-1} + (\mathbf{Q}^0 + (p-1)\mathbf{q})^{-1} \\ &\quad \times \left[\mathbf{m} \left(\mathbf{Q}^* - p\mathbf{m}(\mathbf{Q}^0 + (p-1)\mathbf{q})^{-1}\mathbf{m} \right)^{-1} \mathbf{m} (\mathbf{Q}^0 + (p-1)\mathbf{q})^{-1} - \mathbf{q}(\mathbf{Q}^0 - \mathbf{q})^{-1} \right], \\ \tilde{\mathbf{q}} &= \tilde{\mathbf{Q}}^0 - (\mathbf{Q}^0 - \mathbf{q})^{-1}, \\ \mathbf{m} &= - \left(\mathbf{Q}^* - p\mathbf{m}(\mathbf{Q}^0 + (p-1)\mathbf{q})^{-1}\mathbf{m} \right)^{-1} \mathbf{m} (\mathbf{Q}^0 + (p-1)\mathbf{q})^{-1}. \end{aligned} \tag{B.32}$$

The determinant of \mathbf{Q} is given by

$$\ln \det \mathbf{Q} = (p-1) \ln \det(\mathbf{Q}^0 - \mathbf{q}) + \ln \det(\mathbf{Q}^0 + (p-1)\mathbf{q}) + \ln \det \left(\mathbf{Q}^* - p\mathbf{m}(\mathbf{Q}^0 + (p-1)\mathbf{q})^{-1}\mathbf{m} \right). \tag{B.33}$$

The above results allow us to rewrite

$$\begin{aligned} J(\mathbf{Q}^*, \mathbf{Q}^0, \mathbf{q}, \mathbf{m}) &= \int_{\mathbb{R}^k} d\mathbf{y} \int_{\mathbb{R}^k} \mathcal{D}\xi \exp \left(-\frac{1}{2} \ln \det(2\pi\mathbf{Q}) \right) \int_{\mathbb{R}^k} \mathcal{D}\mathbf{z}^* P_{\text{out}}^*(\mathbf{y}|\mathbf{z}^*) \exp \left(-\frac{1}{2} \mathbf{z}^{*\top} \tilde{\mathbf{Q}}^* \mathbf{z}^* \right) \\ &\quad \times \left[\int_{\mathbb{R}^k} \mathcal{D}\mathbf{z} P_{\text{out}}(\mathbf{y}|\mathbf{z}) \exp \left(-\frac{1}{2} \mathbf{z}^\top (\tilde{\mathbf{Q}}^0 - \tilde{\mathbf{q}}) \mathbf{z} - \mathbf{z}^{*\top} \tilde{\mathbf{m}} \mathbf{z} - \xi^\top \tilde{\mathbf{q}}^{1/2} \mathbf{z} \right) \right]^p. \end{aligned} \tag{B.34}$$

As in the Bayes-optimal case, in order to consider the $p \rightarrow 0^+$ limit, we can rewrite

$$\begin{aligned} \ln I(\hat{\mathbf{Q}}^0, \hat{\mathbf{q}}, \hat{\mathbf{Q}}^*, \hat{\mathbf{m}}) &= p \int_{\mathbb{R}^k} \mathcal{D}\xi \int_{\mathbb{R}^k} d\mathbf{w}^* P_w^*(\mathbf{W}^*) \exp \left(-\frac{1}{2} \mathbf{w}^{*\top} \hat{\mathbf{Q}}^* \mathbf{w}^* \right) \\ &\quad \times \ln \int_{\mathbb{R}^k} d\mathbf{w} P_w(\mathbf{W}) \exp \left(-\frac{1}{2} \mathbf{w}^\top (\hat{\mathbf{Q}}^0 + \hat{\mathbf{q}}) \mathbf{w} + \mathbf{w}^{*\top} \hat{\mathbf{m}} \mathbf{w} + \xi^\top \hat{\mathbf{q}}^{1/2} \mathbf{w} \right) + o(p), \end{aligned} \tag{B.35}$$

$$\begin{aligned} \ln J(\mathbf{Q}^*, \mathbf{Q}^0, \mathbf{q}, \mathbf{m}) &= p \int_{\mathbb{R}^k} d\mathbf{y} \int_{\mathbb{R}^k} \mathcal{D}\xi \exp \left(-\frac{1}{2} \ln \det(2\pi\mathbf{Q}) \right) \int_{\mathbb{R}^k} \mathcal{D}\mathbf{z}^* P_{\text{out}}^*(\mathbf{y}|\mathbf{z}^*) \exp \left(-\frac{1}{2} \mathbf{z}^{*\top} \tilde{\mathbf{Q}}^* \mathbf{z}^* \right) \\ &\quad \times \ln \int_{\mathbb{R}^k} \mathcal{D}\mathbf{z} P_{\text{out}}(\mathbf{y}|\mathbf{z}) \exp \left(-\frac{1}{2} \mathbf{z}^\top (\tilde{\mathbf{Q}}^0 - \tilde{\mathbf{q}}) \mathbf{z} - \mathbf{z}^{*\top} \tilde{\mathbf{m}} \mathbf{z} - \xi^\top \tilde{\mathbf{q}}^{1/2} \mathbf{z} \right) + o(p). \end{aligned} \tag{B.36}$$

B.2. Computing the free entropy

At this point, it is straightforward to compute the free entropy from equation (B.17) by taking the limit $p \rightarrow 0^+$ of equations (B.20)–(B.23)–(B.24) and (B.29)–(B.35)–(B.36). In the Bayes-optimal case, we obtain:

$$\Phi_{\text{BO}}(\alpha) = \text{extr}_{\mathbf{q}, \hat{\mathbf{q}}} \left\{ -\frac{1}{2} \text{Tr}[\mathbf{q}\hat{\mathbf{q}}] + \Psi_w^*(\hat{\mathbf{q}}) + \alpha \Psi_{\text{out}}^*(\mathbf{q}) \right\}, \tag{B.37}$$

$$\begin{aligned} \Psi_w^*(\hat{\mathbf{q}}) &= \mathbb{E}_\xi \left[Z_w^* \left(\hat{\mathbf{q}}^{1/2} \xi, \hat{\mathbf{q}} \right) \ln Z_w^* \left(\hat{\mathbf{q}}^{1/2} \xi, \hat{\mathbf{q}} \right) \right], \\ \Psi_{\text{out}}^*(\mathbf{q}) &= \mathbb{E}_{\mathbf{y}, \xi} \left[Z_{\text{out}}^* \left(\mathbf{y}, \mathbf{q}^{1/2} \xi, \mathbf{Q}^* - \mathbf{q} \right) \ln Z_{\text{out}}^* \left(\mathbf{y}, \mathbf{q}^{1/2} \xi, \mathbf{Q}^* - \mathbf{q} \right) \right]. \end{aligned} \tag{B.38}$$

In the non Bayes-optimal case, we obtain:

$$\Phi_{\text{non-BO}}(\alpha) = \text{extr}_{\mathbf{Q}^0, \mathbf{q}, \mathbf{m}, \hat{\mathbf{Q}}^0, \hat{\mathbf{q}}, \hat{\mathbf{m}}} \left\{ -\text{Tr}[\mathbf{m}\hat{\mathbf{m}}] + \frac{1}{2} \text{Tr} \left[\mathbf{Q}^0 \hat{\mathbf{Q}}^0 \right] + \frac{1}{2} \text{Tr}[\mathbf{q}\hat{\mathbf{q}}] \right. \tag{B.39}$$

$$\left. + \Psi_w(\hat{\mathbf{Q}}^0, \hat{\mathbf{q}}, \hat{\mathbf{m}}) + \alpha \Psi_{\text{out}}(\mathbf{Q}^*, \mathbf{Q}^0, \mathbf{q}, \mathbf{m}) \right\}, \tag{B.40}$$

$$\begin{aligned}\Psi_w(\hat{Q}^0, \hat{q}, \hat{m}) &= \mathbb{E}_\xi \left[Z_w^* \left(\hat{m} \hat{q}^{-1/2} \xi, \hat{m} \hat{q}^{-1} \hat{m} \right) \ln Z_w \left(\hat{q}^{1/2} \xi, \hat{Q}^0 + \hat{q} \right) \right], \\ \Psi_{\text{out}}(\mathbf{Q}^*, \mathbf{Q}^0, \mathbf{q}, \mathbf{m}) &= \mathbb{E}_{\mathbf{y}, \xi} \left[Z_{\text{out}}^* \left(\mathbf{y}; \mathbf{m} \mathbf{q}^{-1/2} \xi, -\mathbf{m} \mathbf{q}^{-1} \mathbf{m} \right) \ln Z_{\text{out}} \left(\mathbf{y}; \mathbf{q}^{1/2} \xi, \mathbf{Q}^0 - \mathbf{q} \right) \right],\end{aligned}\quad (\text{B.41})$$

where we remind that in both cases \mathbf{Q}^* is fixed given the teacher prior. The above equations make use of a series of auxiliary functions $Z_w^*, Z_w, Z_{\text{out}}^*, Z_{\text{out}}$ that simply come from a more compact way of writing equations (B.23)–(B.24) and (B.35)–(B.36), i.e.

$$\mathcal{Z}_w(\gamma, \Lambda) = \int_{\mathbb{R}^k} d\mathbf{w} P_w(\mathbf{w}) e^{-\frac{1}{2} \mathbf{w}^\top \Lambda \mathbf{w} + \gamma^\top \mathbf{w}}, \quad (\text{B.42a})$$

$$\mathcal{Z}_{\text{out}}(\mathbf{y}; \omega, \mathbf{V}) = \int_{\mathbb{R}^k} d\mathbf{z} \frac{e^{-\frac{1}{2} (\mathbf{z} - \omega)^\top \mathbf{V}^{-1} (\mathbf{z} - \omega)}}{\sqrt{\det(2\pi \mathbf{V})}} P_{\text{out}}(\mathbf{y} | \mathbf{z}), \quad (\text{B.42b})$$

and Z_w^*, Z_{out}^* are defined in the exact same way provided that the student distributions P_w, P_{out} are replaced by the teacher distributions P_w^*, P_{out}^* .

Appendix C. Update equations for the overlap parameters

We can now compute the update equations for the overlap parameters both in the Bayes and non Bayes-optimal settings by taking the derivatives of equation (B.37) with respect to (\mathbf{q}, \hat{q}) and of equation (B.39) with respect to $(\mathbf{Q}^0, \mathbf{q}, \mathbf{m}, \hat{Q}^0, \hat{q}, \hat{m})$, and setting them to zero. In the Bayes-optimal setting the update equations are therefore given by:

$$\mathbf{q} = \mathbb{E}_\xi \left[Z_w^* (\hat{q}^{1/2} \xi, \hat{q}) f_w^* (\hat{q}^{1/2} \xi, \hat{q}) f_w^* (\hat{q}^{1/2} \xi, \hat{q})^\top \right], \quad (\text{C.1})$$

$$\hat{q} = \alpha \mathbb{E}_{\mathbf{y}, \xi} \left[Z_{\text{out}}^* (\mathbf{y}; \mathbf{q}^{1/2} \xi, \mathbf{Q}^* - \mathbf{q}) f_{\text{out}}^* (\mathbf{y}; \mathbf{q}^{1/2} \xi, \mathbf{Q}^* - \mathbf{q}) f_{\text{out}}^* (\mathbf{y}; \mathbf{q}^{1/2} \xi, \mathbf{Q}^* - \mathbf{q})^\top \right]. \quad (\text{C.2})$$

In the non-Bayes optimal setting, we define for simplicity: $\mathbf{V} = \mathbf{Q}^0 - \mathbf{q}$, $\hat{\mathbf{V}} = \hat{Q}^0 + \hat{q}$, and we find

$$\mathbf{m} = \mathbb{E}_\xi \left[Z_w^* \times f_w^* (\hat{m} \hat{q}^{-1/2} \xi, \hat{m} \hat{q}^{-1} \hat{m}) f_w (\hat{q}^{1/2} \xi, \hat{\mathbf{V}})^\top \right], \quad (\text{C.3})$$

$$\mathbf{q} = \mathbb{E}_\xi \left[Z_w^* (\hat{m} \hat{q}^{-1/2} \xi, \hat{m} \hat{q}^{-1} \hat{m}) f_w (\hat{q}^{1/2} \xi, \hat{\mathbf{V}}) f_w (\hat{q}^{1/2} \xi, \hat{\mathbf{V}})^\top \right], \quad (\text{C.4})$$

$$\mathbf{V} = \mathbb{E}_\xi \left[Z_w^* (\hat{m} \hat{q}^{-1/2} \xi, \hat{m} \hat{q}^{-1} \hat{m}) \partial_\gamma f_w (\hat{q}^{1/2} \xi, \hat{\mathbf{V}}) \right], \quad (\text{C.5})$$

$$\hat{\mathbf{m}} = \alpha \mathbb{E}_{\mathbf{y}, \xi} \left[Z_{\text{out}}^* f_{\text{out}}^* (\mathbf{y}; \mathbf{m} \mathbf{q}^{-1/2} \xi, \mathbf{Q}^* - \mathbf{m}^\top \mathbf{q}^{-1} \mathbf{m}) f_{\text{out}} (\mathbf{y}; \mathbf{q}^{1/2} \xi, \mathbf{V})^\top \right], \quad (\text{C.6})$$

$$\hat{q} = \alpha \mathbb{E}_{\mathbf{y}, \xi} \left[Z_{\text{out}}^* (\mathbf{y}; \mathbf{m} \mathbf{q}^{-1/2} \xi, \mathbf{Q}^* - \mathbf{m}^\top \mathbf{q}^{-1} \mathbf{m}) f_{\text{out}} (\mathbf{y}; \mathbf{q}^{1/2} \xi, \mathbf{V}) f_{\text{out}} (\mathbf{y}; \mathbf{q}^{1/2} \xi, \mathbf{V})^\top \right], \quad (\text{C.7})$$

$$\hat{\mathbf{V}} = -\alpha \mathbb{E}_{\mathbf{y}, \xi} \left[Z_{\text{out}}^* (\mathbf{y}; \mathbf{m} \mathbf{q}^{-1/2} \xi, \mathbf{Q}^* - \mathbf{m}^\top \mathbf{q}^{-1} \mathbf{m}) \partial_w f_{\text{out}} (\mathbf{y}; \mathbf{q}^{1/2} \xi, \mathbf{V}) \right], \quad (\text{C.8})$$

where in both settings we have made use of the following definitions.

C.1. Definitions of the update functions

For $\mathbf{w} \in \mathbb{R}^k$, let

$$Q_w(\mathbf{w}; \gamma, \mathbf{\Lambda}) \equiv \frac{P_w(\mathbf{w})}{\mathcal{Z}_w(\gamma, \mathbf{\Lambda})} e^{-\frac{1}{2} \mathbf{w}^\top \mathbf{\Lambda} \mathbf{w} + \gamma^\top \mathbf{w}}, \tag{C.9a}$$

with

$$\mathbf{f}_w(\gamma, \mathbf{\Lambda}) \equiv \partial_\gamma \log \mathcal{Z}_w(\gamma, \mathbf{\Lambda}) = \mathbb{E}_{Q_w}[\mathbf{w}], \tag{C.9b}$$

and for $\mathbf{z} \in \mathbb{R}^k$, let

$$Q_{\text{out}}(\mathbf{z}; \mathbf{y}, \boldsymbol{\omega}, \mathbf{V}) \equiv \frac{P_{\text{out}}(\mathbf{y}|\mathbf{w})}{\mathcal{Z}_{\text{out}}(\mathbf{y}, \boldsymbol{\omega}, \mathbf{V})} \frac{e^{-\frac{1}{2}(\mathbf{z}-\boldsymbol{\omega})^\top \mathbf{V}^{-1}(\mathbf{z}-\boldsymbol{\omega})}}{\sqrt{\det(2\pi\mathbf{V})}}, \tag{C.9c}$$

with

$$\mathbf{f}_{\text{out}}(\mathbf{y}, \boldsymbol{\omega}, \mathbf{V}) \equiv \partial_{\boldsymbol{\omega}} \log \mathcal{Z}_{\text{out}}(\mathbf{y}, \boldsymbol{\omega}, \mathbf{V}) = \mathbf{V}^{-1} \mathbb{E}_{Q_{\text{out}}}[\mathbf{z} - \boldsymbol{\omega}], \tag{C.9d}$$

where the definitions of $\mathbf{f}_w^*, \mathbf{f}_{\text{out}}^*$ are identical, provided that P_w, P_{out} are replaced by P_w^*, P_{out}^* . The functions \mathcal{Z}_w and \mathcal{Z}_{out} are given by equations (B.42).

The explicit expressions of the auxiliary functions depend on the choice of the teacher and student distributions. We evaluate these expressions for the special cases under consideration in the following sections.

C.2. Bayes-optimal update functions

In this section, we evaluate the Bayes-optimal update functions. We consider directly the expressions obtained after performing the mapping described in appendix A.

C.2.1. Gaussian prior terms with the dimensional reduction A

In the case of Gaussian teacher prior, it is straightforward to notice that, after the application of the mapping A, the prior over the weights is still Gaussian with covariance $\tilde{\boldsymbol{\Sigma}} = \mathbf{I}_{k-1} + \mathbf{1}_{k-1} \mathbf{1}_{k-1}^\top$:

$$\begin{aligned} \mathcal{Z}_w^*(\gamma, \mathbf{\Lambda}) &= \int_{\mathbb{R}^{k-1}} \frac{d\mathbf{w}}{\sqrt{(2\pi)^{k-1} \det(\tilde{\boldsymbol{\Sigma}})}} \exp \left[-\frac{1}{2} \mathbf{w}^\top (\tilde{\boldsymbol{\Sigma}}^{-1} + \mathbf{\Lambda}) \mathbf{w} + \gamma^\top \mathbf{w} \right] \\ &= \frac{1}{\sqrt{\det(\tilde{\boldsymbol{\Sigma}}) \det(\tilde{\boldsymbol{\Sigma}}^{-1} + \mathbf{\Lambda})}} \exp \left[\frac{1}{2} \gamma^\top (\tilde{\boldsymbol{\Sigma}}^{-1} + \mathbf{\Lambda})^{-1} \gamma \right], \end{aligned} \tag{C.10a}$$

leading to

$$\mathbf{f}_w^*(\gamma, \mathbf{\Lambda}) = \partial_\gamma \log \mathcal{Z}_w^*(\gamma, \mathbf{\Lambda}) = (\tilde{\boldsymbol{\Sigma}}^{-1} + \mathbf{\Lambda})^{-1} \gamma, \tag{C.10b}$$

$$\partial_\gamma \mathbf{f}_w^*(\gamma, \mathbf{\Lambda}) = (\tilde{\boldsymbol{\Sigma}}^{-1} + \mathbf{\Lambda})^{-1}. \tag{C.10c}$$

For $k = 3$, the reduced covariance matrix is given by

$$\tilde{\boldsymbol{\Sigma}} = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}. \tag{C.11}$$

C.2.2. Rademacher prior terms with the dimensional reduction A

Considering $k = 3$, the reduced prior given by equation (A.11) becomes

$$\begin{aligned} P_{\tilde{\mathbf{w}}}(\tilde{w}_1^*, \tilde{w}_2^*) &= \frac{1}{2^3} [2\delta(\tilde{w}_1^*)\delta(\tilde{w}_2^*) + \delta(\tilde{w}_1^*)\delta(\tilde{w}_2^* + 2) + \delta(\tilde{w}_1^* + 2)\delta(\tilde{w}_2^*) + \delta(\tilde{w}_1^*)\delta(\tilde{w}_2^* - 2) \\ &\quad + \delta(\tilde{w}_1^* - 2)\delta(\tilde{w}_2^*) + \delta(\tilde{w}_1^* + 2)\delta(\tilde{w}_2^* + 2) + \delta(\tilde{w}_1^* - 2)\delta(\tilde{w}_2^* - 2)]. \end{aligned} \tag{C.12}$$

The denoising functions for this case are computed numerically, via Monte Carlo sampling of the distribution given equation (C.12).

C.2.3. Output terms with the dimensional reduction **A**

Considering directly the mapping to dimension $k - 1$, we can write the Bayes-optimal model distribution as

$$P_{\text{out}}^*(\mathbf{y}|\mathbf{z}) = \sum_{l=1}^{k-1} \delta_{\mathbf{y}, \mathbf{e}_l} \Theta(z_l) \prod_{h \neq l, h=1}^{k-1} \Theta(z_l - z_h) + \delta_{\mathbf{y}, \mathbf{e}_k} \prod_{l=1}^{k-1} \Theta(-z_l). \tag{C.13}$$

Therefore, the auxiliary functions Z_{out}^* , and so on, are composed by $k - 1$ contributions according to the membership of the label \mathbf{y} in the argument. For instance, in the case $k = 3$, we have

$$\begin{aligned} Z_{\text{out}}^*(\mathbf{y}; \boldsymbol{\omega}, \mathbf{V}) &= \delta_{\mathbf{y}, \mathbf{e}_1} \int_0^{+\infty} dz_1 \int_{-\infty}^{z_1} dz_2 \frac{e^{-\frac{1}{2}(\mathbf{z}-\boldsymbol{\omega})^\top \mathbf{V}^{-1}(\mathbf{z}-\boldsymbol{\omega})}}{\sqrt{\det(2\pi \mathbf{V})}} + \delta_{\mathbf{y}, \mathbf{e}_2} \int_0^{+\infty} dz_2 \int_{-\infty}^{z_2} dz_1 \frac{e^{-\frac{1}{2}(\mathbf{z}-\boldsymbol{\omega})^\top \mathbf{V}^{-1}(\mathbf{z}-\boldsymbol{\omega})}}{\sqrt{\det(2\pi \mathbf{V})}} \\ &+ \delta_{\mathbf{y}, \mathbf{e}_3} \int_{-\infty}^0 dz_1 \int_{-\infty}^0 dz_2 \frac{e^{-\frac{1}{2}(\mathbf{z}-\boldsymbol{\omega})^\top \mathbf{V}^{-1}(\mathbf{z}-\boldsymbol{\omega})}}{\sqrt{\det(2\pi \mathbf{V})}}. \end{aligned} \tag{C.14}$$

Similarly, for f_{out}^* we need to change the integration bounds in order to take into account all the possibilities for the label. For each term, we can only compute analytically the inner integral, while we have to estimate the outer ones via Monte Carlo sampling. Therefore, applying the mapping in **A** is useful in order to reduce the number of integrals to be performed numerically and speed up the whole procedure.

C.3. ERM update functions

The update equations for ERM can be derived as a special case of the non Bayes-optimal equations (C.1) and so on. In particular, this can be seen by rewriting the solution of the optimization problem as the *ground state* of the following measure

$$\begin{aligned} P_\beta(\mathbf{W}|\mathbf{X}, \mathbf{Y}) &= \frac{1}{Z_d(\beta)} \exp(-\beta r_\lambda(\mathbf{W})) \exp(-\beta \mathcal{L}(\mathbf{W}; \mathbf{X}, \mathbf{Y})) \\ &= \frac{1}{Z_d(\beta)} \prod_{l=1}^k \exp\left(-\frac{\beta \lambda}{2} \|\mathbf{w}_l\|_2^2\right) \prod_{\mu=1}^n \exp\left(-\beta \ell(\mathbf{W}^\top \mathbf{x}_\mu, \mathbf{y}_\mu)\right), \end{aligned} \tag{C.15}$$

i.e. the solution in the limit $\beta \rightarrow \infty$. Therefore, we can express the prior and model distributions of a student learning via ERM as

$$P_w(\mathbf{W}) \propto \exp\left(-\frac{\beta \lambda}{2} \|\mathbf{w}\|_2^2\right), \quad P_{\text{out}}(\mathbf{y}|\mathbf{W}^\top \mathbf{x}) \propto \exp\left(-\beta \ell(\mathbf{W}^\top \mathbf{x}, \mathbf{y})\right). \tag{C.16}$$

C.3.1. Prior terms with the dimensional reduction **A**

The ERM ridge-regularization prior can therefore be seen as i.i.d. Gaussian-distributed with variance $1/\beta\lambda$. This means that, applying the mapping **A**, we have

$$P_w(\mathbf{W}) = \frac{1}{(2\pi)^{(k-1)/2} \sqrt{\det(\mathbf{C}/(\beta\lambda))}} \exp\left(-\frac{\beta \lambda}{2} \mathbf{w}^\top \mathbf{C}^{-1} \mathbf{w}\right), \tag{C.17}$$

where again \mathbf{C} is the prior covariance in the reduced setting, i.e. $\mathbf{C} = [2, 1; 1, 2]$ for $k = 3$. Let we rescale $\boldsymbol{\gamma} \leftarrow \beta \boldsymbol{\gamma}$ and $\boldsymbol{\Lambda} \leftarrow \beta \boldsymbol{\Lambda}$. We will see that this would correspond to the rescaling: $\hat{\boldsymbol{\gamma}} \leftarrow \beta^2 \hat{\boldsymbol{\gamma}}$ and $\hat{\mathbf{V}} \leftarrow \beta \hat{\mathbf{V}}$. We obtain

$$\begin{aligned} Z_w(\boldsymbol{\gamma}, \boldsymbol{\Lambda}) &= \int_{\mathbb{R}^{k-1}} d\mathbf{w} \frac{e^{-\frac{\beta}{2} \mathbf{w}^\top (\lambda \mathbf{C}^{-1} + \boldsymbol{\Lambda}) \mathbf{w} + \beta \boldsymbol{\gamma}^\top \mathbf{w}}}{(2\pi)^{(k-1)/2} \sqrt{\det(\mathbf{C}/\beta\lambda)}} \\ &= \frac{1}{\sqrt{\det(\mathbf{C}/\beta\lambda) \det(\beta \lambda \mathbf{C}^{-1} + \beta \boldsymbol{\Lambda})}} \exp\left(\frac{\beta}{2} \boldsymbol{\gamma}^\top (\lambda \mathbf{C}^{-1} + \boldsymbol{\Lambda})^{-1} \boldsymbol{\gamma}\right), \end{aligned} \tag{C.18}$$

$$\mathbf{f}_w(\boldsymbol{\gamma}, \boldsymbol{\Lambda}) = \beta (\lambda \mathbf{C}^{-1} + \boldsymbol{\Lambda})^{-1} \boldsymbol{\gamma}, \tag{C.19}$$

$$\partial_{\boldsymbol{\gamma}} \mathbf{f}_w(\boldsymbol{\gamma}, \boldsymbol{\Lambda}) = \beta (\lambda \mathbf{C}^{-1} + \boldsymbol{\Lambda})^{-1}. \tag{C.20}$$

Substituting the expressions above in equations (10), we find

$$\mathbf{m} = \frac{1}{\sqrt{\det(C) \det(C^{-1} + \hat{\mathbf{m}}\hat{\mathbf{q}}^{-1}\hat{\mathbf{m}})} \sqrt{\det(I - \hat{\mathbf{q}}^{-1/2}\hat{\mathbf{m}}(C^{-1} + \hat{\mathbf{m}}\hat{\mathbf{q}}^{-1}\hat{\mathbf{m}})^{-1}\hat{\mathbf{m}}\hat{\mathbf{q}}^{-1/2)}}} \tag{C.21}$$

$$\begin{aligned} &\times (C^{-1} + \hat{\mathbf{m}}\hat{\mathbf{q}}^{-1}\hat{\mathbf{m}})^{-1}\hat{\mathbf{m}}\hat{\mathbf{q}}^{-1/2} \left(I - \hat{\mathbf{q}}^{-1/2}\hat{\mathbf{m}}(C^{-1} + \hat{\mathbf{m}}\hat{\mathbf{q}}^{-1}\hat{\mathbf{m}})^{-1}\hat{\mathbf{m}}\hat{\mathbf{q}}^{-1/2} \right)^{-1} \\ &\times \hat{\mathbf{q}}^{1/2}(\lambda C^{-1} + \hat{\mathbf{V}})^{-1}, \end{aligned} \tag{C.22}$$

$$\mathbf{q} = \frac{1}{\sqrt{\det(C) \det(C^{-1} + \hat{\mathbf{m}}\hat{\mathbf{q}}^{-1}\hat{\mathbf{m}})} \sqrt{\det(I - \hat{\mathbf{q}}^{-1/2}\hat{\mathbf{m}}(C^{-1} + \hat{\mathbf{m}}\hat{\mathbf{q}}^{-1}\hat{\mathbf{m}})^{-1}\hat{\mathbf{m}}\hat{\mathbf{q}}^{-1/2)}}} \tag{C.23}$$

$$\begin{aligned} &\times (\lambda C^{-1} + \hat{\mathbf{V}})^{-1}\hat{\mathbf{q}}^{1/2} \left(I - \hat{\mathbf{q}}^{-1/2}\hat{\mathbf{m}}(C^{-1} + \hat{\mathbf{m}}\hat{\mathbf{q}}^{-1}\hat{\mathbf{m}})^{-1}\hat{\mathbf{m}}\hat{\mathbf{q}}^{-1/2} \right)^{-1} \\ &\times \hat{\mathbf{q}}^{1/2}(\lambda C^{-1} + \hat{\mathbf{V}})^{-1}, \end{aligned} \tag{C.24}$$

$$\mathbf{V} = \frac{(\lambda C^{-1} + \hat{\mathbf{V}})^{-1}}{\sqrt{\det(C) \det(C^{-1} + \hat{\mathbf{m}}\hat{\mathbf{q}}^{-1}\hat{\mathbf{m}})} \sqrt{\det(I - \hat{\mathbf{q}}^{-1/2}\hat{\mathbf{m}}(C^{-1} + \hat{\mathbf{m}}\hat{\mathbf{q}}^{-1}\hat{\mathbf{m}})^{-1}\hat{\mathbf{m}}\hat{\mathbf{q}}^{-1/2)}}}, \tag{C.25}$$

where additionally we have rescaled: $\mathbf{m} \leftarrow \beta\mathbf{m}$, $\mathbf{q} \leftarrow \beta^2\mathbf{q}$, $\mathbf{V} \leftarrow \beta\mathbf{V}$. The equations are now independent of the parameter β . We will see that the above rescaling is consistent and leads to a set of well-defined equations in the $\beta \rightarrow \infty$ limit.

C.3.2. Output terms with the dimensional reduction A

We remind that we have performed the rescaling $\mathbf{V} \rightarrow \beta^{-1}\mathbf{V}$. In the $\beta \rightarrow \infty$ limit, the ERM output term Z_{out} becomes

$$Z_{\text{out}}(\mathbf{y}; \boldsymbol{\omega}, \mathbf{V}) \propto \sqrt{\beta^{k-1}} \int_{\mathbb{R}^{k-1}} d\mathbf{z} \frac{e^{-\beta[-\frac{1}{2}(\mathbf{z}-\boldsymbol{\omega})^\top \mathbf{V}^{-1}(\mathbf{z}-\boldsymbol{\omega}) + \mathcal{L}(\mathbf{y}, \mathbf{z})]}}{\sqrt{\det(2\pi\mathbf{V})}} \xrightarrow{\beta \rightarrow \infty} \sqrt{\frac{\beta^{k-1}}{\det(2\pi\mathbf{V})}} e^{-\beta\mathcal{M}_{\mathbf{V}\mathcal{L}(\mathbf{y}, \cdot)}(\boldsymbol{\omega})}, \tag{C.26}$$

where \mathcal{M} is a Moreau envelope associated with the loss \mathcal{L} ,

$$\mathcal{M}_{\mathbf{V}\mathcal{L}(\mathbf{y}, \cdot)}(\boldsymbol{\omega}) = \inf_{\mathbf{z} \in \mathbb{R}^{k-1}} \left[\frac{1}{2}(\mathbf{z} - \boldsymbol{\omega})^\top \mathbf{V}^{-1}(\mathbf{z} - \boldsymbol{\omega}) + \mathcal{L}(\mathbf{y}, \mathbf{z}) \right]. \tag{C.27}$$

From equation (C.9d), we have

$$\mathbf{f}_{\text{out}}(\mathbf{y}, \boldsymbol{\omega}, \mathbf{V}) = -\beta \partial_{\boldsymbol{\omega}} \mathcal{M}_{\mathbf{V}\mathcal{L}(\mathbf{y}, \cdot)}(\boldsymbol{\omega}), \tag{C.28}$$

which can be obtained using the proximal operator

$$\text{prox}_{\mathbf{V}\mathcal{L}(\mathbf{y}, \cdot)}(\boldsymbol{\omega}) = \arg \min_{\mathbf{z} \in \mathbb{R}^{k-1}} \left[\frac{1}{2}(\mathbf{z} - \boldsymbol{\omega})^\top \mathbf{V}^{-1}(\mathbf{z} - \boldsymbol{\omega}) + \mathcal{L}(\mathbf{y}, \mathbf{z}) \right]. \tag{C.29}$$

The envelope theorem, $\mathcal{M}'_{\Sigma_f}(\mathbf{x}) = \Sigma^{-1}(\mathbf{x} - \text{prox}_{\Sigma_f}(\mathbf{x}))$ leads to

$$\mathbf{f}_{\text{out}}(\mathbf{y}, \boldsymbol{\omega}, \mathbf{V}) = -\beta \mathbf{V}^{-1} \left(\boldsymbol{\omega} - \text{prox}_{\mathbf{V}\mathcal{L}(\mathbf{y}, \cdot)}(\boldsymbol{\omega}) \right), \tag{C.30}$$

$$\partial_{\boldsymbol{\omega}} \mathbf{f}_{\text{out}}(\mathbf{y}, \boldsymbol{\omega}, \mathbf{V}) = -\beta \mathbf{V}^{-1} \left(\mathbf{I} - \partial_{\boldsymbol{\omega}} \text{prox}_{\mathbf{V}\mathcal{L}(\mathbf{y}, \cdot)}(\boldsymbol{\omega}) \right), \tag{C.31}$$

$$\partial_{\boldsymbol{\omega}} \text{prox}_{\mathbf{V}\mathcal{L}(\mathbf{y}, \cdot)}(\boldsymbol{\omega}) = \partial_{\boldsymbol{\omega}} \mathbf{z}^*(\boldsymbol{\omega}) = (\mathbf{V}^{-1} + \partial_{\mathbf{z}}^2 \mathcal{L})^{-1}, \tag{C.32}$$

consistently with the rescaling previously adopted, which leads to the final equations equations (10) in the main text, holding in the limit $\beta \rightarrow \infty$.

C.3.3. Special case: square loss

The proximal operator for the square loss can be computed analytically:

$$\text{prox}_{\mathcal{V}\mathcal{L}(y,\cdot)}^{\text{SL}}(\omega) = (\mathbf{I} + \mathbf{V})^{-1}(\mathbf{W} + \mathbf{V}\mathbf{y}), \quad (\text{C.33})$$

$$\partial_{\mathbf{w}}\text{prox}_{\mathcal{V}\mathcal{L}(y,\cdot)}^{\text{SL}}(\omega) = (\mathbf{I} + \mathbf{V})^{-1}. \quad (\text{C.34})$$

Appendix D. Proof of the main theorem

In this section we prove the main theorem in a slightly more general setup than what is presented in the main part of the paper. 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.

D.1. The learning problem

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

$$\mathbf{Y} = \phi_{\text{out}}\left(\frac{1}{\sqrt{d}}\mathbf{X}\mathbf{W}^*\right) \quad (\text{D.1})$$

where $\mathbf{Y} \in \mathbb{R}^{n \times k}$, $\mathbf{X} \sim \mathcal{N}(0, 1) \in \mathbb{R}^{n \times d}$ and $\mathbf{W}^* \in \mathbb{R}^{d \times k}$. The goal is to try to learn an estimator of \mathbf{W}^* using a generalised linear model defined by the optimisation problem

$$\hat{\mathbf{W}} \in \arg \min_{\mathbf{W} \in \mathbb{R}^{d \times k}} \mathcal{L}\left(\mathbf{Y}, \frac{1}{\sqrt{d}}\mathbf{X}\mathbf{W}\right) + r(\mathbf{W}) \quad (\text{D.2})$$

where \mathcal{L}, r are convex functions, and we omit the dependence of the regularisation r on the parameter λ for simplicity. We wish to determine the asymptotic properties of the estimator $\hat{\mathbf{W}}$ in the limit where $n, d \rightarrow \infty$ with fixed ratios $\alpha = n/d$. We now list the necessary assumptions for our main theorem to hold.

D.1.1. Assumptions

- the functions \mathcal{L}, r are proper, closed, lower-semicontinuous, convex functions. The loss function \mathcal{L} is differentiable and pseudo-Lipschitz of order 2 in both its arguments. We assume additionally that the regularisation r is strongly convex, differentiable and pseudo-Lipschitz of order 2.
- the dimensions n, d grow linearly with finite ratios $\alpha = n/d$, and the number of classes k is kept constant.
- the lines of the ground truth matrix $\mathbf{W}^* \in \mathbb{R}^{d \times k}$ are sampled i.i.d. from a sub-Gaussian probability distribution in \mathbb{R}^k .

D.2. Reduction to an AMP iteration

We start by reformulating the optimisation problem (D.2) in order to be able to solve it with an AMP iteration. In particular it is useful to separate the design matrix \mathbf{X} in two contributions: one aligned with the ground truth \mathbf{W}^* and one independent on the teacher \mathbf{Y} . To do so we condition \mathbf{X} on the teacher input $\mathbf{X}\mathbf{W}^*$ such that

$$\mathbf{X} = \mathbb{E}[\mathbf{X}|\mathbf{Y}] + \mathbf{X} - \mathbb{E}[\mathbf{X}|\mathbf{Y}] \quad (\text{D.3})$$

$$= \mathbb{E}[\mathbf{X}|\mathbf{X}\mathbf{W}^*] + \mathbf{X} - \mathbb{E}[\mathbf{X}|\mathbf{X}\mathbf{W}^*] \quad (\text{D.4})$$

$$= \mathbf{X}\mathbf{P}_{\mathbf{W}^*} + \tilde{\mathbf{X}}\mathbf{P}_{\mathbf{W}^*}^{\perp} \quad (\text{D.5})$$

where $\tilde{\mathbf{X}}$ is an independent copy of the design matrix \mathbf{X} , $\mathbf{P}_{\mathbf{W}^*}$ denotes the orthogonal projection on the subspace spanned by the columns of \mathbf{W}^* and $\mathbf{P}_{\mathbf{W}^*}^{\perp} = \mathbf{I}_d - \mathbf{P}_{\mathbf{W}^*}$. Furthermore, since we assume that n, d are arbitrarily large and that k remains finite for each instance of the problem, the matrix \mathbf{W}^* has full column rank and the projector $\mathbf{P}_{\mathbf{W}^*} = \mathbf{W}^* ((\mathbf{W}^*)^{\top} \mathbf{W}^*)^{-1} (\mathbf{W}^*)^{\top}$ is always well-defined with high probability. We can then rewrite the original problem as

$$\hat{\mathbf{W}} \in \arg \min_{\mathbf{W} \in \mathbb{R}^{d \times k}} \mathcal{L}\left(\mathbf{Y}, \frac{1}{\sqrt{d}}(\mathbf{X}\mathbf{P}_{\mathbf{W}^*} + \tilde{\mathbf{X}}\mathbf{P}_{\mathbf{W}^*}^{\perp})\mathbf{W}\right) + r(\mathbf{W}). \quad (\text{D.6})$$

The quantity \mathbf{XW}^* is a $\mathbb{R}^{n \times k}$ Gaussian matrix with covariance $(\mathbf{W}^*)^\top \mathbf{W}^*$, and can be represented as $\mathbf{XW}^* = \mathbf{S}((\mathbf{W}^*)^\top \mathbf{W}^*)^{1/2}$ where \mathbf{S} is an $n \times k$ random matrix with i.i.d. standard normal elements. We then have

$$\frac{1}{\sqrt{d}} \mathbf{X} \mathbf{P}_{\mathbf{W}^*} = \frac{1}{\sqrt{d}} \mathbf{W}^* ((\mathbf{W}^*)^\top \mathbf{W}^*)^{-1} (\mathbf{W}^*)^\top \mathbf{W} \quad (\text{D.7})$$

$$= \frac{1}{\sqrt{d}} \mathbf{S} \sqrt{d} \rho^{1/2} \frac{1}{d} \rho^{-1} d \mathbf{m}^\top \quad (\text{D.8})$$

$$= \mathbf{S} \rho^{-1/2} \mathbf{m}^\top \quad (\text{D.9})$$

where we introduced the order parameter $\mathbf{m} = \frac{1}{d} \hat{\mathbf{W}}^\top \mathbf{W}^* \in \mathbb{R}^{k \times k}$ and the quantity $\rho = \frac{1}{d} (\mathbf{W}^*)^\top \mathbf{W}^* \in \mathbb{R}^{k \times k}$. Note that $\mathbf{Y} = \mathbf{S} \rho^{1/2}$, and

$$\hat{\mathbf{W}} \in \arg \min_{\mathbf{W} \in \mathbb{R}^{d \times k}} \mathcal{L} \left(\mathbf{Y}, \mathbf{S} \rho^{-1/2} \mathbf{m}^\top + \frac{1}{\sqrt{d}} \tilde{\mathbf{X}} \mathbf{P}_{\mathbf{W}^*}^\perp \mathbf{W} \right) + r(\mathbf{W}). \quad (\text{D.10})$$

We may then rewrite the optimisation problem equation (D.26) as an equivalent problem under constraint on the definition of \mathbf{m} leading to the Lagrangian formulation

$$\inf_{\mathbf{m}, \mathbf{W}} \sup_{\hat{\mathbf{m}}} \mathcal{L} \left(\mathbf{Y}, \mathbf{S} \rho^{-1/2} \mathbf{m}^\top + \frac{1}{\sqrt{d}} \tilde{\mathbf{X}} \mathbf{P}_{\mathbf{W}^*}^\perp \mathbf{W} \right) + r(\mathbf{W}^* \rho^{-1} \mathbf{m}^\top + \mathbf{P}_{\mathbf{W}^*}^\perp \mathbf{W}) + \text{Tr} \left(\hat{\mathbf{m}}^\top (d \mathbf{m} - \hat{\mathbf{W}}^\top \mathbf{W}^*) \right). \quad (\text{D.11})$$

Letting $\mathbf{U} = \mathbf{P}_{\mathbf{W}^*}^\perp \mathbf{W}$ such that $\mathbf{W} = \mathbf{W}^* \rho^{-1} \mathbf{m}^\top + \mathbf{U}$, the problem becomes

$$\inf_{\mathbf{m}, \mathbf{U}} \sup_{\hat{\mathbf{m}}} \mathcal{L} \left(\mathbf{Y}, \mathbf{S} \rho^{-1/2} \mathbf{m}^\top + \frac{1}{\sqrt{d}} \tilde{\mathbf{X}} \mathbf{U} \right) + r(\mathbf{W}^* \rho^{-1} \mathbf{m}^\top + \mathbf{U}) - \text{Tr} \left(\hat{\mathbf{m}}^\top \mathbf{U}^\top \mathbf{W}^* \right) \quad (\text{D.12})$$

where the initial constraint on \mathbf{m} automatically enforces the orthogonality constraint on \mathbf{U} w.r.t. \mathbf{W}^* . The following lemma then characterises the feasibility sets of $\mathbf{m}, \hat{\mathbf{m}}, \mathbf{U}$.

Lemma 1. Consider the optimisation problem equation (D.12). Then there exist constants $C_U, C_m, C_{\hat{m}}$ such that

$$\frac{1}{\sqrt{d}} \|\mathbf{U}\|_F \leq C_U, \quad \|\mathbf{m}\|_F \leq C_m, \quad \|\hat{\mathbf{m}}\|_F \leq C_{\hat{m}} \quad (\text{D.13})$$

with high probability as $n, d \rightarrow \infty$.

Proof. Consider the optimisation problem defining $\hat{\mathbf{W}}$

$$\hat{\mathbf{W}} \in \arg \min_{\mathbf{W} \in \mathbb{R}^{d \times k}} \mathcal{L}(\mathbf{Y}, \mathbf{XW}) + r(\mathbf{W}). \quad (\text{D.14})$$

From the strong convexity assumption on r , there exists a strictly positive constant λ_2 such that the function $\tilde{r}(\mathbf{W}) := r(\mathbf{W}) - \frac{\lambda_2}{2} \|\mathbf{W}\|_F^2$ is convex (and proper, closed, lower semi-continuous). We can then rewrite the optimisation problem as

$$\hat{\mathbf{W}} \in \arg \min_{\mathbf{W} \in \mathbb{R}^{d \times k}} \mathcal{L}(\mathbf{Y}, \mathbf{XW}) + \tilde{r}(\mathbf{W}) + \frac{\lambda_2}{2} \|\mathbf{W}\|_F^2 \quad (\text{D.15})$$

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

$$\frac{1}{d} \left(\mathcal{L}(\mathbf{Y}, \mathbf{X}\hat{\mathbf{W}}) + \tilde{r}(\hat{\mathbf{W}}) + \frac{\lambda_2}{2} \|\hat{\mathbf{W}}\|_F^2 \right) \leq \frac{1}{d} (\mathcal{L}(\mathbf{Y}) + \tilde{r}(\mathbf{0})). \quad (\text{D.16})$$

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

$$\frac{1}{d} \frac{\lambda_2}{2} \|\hat{\mathbf{W}}\|_F^2 \leq \frac{1}{d} (\mathcal{L}(\mathbf{Y}) + \tilde{r}(\mathbf{0})). \quad (\text{D.17})$$

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

$$\frac{1}{d} \frac{\lambda_2}{2} \|\hat{\mathbf{W}}\|_F^2 \leq \frac{1}{d} \left(C_{\mathcal{L}} \left(1 + \|\mathbf{Y}\|_2^2 \right) \right) + C_{\tilde{r}} \tag{D.18}$$

where, on the right hand side, the term $\|\mathbf{Y}\|_F^2/d = \alpha \|\mathbf{Y}\|_F^2/n$ is bounded since the labels are in $\{-1, +1\}$ and α is finite. Now using the definition of \mathbf{U}

$$\frac{1}{d} \|\mathbf{U}\|_F^2 = \frac{1}{d} \|\mathbf{P}_{\hat{\mathbf{W}}^*}^{\perp} \hat{\mathbf{W}}\|_F^2 \tag{D.19}$$

$$\leq \|\mathbf{P}_{\hat{\mathbf{W}}^*}^{\perp}\|_{op}^2 \frac{1}{d} \|\hat{\mathbf{W}}\|_F^2 \tag{D.20}$$

where the singular values of $\mathbf{P}_{\hat{\mathbf{W}}^*}^{\perp}$ are bounded with probability one. Therefore there exists a constant C_U such that $\|\mathbf{U}\|/\sqrt{d} \leq C_U$. Then, by definition of \mathbf{m} and the Cauchy-Schwarz inequality

$$\|\mathbf{m}\|_F^2 \leq \frac{1}{d} \|\mathbf{c}\|_2^2 \frac{1}{d} \|\mathbf{W}\|_F^2 \tag{D.21}$$

$$\leq \frac{1}{d} \|\mathbf{W}^*\|_2^2 \frac{1}{d} \|\hat{\mathbf{W}}\|_F^2. \tag{D.22}$$

By assumption, the columns of \mathbf{W}^* are sampled from sub-Gaussian distributions, thus, using Bernstein's inequality for sub-exponential random variables there exists a positive constant C_{W^*} such that, with high probability as $n, d \rightarrow +\infty$, $\|\mathbf{W}^*\|_F^2 \leq C_{W^*}$. Combining this with the result on $\hat{\mathbf{W}}$, there exists a positive constant C_m such that $\|\mathbf{m}\|_F \leq C_m$ with high probability as $n, d \rightarrow +\infty$. We finally turn to $\hat{\mathbf{m}}$. The optimality condition for \mathbf{m} in problem equation (D.10) gives

$$\hat{\mathbf{m}} = -\frac{1}{\sqrt{d}} \rho^{-1/2} \mathbf{S}^{\top} \partial \mathcal{L} \left(\mathbf{Y}, \frac{\mathbf{S} \mathbf{m}^{\top}}{\sqrt{\rho}} + \frac{1}{\sqrt{d}} \tilde{\mathbf{X}} \mathbf{C}^{1/2} \mathbf{W}^* \right). \tag{D.23}$$

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

$$\|\hat{\mathbf{m}}\|_2^2 \leq \frac{1}{d} \|\rho^{-1}\|_F \|\mathbf{S}\|_F^2 C_{\partial \mathcal{L}} \left(1 + \frac{1}{d} \|\mathbf{Y}\|_F^2 + \frac{1}{d} \left\| \frac{\mathbf{S} \mathbf{m}^{\top}}{\sqrt{\rho}} + \frac{1}{\sqrt{d}} \tilde{\mathbf{X}} \mathbf{W}^* \right\|_F^2 \right). \tag{D.24}$$

All quantities in the right hand side of the last inequality have bounded scaled norm with high probability, except the operator norm of the random matrix $\tilde{\mathbf{X}}$ which has i.i.d. $\mathcal{N}(0, 1/d)$ elements. Existing results in random matrix theory [38] ensure this operator norm is bounded with high as $n, d \rightarrow +\infty$, which concludes the proof of this lemma. \square

The optimisation problem equation (D.12) is convex and feasible. Furthermore, we may reduce the feasibility sets of $\mathbf{m}, \hat{\mathbf{m}}$ 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 minimisation to obtain the equivalent problem

$$\inf_{\mathbf{m}} \sup_{\hat{\mathbf{m}}} \inf_{\mathbf{U}} \mathcal{L} \left(\mathbf{Y}, \mathbf{S} \rho^{-1/2} \mathbf{m}^{\top} + \frac{1}{\sqrt{d}} \tilde{\mathbf{X}} \mathbf{U} \right) + r(\mathbf{W}^* \rho^{-1} \mathbf{m}^{\top} + \mathbf{U}) - \text{Tr} \left(\hat{\mathbf{m}}^{\top} \mathbf{U}^{\top} \mathbf{W}^* \right) \tag{D.25}$$

and study the optimisation problem in \mathbf{U} at fixed $\mathbf{m}, \hat{\mathbf{m}}$:

$$\inf_{\mathbf{U} \in \mathbb{R}^{d \times k}} \tilde{\mathcal{L}} \left(\frac{1}{\sqrt{d}} \tilde{\mathbf{X}} \mathbf{U} \right) + \tilde{r}(\mathbf{U}) \tag{D.26}$$

where we defined the functions

$$\tilde{\mathcal{L}} : \mathbb{R}^{n \times k} \rightarrow \mathbb{R} \tag{D.27}$$

$$\frac{1}{\sqrt{d}} \tilde{\mathbf{X}} \mathbf{U} \rightarrow \mathcal{L} \left(\mathbf{Y}, \mathbf{S} \rho^{-1/2} \mathbf{m}^{\top} + \frac{1}{\sqrt{d}} \tilde{\mathbf{X}} \mathbf{U} \right) \tag{D.28}$$

$$\tilde{r} : \mathbb{R}^{d \times k} \rightarrow \mathbb{R} \tag{D.29}$$

$$\mathbf{U} \rightarrow r(\mathbf{W}^* \rho^{-1} \mathbf{m}^\top + \mathbf{U}) - \text{Tr}(\hat{\mathbf{m}}^\top \mathbf{U}^\top \mathbf{W}^*) \quad (\text{D.30})$$

and the random matrix $\tilde{\mathbf{X}}$ 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 optimisation problem can now be studied with a non-separable, matrix-valued approximate message passing iteration. The AMP iteration solving problem equation (D.26) is given in the following lemma

Lemma 2. Consider the following AMP iteration

$$\mathbf{u}^{t+1} = \tilde{\mathbf{X}}^\top \mathbf{h}_t(\mathbf{v}^t) - \mathbf{e}_t(\mathbf{u}^t) \langle \mathbf{h}'_t \rangle^\top \quad (\text{D.31})$$

$$\mathbf{v}^t = \tilde{\mathbf{X}} \mathbf{e}_t(\mathbf{u}^t) - \mathbf{h}_{t-1}(\mathbf{v}^{t-1}) \langle \mathbf{e}'_t \rangle^\top \quad (\text{D.32})$$

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

$$\mathbf{h}_t(\mathbf{v}^t) = \left(\mathbf{R}_{\mathcal{L}(\mathbf{Y}, \cdot), \mathbf{S}^t} (\mathbf{S} \rho^{-1/2} \mathbf{m}^\top + \mathbf{v}^t) - (\mathbf{S} \rho^{-1/2} \mathbf{m}^\top + \mathbf{v}^t) \right) (\mathbf{S}^t)^{-1} \quad (\text{D.33})$$

$$\mathbf{e}_t(\mathbf{u}^t) = \mathbf{R}_{r(\cdot), \hat{\mathbf{S}}^t} \left(\mathbf{u}^t \hat{\mathbf{S}}^t + \mathbf{W}^* \hat{\mathbf{m}}^\top \hat{\mathbf{S}}^t + \mathbf{W}^* \rho^{-1} \mathbf{m}^\top \right) - \mathbf{W}^* \rho^{-1} \mathbf{m}^\top \quad (\text{D.34})$$

$$\text{and } \mathbf{S}^t = \langle \langle \mathbf{e}^t \rangle' \rangle^\top, \quad \hat{\mathbf{S}}^t = - \langle \langle \mathbf{h}^t \rangle' \rangle^\top^{-1}. \quad (\text{D.35})$$

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

$$\mathbf{R}_{r(\cdot), \hat{\mathbf{S}}^\infty} \left(\mathbf{u}^\infty \hat{\mathbf{S}}^\infty + \mathbf{W}^* \hat{\mathbf{m}}^\top \hat{\mathbf{S}}^\infty + \mathbf{W}^* \rho^{-1} \mathbf{m}^\top \right) - \mathbf{W}^* \rho^{-1} \mathbf{m}^\top = \mathbf{U}^* \quad (\text{D.36})$$

$$\mathbf{R}_{\mathcal{L}(\mathbf{Y}, \cdot), \mathbf{S}^\infty} (\mathbf{S} \rho^{-1/2} \mathbf{m}^\top + \mathbf{v}^\infty) - \mathbf{S} \rho^{-1/2} \mathbf{m}^\top = \tilde{\mathbf{X}} \mathbf{U}^* \quad (\text{D.37})$$

where \mathbf{U}^* is the unique solution to the optimisation problem equation (D.26).

Proof. To find the correct form of the non-linearities in the AMP iteration, we match the optimality condition of problem equation (D.26) with the generic form of the fixed point of the AMP iteration equation (D.101). In the subsequent derivation, we absorb the scaling $1/\sqrt{d}$ in the matrix $\tilde{\mathbf{X}}$, such that its elements are i.i.d. $\mathcal{N}(0, 1/d)$, and omit time indices for simplicity. Going back to problem equation (D.26), its optimality condition reads :

$$\tilde{\mathbf{X}}^\top \partial \tilde{\mathcal{L}}(\tilde{\mathbf{X}} \mathbf{U}) + \partial \tilde{r}(\mathbf{U}) = 0. \quad (\text{D.38})$$

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

$$\tilde{\mathbf{X}}^\top \left(\partial \tilde{\mathcal{L}}(\tilde{\mathbf{X}} \mathbf{U}) \mathbf{S} + \tilde{\mathbf{X}} \mathbf{U} \right) \mathbf{S}^{-1} + \left(\partial \tilde{r}(\mathbf{U}) \hat{\mathbf{S}} + \mathbf{U} \right) \hat{\mathbf{S}}^{-1} = \tilde{\mathbf{X}}^\top \tilde{\mathbf{X}} \mathbf{U} \mathbf{S}^{-1} + \mathbf{U} \hat{\mathbf{S}}^{-1} \quad (\text{D.39})$$

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:

$$\hat{\mathbf{v}} = \partial \tilde{\mathcal{L}}(\tilde{\mathbf{X}} \mathbf{U}) \mathbf{S} + \tilde{\mathbf{X}} \mathbf{U} \iff \tilde{\mathbf{X}} \mathbf{U} = \mathbf{R}_{\tilde{\mathcal{L}}, \mathbf{S}}(\hat{\mathbf{v}}) \quad (\text{D.40})$$

such that

$$\mathbf{R}_{\tilde{\mathcal{L}}, \mathbf{S}}(\hat{\mathbf{v}}) = (\text{Id} + \partial \tilde{\mathcal{L}}(\bullet) \mathbf{S})^{-1}(\hat{\mathbf{v}}) = \arg \min_{\mathbf{T} \in \mathbb{R}^{n \times k}} \left\{ \tilde{\mathcal{L}}(\mathbf{T}) + \frac{1}{2} \text{tr}((\mathbf{T} - \hat{\mathbf{v}}) \mathbf{S}^{-1} (\mathbf{T} - \hat{\mathbf{v}})^\top) \right\}. \quad (\text{D.41})$$

Similarly for the regularisation, introduce

$$\hat{\mathbf{u}} \equiv \left(\mathbf{I} + \partial\tilde{r}(\bullet)\hat{\mathbf{S}} \right) (\mathbf{U}) \quad \mathbf{U} = \mathbf{R}_{\tilde{r},\hat{\mathbf{S}}}(\hat{\mathbf{u}}) \quad (\text{D.42})$$

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

$$\mathbf{R}_{\tilde{r},\hat{\mathbf{S}}}(\hat{\mathbf{V}}) = \left(\mathbf{I} + \partial\tilde{r}(\bullet)\hat{\mathbf{S}} \right)^{-1} (\hat{\mathbf{V}}) = \arg \min_{\mathbf{T} \in \mathbb{R}^{d \times k}} \left\{ \tilde{r}(\mathbf{T}) + \frac{1}{2} \text{tr} \left((\mathbf{T} - \hat{\mathbf{v}}) \hat{\mathbf{S}}^{-1} (\mathbf{T} - \hat{\mathbf{v}})^\top \right) \right\} \quad (\text{D.43})$$

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 equation (D.39) may then be rewritten as:

$$\tilde{\mathbf{X}}^\top \left(\mathbf{R}_{\tilde{\mathcal{L}},\mathbf{S}}(\hat{\mathbf{V}}) - \hat{\mathbf{V}} \right) \mathbf{S}^{-1} = (\hat{\mathbf{u}} - \mathbf{R}_{\tilde{r},\hat{\mathbf{S}}}(\hat{\mathbf{u}})) \hat{\mathbf{S}}^{-1} \quad (\text{D.44})$$

$$\tilde{\mathbf{X}} \mathbf{R}_{\tilde{r},\hat{\mathbf{S}}}(\hat{\mathbf{u}}) = \mathbf{R}_{\tilde{\mathcal{L}},\mathbf{S}}(\hat{\mathbf{v}}) \quad (\text{D.45})$$

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 equation (D.101), omitting the time indices for simplicity, reads:

$$\mathbf{u} + \mathbf{e}(\mathbf{u}) \langle \mathbf{h}' \rangle^\top = \tilde{\mathbf{X}}^\top \mathbf{h}(\mathbf{v}) \quad (\text{D.46})$$

$$\mathbf{v} + \mathbf{h}(\mathbf{v}) \langle \mathbf{e}' \rangle^\top = \tilde{\mathbf{X}} \mathbf{e}(\mathbf{u}). \quad (\text{D.47})$$

Matching this fixed point with the optimality condition equation (D.44) suggests the following mapping:

$$\begin{aligned} \mathbf{h}(\mathbf{v}) &= \left(\mathbf{R}_{\tilde{\mathcal{L}},\mathbf{S}}(\mathbf{v}) - \mathbf{v} \right) \mathbf{S}^{-1}, & \mathbf{S} &= \langle \mathbf{e}' \rangle^\top, \\ \mathbf{e}(\mathbf{u}) &= \mathbf{R}_{\tilde{r},\hat{\mathbf{S}}}(\mathbf{u}\hat{\mathbf{S}}), & \hat{\mathbf{S}} &= -(\langle \mathbf{h}' \rangle^\top)^{-1}, \end{aligned} \quad (\text{D.48})$$

where we redefined $\hat{\mathbf{u}} \equiv \hat{\mathbf{u}}\hat{\mathbf{S}}$ in (D.42). 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

$$\mathbf{R}_{\tilde{\mathcal{L}},\mathbf{S}}(\mathbf{v}) = \arg \min_{\mathbf{x} \in \mathbb{R}^{n \times k}} \left\{ \mathcal{L} \left(\phi_{\text{out}}(\sqrt{\rho}\mathbf{s}), \mathbf{S}\rho^{-1/2}\mathbf{m}^\top + \mathbf{x} \right) + \frac{1}{2} \text{tr} \left((\mathbf{x} - \mathbf{v}) \mathbf{S}^{-1} (\mathbf{x} - \mathbf{v})^\top \right) \right\} \quad (\text{D.49})$$

letting $\tilde{\mathbf{x}} = \mathbf{S}\rho^{-1/2}\mathbf{m}^\top + \mathbf{x}$,

$$\begin{aligned} \mathbf{R}_{\tilde{\mathcal{L}},\mathbf{S}}(\mathbf{v}) &= \arg \min_{\tilde{\mathbf{x}} \in \mathbb{R}^{n \times k}} \left\{ \mathcal{L}(\phi_{\text{out}}(\sqrt{\rho}\mathbf{s}), \tilde{\mathbf{x}}) \right. \\ &\quad \left. + \frac{1}{2} \text{tr} \left((\tilde{\mathbf{x}} - (\mathbf{S}\rho^{-1/2}\mathbf{m}^\top + \mathbf{v}))\mathbf{S}^{-1}(\tilde{\mathbf{x}} - (\mathbf{S}\rho^{-1/2}\mathbf{m}^\top + \mathbf{v}))^\top \right) \right\} - \mathbf{S}\rho^{-1}\mathbf{m}^\top \end{aligned} \quad (\text{D.50})$$

$$= \mathbf{R}_{\mathcal{L}(\mathbf{Y},\cdot),\mathbf{S}}(\mathbf{S}\rho^{-1/2}\mathbf{m}^\top + \mathbf{v}) - \mathbf{S}\rho^{-1}\mathbf{m}^\top \quad (\text{D.51})$$

and the corresponding non-linearity will then be

$$\mathbf{h}(\mathbf{v}) = \left(\mathbf{R}_{\mathcal{L}(\mathbf{Y},\cdot),\mathbf{S}}(\mathbf{S}\rho^{-1/2}\mathbf{m}^\top + \mathbf{v}) - (\mathbf{S}\rho^{-1/2}\mathbf{m}^\top + \mathbf{v}) \right) \mathbf{S}^{-1}. \quad (\text{D.52})$$

Moving to the regularisation, the resolvent reads

$$\mathbf{R}_{\tilde{r},\hat{\mathbf{S}}}(\mathbf{u}) = \arg \min_{\mathbf{x} \in \mathbb{R}^{d \times k}} \left\{ r(\mathbf{W}^* \rho^{-1} \mathbf{m}^\top + \mathbf{x}) - \text{Tr}(\hat{\mathbf{m}}^\top \mathbf{x}^\top \mathbf{W}^*) + \frac{1}{2} \text{tr} \left((\mathbf{x} - \mathbf{u}) \hat{\mathbf{S}}^{-1} (\mathbf{x} - \mathbf{u})^\top \right) \right\} \quad (\text{D.53})$$

letting $\tilde{\mathbf{x}} = \mathbf{W}^* \rho^{-1} \mathbf{m}^\top + \mathbf{x}$, we obtain

$$\mathbf{R}_{\tilde{r},\hat{\mathbf{S}}}(\mathbf{u}) = \arg \min_{\tilde{\mathbf{x}} \in \mathbb{R}^{d \times k}} \left\{ r(\tilde{\mathbf{x}}) - \hat{\mathbf{m}}^\top \tilde{\mathbf{x}}^\top \mathbf{W}^* \right. \quad (\text{D.54})$$

$$\left. + \frac{1}{2} \text{tr} \left((\tilde{\mathbf{x}} - (\mathbf{u} + \mathbf{W}^* \rho^{-1} \mathbf{m}^\top)) \hat{\mathbf{S}}^{-1} (\tilde{\mathbf{x}} - (\mathbf{u} + \mathbf{W}^* \rho^{-1} \mathbf{m}^\top))^\top \right) \right\} - \mathbf{W}^* \rho^{-1} \mathbf{m}^\top \quad (\text{D.55})$$

$$= \arg \min_{\tilde{\mathbf{x}} \in \mathbb{R}^{d \times k}} \left\{ r(\tilde{\mathbf{x}}) \right. \quad (\text{D.56})$$

$$\left. + \frac{1}{2} \text{tr} \left((\tilde{\mathbf{x}} - (\mathbf{u} + \mathbf{W}^* \hat{\mathbf{m}}^\top \hat{\mathbf{S}} + \mathbf{W}^* \rho^{-1} \mathbf{m}^\top)) \hat{\mathbf{S}}^{-1} (\tilde{\mathbf{x}} - (\mathbf{u} + \mathbf{W}^* \hat{\mathbf{m}}^\top \hat{\mathbf{S}} + \mathbf{W}^* \rho^{-1} \mathbf{m}^\top))^\top \right) \right\} \quad (\text{D.57})$$

$$- \mathbf{W}^* \rho^{-1} \mathbf{m}^\top \quad (\text{D.58})$$

$$\mathbf{R}_{r(\cdot),\hat{\mathbf{S}}}(\mathbf{u} + \mathbf{W}^* \hat{\mathbf{m}}^\top \hat{\mathbf{S}} + \mathbf{W}^* \rho^{-1} \mathbf{m}^\top) - \mathbf{W}^* \rho^{-1} \mathbf{m}^\top \quad (\text{D.59})$$

which gives the following non-linearity for the AMP iteration

$$\mathbf{e}(\mathbf{u}) = \mathbf{R}_{r(\cdot),\hat{\mathbf{S}}}(\mathbf{u} \hat{\mathbf{S}} + \mathbf{W}^* \hat{\mathbf{m}}^\top \hat{\mathbf{S}} + \mathbf{W}^* \rho^{-1} \mathbf{m}^\top) - \mathbf{W}^* \rho^{-1} \mathbf{m}^\top. \quad (\text{D.60})$$

□

The following lemma then gives the exact asymptotics at each time step of the AMP iteration solving problem equation (D.26): its *state evolution equations*.

Lemma 3. Consider the AMP iteration equations (D.31)–(D.35). Assume it is initialised with \mathbf{u}^0 such that $\lim_{d \rightarrow \infty} \frac{1}{d} \|\mathbf{e}_0(\mathbf{u}^0)^\top \mathbf{e}_0(\mathbf{u}^0)\|_F$ exists, a positive definite matrix $\hat{\mathbf{S}}_0$, and $\mathbf{h}_{-1} \equiv 0$. Then for any $t \in \mathbb{N}$, and any pair of sequences of uniformly pseudo-Lipschitz functions $\phi_{1,n} : \mathbb{R}^{d \times k}$ and $\phi_{2,n} : \mathbb{R}^{n \times k}$, the following holds

$$\phi_{1,n}(\mathbf{u}^t) \stackrel{P}{\simeq} \mathbb{E} \left[\phi_{1,n}(\mathbf{G}(\hat{\mathbf{Q}}^t)^{1/2}) \right] \quad (\text{D.61})$$

$$\phi_{2,n}(\mathbf{v}^t) \stackrel{P}{\simeq} \mathbb{E} \left[\phi_{2,n}(\mathbf{H}(\mathbf{Q}^t)^{1/2}) \right] \quad (\text{D.62})$$

where $\mathbf{G} \in \mathbb{R}^{d \times k}$ 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

$$\begin{aligned} \mathbf{Q}^t &= \frac{1}{d} \mathbb{E} \left[\left(\mathbf{R}_{r(\cdot),(\hat{\mathbf{V}}^t)^{-1}}(\mathbf{G}(\hat{\mathbf{Q}}^t)^{1/2}(\hat{\mathbf{V}}^t)^{-1} + \mathbf{W}^* \hat{\mathbf{m}}^\top (\hat{\mathbf{V}}^t)^{-1} + \mathbf{W}^* \rho^{-1} \mathbf{m}^\top) - \mathbf{W}^* \rho^{-1} \mathbf{m}^\top \right)^\top \right. \\ &\quad \left. \times \left(\mathbf{R}_{r(\cdot),(\hat{\mathbf{V}}^t)^{-1}}(\mathbf{G}(\hat{\mathbf{Q}}^t)^{1/2}(\hat{\mathbf{V}}^t)^{-1} + \mathbf{W}^* \hat{\mathbf{m}}^\top (\hat{\mathbf{V}}^t)^{-1} + \mathbf{W}^* \rho^{-1} \mathbf{m}^\top) - \mathbf{W}^* \rho^{-1} \mathbf{m}^\top \right) \right] \end{aligned} \quad (\text{D.63})$$

$$\hat{\mathbf{Q}}^t = \frac{1}{d} \mathbb{E} \left[\left((\mathbf{R}_{\mathcal{L}(\mathbf{y}, \cdot), \mathbf{v}^{t-1}}(\cdot) - Id) \left(\mathbf{S} \rho^{-1/2} \mathbf{m}^\top + \mathbf{H}(\mathbf{Q}^{t-1})^{1/2} \right) (\mathbf{V}^{t-1})^{-1} \right)^\top \right. \\ \left. \times \left((\mathbf{R}_{\mathcal{L}(\mathbf{y}, \cdot), \mathbf{v}^{t-1}}(\cdot) - Id) \left(\mathbf{S} \rho^{-1/2} \mathbf{m}^\top + \mathbf{H}(\mathbf{Q}^{t-1})^{1/2} \right) (\mathbf{V}^{t-1})^{-1} \right) \right] \quad (\text{D.64})$$

$$\mathbf{V}^t = \frac{1}{d} \mathbb{E} \left[(\hat{\mathbf{Q}}^t)^{-1/2} \mathbf{G}^\top \mathbf{R}_{r(\cdot), (\hat{\mathbf{V}}^t)^{-1}} \left(\mathbf{G}(\hat{\mathbf{Q}}^t)^{1/2} (\hat{\mathbf{V}}^t)^{-1} + \mathbf{W}^* \hat{\mathbf{m}}^\top (\hat{\mathbf{V}}^t)^{-1} + \mathbf{W}^* \rho^{-1} \mathbf{m}^\top \right) \right] \quad (\text{D.65})$$

$$\hat{\mathbf{V}}^t = -\frac{1}{d} \mathbb{E} \left[(\mathbf{Q}^{t-1})^{-1/2} \mathbf{H}^\top \left((\mathbf{R}_{\mathcal{L}(\mathbf{y}, \cdot), \mathbf{v}^{t-1}}(\cdot) - Id) \left(\mathbf{S} \rho^{-1/2} \mathbf{m}^\top + \mathbf{H}(\mathbf{Q}^{t-1})^{1/2} \right) \right) (\mathbf{V}^{t-1})^{-1} \right]. \quad (\text{D.66})$$

Proof. Owing to the properties of Bregman proximity operators [39, 40], the update functions in the AMP iteration equations (D.31)–(D.35) are Lipschitz continuous. Thus under the assumptions made on the initialisation, the assumptions of Theorem D.1 are verified, which gives the desired result. \square

Lemma 4. Consider iteration equations (D.31)–(D.35), where the parameters $\mathbf{Q}, \hat{\mathbf{Q}}, \mathbf{V}, \hat{\mathbf{V}}$ are initialised at any fixed point of the state evolution equations of Lemma 3. For any sequence initialised with $\hat{\mathbf{V}}_0 = \hat{\mathbf{V}}$ and \mathbf{u}_0 such that

$$\lim_{d \rightarrow \infty} \frac{1}{d} \mathbf{e}_0(\mathbf{u}_0)^\top \mathbf{e}_0(\mathbf{u}_0) = \mathbf{Q} \quad (\text{D.67})$$

the following holds

$$\lim_{t \rightarrow \infty} \lim_{d \rightarrow \infty} \frac{1}{\sqrt{d}} \|\mathbf{u}^t - \mathbf{u}^*\|_F = 0 \quad \lim_{t \rightarrow \infty} \lim_{d \rightarrow \infty} \frac{1}{\sqrt{d}} \|\mathbf{v}^t - \mathbf{v}^*\|_F = 0. \quad (\text{D.68})$$

Proof. The proof of this lemma is identical to that of Lemma 7 from [20]. \square

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

Lemma 5. For any fixed \mathbf{m} and $\hat{\mathbf{m}}$ in their feasibility sets, let \mathbf{U}^* be the unique solution to the optimisation problem equation (D.26). Then, for any sequences (in the problem dimension) of pseudo-Lipschitz functions of order 2 $\phi_{1,n} : \mathbb{R}^{n \times k} \rightarrow \mathbb{R}$ and $\phi_{2,n} : \mathbb{R}^{d \times k} \rightarrow \mathbb{R}$, the following holds

$$\phi_{1,n}(\mathbf{U}^*) \stackrel{P}{\simeq} \mathbb{E} \left[\phi_{1,n} \left(\mathbf{R}_{r(\cdot), \hat{\mathbf{V}}^{-1}} \left(\mathbf{G} \hat{\mathbf{Q}}^{1/2} \hat{\mathbf{V}}^{-1} + \mathbf{W}^* \hat{\mathbf{m}}^\top \hat{\mathbf{V}}^{-1} + \mathbf{W}^* \rho^{-1} \mathbf{m}^\top \right) - \mathbf{W}^* \rho^{-1} \mathbf{m}^\top \right) \right] \quad (\text{D.69})$$

$$\phi_{2,n} \left(\frac{1}{\sqrt{d}} \tilde{\mathbf{X}} \mathbf{U}^* \right) \stackrel{P}{\simeq} \mathbb{E} \left[\phi_{2,n} \left(\mathbf{R}_{\mathcal{L}(\mathbf{y}, \cdot), \mathbf{V}} \left(\mathbf{S} \rho^{-1/2} \mathbf{m}^\top + \mathbf{H} \hat{\mathbf{Q}}^{1/2} \right) - \mathbf{S} \rho^{-1} \mathbf{m}^\top \right) \right] \quad (\text{D.70})$$

where $\mathbf{G} \in \mathbb{R}^{d \times k}$ 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 of the following set of self consistent equations

$$\mathbf{Q} = \frac{1}{d} \mathbb{E} \left[\left(\mathbf{R}_{r(\cdot), \hat{\mathbf{V}}^{-1}} \left(\mathbf{G} \hat{\mathbf{Q}}^{1/2} \hat{\mathbf{V}}^{-1} + \mathbf{W}^* \hat{\mathbf{m}}^\top \hat{\mathbf{V}}^{-1} + \mathbf{W}^* \rho^{-1} \mathbf{m}^\top \right) - \mathbf{W}^* \rho^{-1} \mathbf{m}^\top \right)^\top \right. \\ \left. \times \left(\mathbf{R}_{r(\cdot), \hat{\mathbf{V}}^{-1}} \left(\mathbf{G} \hat{\mathbf{Q}}^{1/2} \hat{\mathbf{V}}^{-1} + \mathbf{W}^* \hat{\mathbf{m}}^\top \hat{\mathbf{V}}^{-1} + \mathbf{W}^* \rho^{-1} \mathbf{m}^\top \right) - \mathbf{W}^* \rho^{-1} \mathbf{m}^\top \right) \right] \quad (\text{D.71})$$

$$\hat{\mathbf{Q}} = \frac{1}{d} \mathbb{E} \left[\left((\mathbf{R}_{\mathcal{L}(\mathbf{y}, \cdot), \mathbf{V}}(\cdot) - Id) \left(\mathbf{S} \rho^{-1/2} \mathbf{m}^\top + \mathbf{H} \mathbf{Q}^{1/2} \right) \mathbf{V}^{-1} \right)^\top \right. \\ \left. \times \left((\mathbf{R}_{\mathcal{L}(\mathbf{y}, \cdot), \mathbf{V}}(\cdot) - Id) \left(\mathbf{S} \rho^{-1/2} \mathbf{m}^\top + \mathbf{H} \mathbf{Q}^{1/2} \right) \mathbf{V}^{-1} \right) \right] \quad (\text{D.72})$$

$$\mathbf{V} = \frac{1}{d} \mathbb{E} \left[\hat{\mathbf{Q}}^{-1/2} \mathbf{G}^\top \mathbf{R}_{r(\cdot), \hat{\mathbf{V}}^{-1}} \left(\mathbf{G} \hat{\mathbf{Q}}^{1/2} \hat{\mathbf{V}}^{-1} + \mathbf{W}^* \hat{\mathbf{m}}^\top \hat{\mathbf{V}}^{-1} + \mathbf{W}^* \rho^{-1} \mathbf{m}^\top \right) \right] \quad (\text{D.73})$$

$$\hat{\mathbf{V}} = -\frac{1}{d} \mathbb{E} \left[\mathbf{Q}^{-1/2} \mathbf{H}^\top \left((\mathbf{R}_{\mathcal{L}(\mathbf{y}, \cdot), \mathbf{V}}(\cdot) - Id) \left(\mathbf{S} \rho^{-1/2} \mathbf{m}^\top + \mathbf{H} \mathbf{Q}^{1/2} \right) \mathbf{V}^{-1} \right) \right]. \quad (\text{D.74})$$

Proof. Combining the results of the previous lemmas, this proof is close to that of Theorem 1.5 in [29]. \square

Returning to the optimisation problem on $\mathbf{m}, \hat{\mathbf{m}}$ in equation (D.25), the solution \mathbf{U}^* , at any dimension, verifies the zero gradient conditions on $\mathbf{m}, \hat{\mathbf{m}}$:

$$\partial \hat{\mathbf{m}} = 0 \iff (\mathbf{U}^*)^\top \mathbf{W}^* = 0 \tag{D.75}$$

$$\begin{aligned} \partial \mathbf{m} = 0 \iff & \mathbf{m} \rho^{-1/2} \mathbf{S}^\top \partial \mathcal{L} \left(\phi_{\text{out}}(\sqrt{\rho} \mathbf{s}), \mathbf{S} \rho^{-1/2} \mathbf{m}^\top + \frac{1}{\sqrt{d}} \tilde{\mathbf{X}} \mathbf{U} \right) \\ & + \rho^{-1} (\mathbf{W}^*)^\top \partial r((\mathbf{W}^* \rho^{-1} \mathbf{m}^\top + \mathbf{U})) = 0. \end{aligned} \tag{D.76}$$

Using Lemma 5 with the assumption that the gradients of \mathcal{L}, r are pseudo-Lipschitz, we obtain for \mathbf{m}

$$\frac{1}{d} \mathbb{E} \left[\left(\mathbf{R}_{r(\cdot), \hat{\mathbf{V}}^{-1}} \left(\mathbf{G} \hat{\mathbf{Q}}^{1/2} \hat{\mathbf{V}}^{-1} + \mathbf{W}^* \hat{\mathbf{m}}^\top \hat{\mathbf{V}}^{-1} + \mathbf{W}^* \rho^{-1} \mathbf{m}^\top \right) - \mathbf{W}^* \rho^{-1} \mathbf{m}^\top \right)^\top \mathbf{W}^* \right] = 0 \tag{D.77}$$

$$\iff \mathbf{m} = \frac{1}{d} \mathbb{E} \left[(\mathbf{W}^*)^\top \mathbf{R}_{r(\cdot), \hat{\mathbf{V}}^{-1}} \left(\mathbf{G} \hat{\mathbf{Q}}^{1/2} \hat{\mathbf{V}}^{-1} + \mathbf{W}^* \hat{\mathbf{m}}^\top \hat{\mathbf{V}}^{-1} + \mathbf{W}^* \rho^{-1} \mathbf{m}^\top \right) \right] \tag{D.78}$$

and for $\hat{\mathbf{m}}$

$$\frac{1}{d} \mathbb{E} \left[\mathbf{m} \rho^{-1/2} \mathbf{S}^\top \partial \mathcal{L} \left(\phi_{\text{out}}(\sqrt{\rho} \mathbf{s}), \mathbf{R}_{\mathcal{L}(Y, \cdot), \mathbf{V}}(\mathbf{S} \rho^{-1/2} \mathbf{m}^\top + \mathbf{H} \hat{\mathbf{Q}}^{1/2}) \right) \right] \tag{D.79}$$

$$+ \rho^{-1} (\mathbf{W}^*)^\top \partial r \left(\left(\mathbf{R}_{r(\cdot), \hat{\mathbf{V}}^{-1}} \left(\mathbf{G} \hat{\mathbf{Q}}^{1/2} \hat{\mathbf{V}}^{-1} + \mathbf{W}^* \hat{\mathbf{m}}^\top \hat{\mathbf{V}}^{-1} + \mathbf{W}^* \rho^{-1} \mathbf{m}^\top \right) \right) \right) = 0. \tag{D.80}$$

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

$$\frac{1}{d} \mathbb{E} \left[\mathbf{m} \rho^{-1/2} \mathbf{S}^\top \left(\text{Id} - \mathbf{R}_{\mathcal{L}(Y, \cdot), \mathbf{V}}(\cdot) \right) \left(\mathbf{S} \rho^{-1/2} \mathbf{m}^\top + \mathbf{H} \hat{\mathbf{Q}}^{1/2} \right) \mathbf{V}^{-1} \right] \tag{D.81}$$

$$+ \rho^{-1} (\mathbf{W}^*)^\top \left(\text{Id} - \mathbf{R}_{r(\cdot), \hat{\mathbf{V}}^{-1}}(\cdot) \right) \left(\mathbf{G} \hat{\mathbf{Q}}^{1/2} \hat{\mathbf{V}}^{-1} + \mathbf{W}^* \hat{\mathbf{m}}^\top \hat{\mathbf{V}}^{-1} + \mathbf{W}^* \rho^{-1} \mathbf{m}^\top \right) \hat{\mathbf{V}} = 0 \tag{D.82}$$

which simplifies to

$$\hat{\mathbf{m}}^\top = -\frac{1}{d} \mathbb{E} \left[\mathbf{m} \rho^{-1/2} \mathbf{S}^\top \left(\text{Id} - \mathbf{R}_{\mathcal{L}(Y, \cdot), \mathbf{V}}(\cdot) \right) \left(\mathbf{S} \rho^{-1/2} \mathbf{m}^\top + \mathbf{H} \hat{\mathbf{Q}}^{1/2} \right) \mathbf{V}^{-1} \right] \tag{D.83}$$

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

$$\begin{aligned} \mathbf{Q} = & \frac{1}{d} \mathbb{E} \left[\left(\mathbf{R}_{r(\cdot), \hat{\mathbf{V}}^{-1}} \left(\mathbf{G} \hat{\mathbf{Q}}^{1/2} \hat{\mathbf{V}}^{-1} + \mathbf{W}^* \hat{\mathbf{m}}^\top \hat{\mathbf{V}}^{-1} + \mathbf{W}^* \rho^{-1} \mathbf{m}^\top \right) - \mathbf{W}^* \rho^{-1} \mathbf{m}^\top \right)^\top \right. \\ & \left. \times \left(\mathbf{R}_{r(\cdot), \hat{\mathbf{V}}^{-1}} \left(\mathbf{G} \hat{\mathbf{Q}}^{1/2} \hat{\mathbf{V}}^{-1} + \mathbf{W}^* \hat{\mathbf{m}}^\top \hat{\mathbf{V}}^{-1} + \mathbf{W}^* \rho^{-1} \mathbf{m}^\top \right) - \mathbf{W}^* \rho^{-1} \mathbf{m}^\top \right) \right] \end{aligned} \tag{D.84}$$

$$\begin{aligned} \hat{\mathbf{Q}} = & \frac{1}{d} \mathbb{E} \left[\left(\left(\mathbf{R}_{\mathcal{L}(Y, \cdot), \mathbf{V}}(\cdot) - \text{Id} \right) \left(\mathbf{S} \rho^{-1/2} \mathbf{m}^\top + \mathbf{H} \hat{\mathbf{Q}}^{1/2} \right) \mathbf{V}^{-1} \right)^\top \right. \\ & \left. \times \left(\left(\mathbf{R}_{\mathcal{L}(Y, \cdot), \mathbf{V}}(\cdot) - \text{Id} \right) \left(\mathbf{S} \rho^{-1/2} \mathbf{m}^\top + \mathbf{H} \hat{\mathbf{Q}}^{1/2} \right) \mathbf{V}^{-1} \right) \right] \end{aligned} \tag{D.85}$$

$$\mathbf{V} = \frac{1}{d} \mathbb{E} \left[\hat{\mathbf{Q}}^{-1/2} \mathbf{G}^\top \mathbf{R}_{r(\cdot), \hat{\mathbf{V}}^{-1}} \left(\mathbf{G} \hat{\mathbf{Q}}^{1/2} \hat{\mathbf{V}}^{-1} + \mathbf{W}^* \hat{\mathbf{m}}^\top \hat{\mathbf{V}}^{-1} + \mathbf{W}^* \rho^{-1} \mathbf{m}^\top \right) \right] \tag{D.86}$$

$$\hat{\mathbf{V}} = -\frac{1}{d} \mathbb{E} \left[\mathbf{Q}^{-1/2} \mathbf{H}^\top \left(\left(\mathbf{R}_{\mathcal{L}(Y, \cdot), \mathbf{V}}(\cdot) - \text{Id} \right) \left(\mathbf{S} \rho^{-1/2} \mathbf{m}^\top + \mathbf{H} \hat{\mathbf{Q}}^{1/2} \right) \mathbf{V}^{-1} \right) \right] \tag{D.87}$$

$$\mathbf{m} = \frac{1}{d} \mathbb{E} \left[(\mathbf{W}^*)^\top \mathbf{R}_{r(\cdot), \hat{\mathbf{V}}^{-1}} \left(\mathbf{G} \hat{\mathbf{Q}}^{1/2} \hat{\mathbf{V}}^{-1} + \mathbf{W}^* \hat{\mathbf{m}}^\top \hat{\mathbf{V}}^{-1} + \mathbf{W}^* \rho^{-1} \mathbf{m}^\top \right) \right] \tag{D.88}$$

$$\hat{\mathbf{m}}^\top = -\frac{1}{d} \mathbb{E} \left[\mathbf{m} \rho^{-1/2} \mathbf{S}^\top \left(\text{Id} - \mathbf{R}_{\mathcal{L}(Y, \cdot), \mathbf{V}}(\cdot) \right) \left(\mathbf{S} \rho^{-1/2} \mathbf{m}^\top + \mathbf{H} \hat{\mathbf{Q}}^{1/2} \right) \mathbf{V}^{-1} \right]. \tag{D.89}$$

These equations then characterise the asymptotic properties of the quantities $\hat{\mathbf{U}}$ and $\tilde{\mathbf{X}} \hat{\mathbf{U}} / \sqrt{d}$. The properties of $\hat{\mathbf{W}}$ and $\tilde{\mathbf{X}} \hat{\mathbf{W}} / \sqrt{d}$ are then obtained by using the definition of \mathbf{U} in terms of orthogonal decompositions.

Note that To match these equations with the replica ones, we first need to assume the loss and cost functions are separable. The proximal operators are then separable as well across lines of the input matrices. All arguments then have i.i.d. lines (Gaussian matrices with $k \times k$ covariances, or lines of the teacher matrix, which are i.i.d. multiplied with $k \times k$ matrices), and the $1/d$ averages simplify, leaving the aspect ratio in the quantities defined over arguments in $\mathbb{R}^{n \times k}$. The rest of the matching then boils down to identifying the proximal operators with the replica notations, done in appendix A, and standard Gaussian integration, as done for instance in [8], appendix III.3.

D.3. Toolbox

In this section, we reproduce part of the appendix of [20] 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.

D.3.1. Notations

For a given function $\phi: \mathbb{R}^{d \times k} \rightarrow \mathbb{R}^{n \times k}$, we write:

$$\phi(\mathbf{X}) = \begin{bmatrix} \phi^1(\mathbf{X}) \\ \vdots \\ \phi^d(\mathbf{X}) \end{bmatrix} \in \mathbb{R}^{d \times k} \tag{D.90}$$

where each $\phi^i: \mathbb{R}^{d \times k} \rightarrow \mathbb{R}^k$. We then write the $k \times k$ Jacobian

$$\frac{\partial \phi^i}{\partial \mathbf{X}_j}(\mathbf{X}) = \begin{bmatrix} \frac{\partial \phi^i_1(\mathbf{X})}{\partial X_{j1}} & \dots & \frac{\partial \phi^i_1(\mathbf{X})}{\partial X_{jk}} \\ \vdots & \ddots & \vdots \\ \frac{\partial \phi^i_k(\mathbf{X})}{\partial X_{j1}} & \dots & \frac{\partial \phi^i_k(\mathbf{X})}{\partial X_{jk}} \end{bmatrix} \in \mathbb{R}^{k \times k}. \tag{D.91}$$

For a given matrix $\mathbf{Q} \in \mathbb{R}^{k \times k}$, we write $\mathbf{Z} \in \mathbb{R}^{n \times k} \sim \mathcal{N}(\mathbf{0}, \mathbf{Q} \otimes \mathbf{I}_n)$ to denote that the lines of \mathbf{Z} are sampled i.i.d. from $\mathcal{N}(\mathbf{0}, \mathbf{Q})$. Note that this is equivalent to saying that $\mathbf{Z} = \tilde{\mathbf{Z}}\mathbf{Q}^{1/2}$ where $\tilde{\mathbf{Z}} \in \mathbb{R}^{n \times k}$ is an i.i.d. standard normal random matrix. The notation $\stackrel{p}{\simeq}$ denotes convergence in probability. We start with some definitions that commonly appear in the approximate message-passing literature, see e.g. [30, 41]. 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 1 (Pseudo-Lipschitz function). For $K, k \in \mathbb{N}^*$ and any $n, d \in \mathbb{N}^*$, a function $\phi: \mathbb{R}^{d \times k} \rightarrow \mathbb{R}^{n \times k}$ is called a *pseudo-Lipschitz of order K* if there exists a constant $L(K, k)$ such that for any $\mathbf{X}, \mathbf{Y} \in \mathbb{R}^{d \times k}$,

$$\frac{\|\phi(\mathbf{X}) - \phi(\mathbf{Y})\|_F}{\sqrt{n}} \leq L \left(1 + \left(\frac{\|\mathbf{X}\|_F}{\sqrt{d}} \right)^{K-1} + \left(\frac{\|\mathbf{Y}\|_F}{\sqrt{d}} \right)^{K-1} \right) \frac{\|\mathbf{X} - \mathbf{Y}\|_F}{\sqrt{d}} \tag{D.92}$$

where $\|\bullet\|_F$ denotes the Frobenius norm. Since k will be kept finite, it can be absorbed in any of the constants.

For example, the function $f: \mathbb{R}^{d \times k} \rightarrow \mathbb{R}, \mathbf{X} \mapsto \|\mathbf{X}\|_F^2/d$ is pseudo-Lipshitz of order 2.

D.3.2. 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. [42, 43]. 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 [20, 21]. 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 $f: \mathbb{R}^{d \times k} \rightarrow \mathbb{R}$, a given matrix $\mathbf{X} \in \mathbb{R}^{d \times k}$ and a given symmetric positive definite matrix $\mathbf{V} \in \mathbb{R}^{k \times k}$ with bounded spectral norm, we will consider operators of the type

$$\arg \min_{\mathbf{T} \in \mathbb{R}^{d \times k}} \left\{ f(\mathbf{T}) + \frac{1}{2} \text{tr} \left((\mathbf{T} - \mathbf{X}) \mathbf{V}^{-1} (\mathbf{T} - \mathbf{X})^\top \right) \right\}. \tag{D.93}$$

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

$$\text{Prox}_{f(\bullet, \mathbf{V}^{1/2})}(\mathbf{X}\mathbf{V}^{-1/2})\mathbf{V}^{1/2} \in \mathbb{R}^{d \times k} \tag{D.94}$$

or as a Bregman proximal operator [39] defined with the Bregman distance induced by the strictly convex, coercive function (for positive definite V)

$$\mathbf{X} \mapsto \frac{1}{2} \text{tr}(\mathbf{XV}^{-1}\mathbf{X}^\top) \tag{D.95}$$

which justifies the use of the Bregman resolvent

$$\arg \min_{T \in \mathbb{R}^{d \times k}} \left\{ f(T) + \frac{1}{2} \text{tr}((T - \mathbf{X})V^{-1}(T - \mathbf{X})^\top) \right\} = (I + \partial f(\bullet)V)^{-1}(\mathbf{X}). \tag{D.96}$$

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 (D.95), see e.g. [39, 40]. 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 $g : \mathcal{X} \rightarrow \mathbb{R}$, where \mathcal{X} is a given input Hilbert space. Let T be the associated Bregman proximal of a given convex function $f : \mathcal{X} \rightarrow \mathbb{R}$, i.e. for any $\mathbf{x} \in \mathcal{X}$

$$T(\mathbf{x}) = \arg \min_{\mathbf{y} \in \mathcal{X}} \{ f(\mathbf{x}) + D_g(\mathbf{x}, \mathbf{y}) \}. \tag{D.97}$$

Then T is D -firm, meaning it verifies

$$\langle T\mathbf{x} - T\mathbf{y}, \nabla g(T\mathbf{x}) - \nabla g(T\mathbf{y}) \rangle \leq \langle T\mathbf{x} - T\mathbf{y}, \nabla g(\mathbf{x}) - \nabla g(\mathbf{y}) \rangle \tag{D.98}$$

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

D.3.3. 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 [44]. 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 [30] included in [31]. Consider a sequence Gaussian matrices $\mathbf{A}(n) \in \mathbb{R}^{n \times d}$ 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

$$\{ \mathbf{h}_t : \mathbb{R}^{n \times k} \rightarrow \mathbb{R}^{n \times k} \}_{t \in \mathbb{N}} \quad \{ \mathbf{e}_t : \mathbb{R}^{d \times k} \rightarrow \mathbb{R}^{d \times k} \}_{t \in \mathbb{N}} \tag{D.99}$$

initialised on $\mathbf{u}^0 \in \mathbb{R}^{d \times k}$ in such a way that the limit

$$\lim_{d \rightarrow \infty} \frac{1}{d} \| \mathbf{e}_0(\mathbf{u}^0)^\top \mathbf{e}_0(\mathbf{u}^0) \|_F \tag{D.100}$$

exists, and recursively define:

$$\mathbf{u}^{t+1} = \mathbf{A}^\top \mathbf{h}_t(\mathbf{v}^t) - \mathbf{e}_t(\mathbf{u}^t) \langle \mathbf{h}'_t \rangle^\top \tag{D.101}$$

$$\mathbf{v}^t = \mathbf{A} \mathbf{e}_t(\mathbf{u}^t) - \mathbf{h}_{t-1}(\mathbf{v}^{t-1}) \langle \mathbf{e}'_t \rangle^\top \tag{D.102}$$

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

$$\langle \mathbf{h}'_t \rangle = \frac{1}{d} \sum_{i=1}^n \frac{\partial \mathbf{h}_t^i}{\partial \mathbf{v}_i}(\mathbf{v}^t) \in \mathbb{R}^{k \times k} \quad \langle \mathbf{e}'_t \rangle = \frac{1}{d} \sum_{i=1}^d \frac{\partial \mathbf{e}_t^i}{\partial \mathbf{u}_i}(\mathbf{u}^t) \in \mathbb{R}^{k \times k}. \tag{D.103}$$

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

$$\mathbf{Q}_{t+1,s} = \mathbf{Q}_{s,t+1} = \lim_{d \rightarrow \infty} \frac{1}{d} \mathbb{E} \left[\mathbf{e}_s(\hat{\mathbf{Z}}^s)^\top \mathbf{e}_{t+1}(\hat{\mathbf{Z}}^{t+1}) \right] \in \mathbb{R}^{k \times k} \tag{D.104}$$

$$\hat{\mathbf{Q}}_{t+1,s+1} = \hat{\mathbf{Q}}_{s+1,t+1} = \lim_{d \rightarrow \infty} \frac{1}{d} \mathbb{E} \left[\mathbf{h}_s(\mathbf{Z}^s)^\top \mathbf{h}_t(\mathbf{Z}^t) \right] \in \mathbb{R}^{k \times k} \tag{D.105}$$

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

Algorithm 1. Approximate message passing.**Input:** data matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$ and label matrix $\mathbf{Y} \in \mathbb{R}^{n \times k}$ Initialise $\hat{\mathbf{w}}_j^0 \in \mathbb{R}^k$ and $\hat{\mathbf{C}}_j^0, \mathbf{f}_{\text{out},\nu}^0 \in \mathbb{R}^{k \times k}$ for $j = 1, \dots, d$ and $\nu = 1, \dots, n$ at $t = 0$ **repeat**

Channel updates

 Mean $\boldsymbol{\omega}_\nu \in \mathbb{R}^k$ and variance $\mathbf{V}_\nu \in \mathbb{R}^{k \times k}$

$$\mathbf{V}_\nu^t = \frac{1}{d} \sum_{j=1}^d x_{j\nu}^2 \hat{\mathbf{C}}_j^t$$

$$\boldsymbol{\omega}_\nu^t = \frac{1}{\sqrt{d}} \sum_{l=1}^d \hat{\mathbf{w}}_l^t - (\mathbf{V}_\nu^t)^\top \mathbf{f}_{\text{out},\nu}^{t-1}$$

 Denoisers $\mathbf{f}_{\text{out},\nu} \in \mathbb{R}^k$ and $\partial_\omega \mathbf{f}_{\text{out},\nu} \in \mathbb{R}^{k \times k}$

$$\mathbf{f}_{\text{out},\nu}^t \leftarrow \mathbf{f}_{\text{out}}(\mathbf{y}_\nu, \boldsymbol{\omega}_\nu^t, \mathbf{V}_\nu^t)$$

$$\partial_\omega \mathbf{f}_{\text{out},\nu}^t \leftarrow \partial_\omega \mathbf{f}_{\text{out}}(\mathbf{y}_\nu, \boldsymbol{\omega}_\nu^t, \mathbf{V}_\nu^t)$$

Prior updates

 Mean $\boldsymbol{\gamma}_j \in \mathbb{R}^k$ and variance $\boldsymbol{\Lambda}_j \in \mathbb{R}^{k \times k}$

$$\boldsymbol{\Lambda}_j^t = -\frac{1}{d} \sum_{\nu=1}^n x_{j\nu}^2 \partial_\omega \mathbf{f}_{\text{out},\nu}^t$$

$$\boldsymbol{\gamma}_j^t = \frac{1}{\sqrt{d}} \sum_{\nu=1}^n x_{j\nu} \mathbf{f}_{\text{out},\nu}^t + \boldsymbol{\Lambda}_j^t \hat{\mathbf{w}}_j^t$$

 Posterior estimators $\hat{\mathbf{w}}_j \in \mathbb{R}^k$ and $\hat{\mathbf{C}}_j \in \mathbb{R}^{k \times k}$

$$\hat{\mathbf{w}}_j^t = f_w(\boldsymbol{\gamma}_j^t, \boldsymbol{\Lambda}_j^t)$$

$$\hat{\mathbf{C}}_j^t = \partial_\gamma f_w(\boldsymbol{\gamma}_j^t, \boldsymbol{\Lambda}_j^t)$$

 $t \leftarrow t + 1$ **until** Convergence on $\hat{\mathbf{w}}_j$ and $\hat{\mathbf{C}}_j$ **Output:** $\{\hat{\mathbf{w}}_j\}_{j=1}^d$ and $\{\hat{\mathbf{C}}_j\}_{j=1}^d$.**Theorem D.1.** In the setting of the previous paragraph, for any sequence of pseudo-Lipschitz functions $\phi_n : (\mathbb{R}^{n \times K} \times \mathbb{R}^{d \times k})^t \rightarrow \mathbb{R}$, for $n, d \rightarrow +\infty$:

$$\phi_n(\mathbf{u}^0, \mathbf{v}^0, \mathbf{u}^1, \mathbf{v}^1, \dots, \mathbf{v}^{t-1}, \mathbf{u}^t) \stackrel{P}{\simeq} \mathbb{E} \left[\phi_n(\mathbf{u}^0, \mathbf{Z}^0, \hat{\mathbf{Z}}^1, \mathbf{Z}^1, \dots, \mathbf{Z}^{t-1}, \hat{\mathbf{Z}}^t) \right] \quad (\text{D.106})$$

where $(\mathbf{Z}^0, \dots, \mathbf{Z}^{t-1}) \sim \mathcal{N}(\mathbf{0}, \{\mathbf{Q}_{r,s}\}_{0 \leq r,s \leq t-1} \otimes \mathbf{I}_n)$, $(\hat{\mathbf{Z}}^1, \dots, \hat{\mathbf{Z}}^t) \sim \mathcal{N}(\mathbf{0}, \{\hat{\mathbf{Q}}_{r,s}\}_{1 \leq r,s \leq t} \otimes \mathbf{I}_n)$.**Appendix E. Approximate message-passing algorithm: pseudo-code**

In this appendix we present in Algorithm 1 a pseudo-code for the *Approximate message Passing* (AMP) algorithm used in this work.

The update functions $f_w(\boldsymbol{\gamma}, \boldsymbol{\Lambda})$ and $f_{\text{out}}(\mathbf{y}, \boldsymbol{\omega}, \mathbf{V})$ are defined in appendix C.1. For a general and detailed derivation of the algorithm see [37, 44].

Appendix F. AMP implementation: channel and prior updates for $k = 3$

In this appendix we present the expression of the integrals numerically computed for the implementation of Algorithm 1 in the present case.

Apart from the update functions $f_w(\boldsymbol{\gamma}, \boldsymbol{\Lambda})$ and $f_{\text{out}}(\mathbf{y}, \boldsymbol{\omega}, \mathbf{V})$ in appendix C.1, the approximate message passing algorithm also requires the variance updates. Using the mapping from appendix A, the updates are computed though $(k-1) \times (k-1)$ matrices constructed from the derivatives of the denoising functions:

$$\partial_\gamma f_w(\boldsymbol{\gamma}, \boldsymbol{\Lambda}) = \mathbb{E}_{Q_0}[\mathbf{w}\mathbf{w}^\top] - f_w(\boldsymbol{\gamma}, \boldsymbol{\Lambda}) f_w^\top(\boldsymbol{\gamma}, \boldsymbol{\Lambda}), \quad (\text{F.1a})$$

$$\partial_\omega f_{\text{out}}(\mathbf{y}, \boldsymbol{\omega}, \mathbf{V}) = \mathbf{V}^{-1} \mathbb{E}_{Q_{\text{out}}}[(\mathbf{z} - \boldsymbol{\omega})(\mathbf{z} - \boldsymbol{\omega})^\top] - \mathbf{V}^{-1} - f_{\text{out}}(\boldsymbol{\gamma}, \boldsymbol{\Lambda}) f_{\text{out}}^\top(\boldsymbol{\gamma}, \boldsymbol{\Lambda}). \quad (\text{F.1b})$$

F.1. Channel updates

Under the mapping from appendix A, for $k = 3$ we have two sets of integrals related to the channel: one when $\mathbf{y} = [0, 0]^\top$ and other when $\mathbf{y} = [1, 0]^\top$ or $\mathbf{y} = [0, 1]^\top$. A simple flip on the variables distinguishes these two latter cases.

We introduce the notation,

$$\mathbf{V}^{-1} \equiv \begin{bmatrix} \mathcal{V}_{11} & \mathcal{V}_{12} \\ \mathcal{V}_{21} & \mathcal{V}_{22} \end{bmatrix}, \tag{E.2a}$$

as well as the quantities

$$\bar{\mathcal{V}} \equiv \frac{\mathcal{V}_{12} + \mathcal{V}_{21}}{2}, \tag{E.2b}$$

$$\nu \equiv \frac{\mathcal{V}_{11}\mathcal{V}_{22}}{\bar{\mathcal{V}}^2}. \tag{E.2c}$$

Additionally, the standard Gaussian measure is denoted as

$$\mathcal{D}z \equiv \frac{dz}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}, \tag{E.3}$$

and the cumulative distribution function of the standard Gaussian distribution as

$$\Phi(x) \equiv \int_{-\infty}^x \mathcal{D}z. \tag{E.4}$$

F.1.1. Case $\mathbf{y} = [0, 0]^T$

The channel function $f_{\text{out}}(\mathbf{y}, \boldsymbol{\omega}, \mathbf{V})$ is obtained through the numerical computation of the following quantities:

$$\mathcal{Z}_{\text{out}}^{(00)} = \sqrt{\frac{\pi^2}{\mathcal{V}_{22}}} \frac{\omega_1}{\alpha_{12}} \mathcal{J}_0^{(00)}, \tag{E.5a}$$

$$\frac{\partial}{\partial \omega_1} \mathcal{Z}_{\text{out}}^{(00)} = -\sqrt{2\pi} \frac{\bar{\mathcal{V}}}{\sqrt{\mathcal{V}_{11}\mathcal{V}_{22}^2}} e^{-\frac{1}{2}\alpha_{21}^2} \Phi(-\gamma_{12}) + \sqrt{\frac{\pi^2}{\mathcal{V}_{22}}} \mathcal{J}_1^{(00)}, \tag{E.5b}$$

$$\frac{\partial}{\partial \omega_2} \mathcal{Z}_{\text{out}}^{(00)} = -\sqrt{\frac{2\pi}{\mathcal{V}_{11}}} e^{-\frac{1}{2}\alpha_{21}^2} \Phi(-\gamma_{12}), \tag{E.5c}$$

and its derivative $\partial_{\boldsymbol{\omega}} f_{\text{out}}(\mathbf{y}, \boldsymbol{\omega}, \mathbf{V})$ by computing

$$\begin{aligned} \frac{\partial^2}{\partial \omega_1^2} \mathcal{Z}_{\text{out}}^{(00)} &= \sqrt{2\pi} \frac{\bar{\mathcal{V}}^3}{\mathcal{V}_{11}\mathcal{V}_{22}^2} e^{-\frac{1}{2}\alpha_{21}^2} \left[\frac{e^{-\frac{1}{2}\gamma_{12}^2}}{\sqrt{2\pi}} (2\nu - 1) - \frac{\bar{\mathcal{V}}\omega_2}{\sqrt{\mathcal{V}_{11}}} (\nu - 1) \Phi(-\gamma_{12}) \right] \\ &\quad - \frac{\alpha_{12}^2}{\omega_1^2} \mathcal{Z}_{\text{out}}^{(00)} + \sqrt{\frac{\pi^2}{\mathcal{V}_{22}}} \frac{\alpha_{12}}{\omega_1} \mathcal{J}_2^{(00)}, \end{aligned} \tag{E.5d}$$

$$\frac{\partial^2}{\partial \omega_2^2} \mathcal{Z}_{\text{out}}^{(00)} = \sqrt{2\pi} \frac{\bar{\mathcal{V}}}{\mathcal{V}_{11}} e^{-\frac{1}{2}\alpha_{21}^2} \left[\frac{e^{-\frac{1}{2}\gamma_{12}^2}}{\sqrt{2\pi}} + \frac{\sqrt{\mathcal{V}_{11}}}{\omega_2 \bar{\mathcal{V}}} \alpha_{21}^2 \Phi(-\gamma_{12}) \right], \tag{E.5e}$$

$$\frac{\partial^2}{\partial \omega_1 \omega_2} \mathcal{Z}_{\text{out}}^{(00)} = e^{-\frac{1}{2}(\alpha_{21}^2 + \gamma_{12}^2)}, \tag{E.5f}$$

with

$$\mathcal{J}_l^{(00)} \equiv \int_{-\infty}^{-\alpha_{12}} \mathcal{D}z z^l \operatorname{erfc} \left(\frac{-z\bar{\mathcal{V}}\omega_1 + \mathcal{V}_{22}\omega_2}{\sqrt{2\mathcal{V}_{22}}} \right), \tag{E.6}$$

for $l = 0, 1, 2$ and

$$\alpha_{12} \equiv \omega_1 \sqrt{\mathcal{V}_{11} - \frac{\bar{\mathcal{V}}^2}{\mathcal{V}_{22}}}. \tag{E.7a}$$

$$\gamma_{12} \equiv \frac{\mathcal{V}_{11}\omega_1 + \bar{\mathcal{V}}\omega_2}{\sqrt{\mathcal{V}_{11}}}. \tag{E.7b}$$

F.1.2. Case $\mathbf{y} = [1, 0]^T$

The channel function $f_{\text{out}}(\mathbf{y}, \boldsymbol{\omega}, \mathbf{V})$ is obtained through the numerical computation of the following quantities:

$$\mathcal{Z}_{\text{out}}^{(10)} = \sqrt{\frac{\pi^2}{\mathcal{V}_{22}}} \frac{\omega_1}{\alpha_{12}} \mathcal{J}_0^{(10)}, \tag{F.8a}$$

$$\frac{\partial}{\partial \omega_1} \mathcal{Z}_{\text{out}}^{(10)} = -\sqrt{2\pi} \frac{\bar{\mathcal{V}}}{\sqrt{\mathcal{S}_{\mathcal{V}} \mathcal{V}_{22}^2}} e^{-\frac{1}{2}(\Omega^2 - \beta^2)} \tilde{\Phi}(-\beta) + \sqrt{\frac{\pi^2}{\mathcal{V}_{22}}} \mathcal{J}_1^{(10)}, \tag{F.8b}$$

$$\frac{\partial}{\partial \omega_2} \mathcal{Z}_{\text{out}}^{(10)} = -\sqrt{\frac{2\pi}{\mathcal{S}_{\mathcal{V}}}} e^{-\frac{1}{2}(\Omega^2 - \beta^2)} \tilde{\Phi}(-\beta), \tag{F.8c}$$

and its derivative $\partial_{\omega} f_{\text{out}}(\mathbf{y}, \boldsymbol{\omega}, \mathbf{V})$ by computing

$$\begin{aligned} \frac{\partial^2}{\partial \omega_1^2} \mathcal{Z}_{\text{out}}^{(10)} = & -\left(\frac{\alpha_{12}}{\omega_1}\right)^2 \mathcal{Z}_{\text{out}}^{(10)} + \sqrt{\frac{\pi^2}{\mathcal{V}_{22}}} \mathcal{J}_2^{(10)} + \sqrt{2\pi} \frac{\bar{\mathcal{V}}^3}{\mathcal{S}_{\mathcal{V}} \mathcal{V}_{22}^2} \sigma_{12} e^{-\frac{1}{2}(\Omega^2 - \beta^2)} \\ & \times \left[\frac{e^{-\frac{1}{2}\beta^2}}{\sqrt{2\pi}} + \left(\beta - \sqrt{\mathcal{S}_{\mathcal{V}}} \left(\left(1 + \frac{\mathcal{V}_{22}}{\bar{\mathcal{V}} \sigma_{12}} \right) \omega_1 - \frac{\mathcal{V}_{22}}{\bar{\mathcal{V}} \sigma_{12}} \omega_2 \right) \right) \tilde{\Phi}(-\beta) \right], \end{aligned} \tag{F.8d}$$

$$\frac{\partial^2}{\partial \omega_2^2} \mathcal{Z}_{\text{out}}^{(10)} = -\sqrt{2\pi} \frac{\bar{\mathcal{V}} + \mathcal{V}_{22}}{\mathcal{S}_{\mathcal{V}}} e^{-\frac{1}{2}(\Omega^2 - \beta^2)} \left[\frac{e^{-\frac{1}{2}\beta^2}}{\sqrt{2\pi}} + \left(\beta - \sqrt{\mathcal{S}_{\mathcal{V}}} \left(\frac{\bar{\mathcal{V}} \omega_1 + \mathcal{V}_{22} \omega_2}{\bar{\mathcal{V}} + \mathcal{V}_{22}} \right) \right) \tilde{\Phi}(-\beta) \right], \tag{F.8e}$$

$$\frac{\partial^2}{\partial \omega_1 \partial \omega_2} \mathcal{Z}_{\text{out}}^{(10)} = -\sqrt{2\pi} \frac{\bar{\mathcal{V}} + \mathcal{V}_{11}}{\mathcal{S}_{\mathcal{V}}} e^{-\frac{1}{2}(\Omega^2 - \beta^2)} \left[\frac{e^{-\frac{1}{2}\beta^2}}{\sqrt{2\pi}} - \frac{\mathcal{V}_{11} \mathcal{V}_{22} - \bar{\mathcal{V}}^2}{\mathcal{S}_{\mathcal{V}} (\bar{\mathcal{V}} + \mathcal{V}_{11})} \sqrt{\mathcal{S}_{\mathcal{V}}} (\omega_1 - \omega_2) \tilde{\Phi}(-\beta) \right], \tag{F.8f}$$

where

$$\mathcal{J}_l^{(10)} \equiv \int_{-\alpha_{12}}^{\infty} \mathcal{D}z z^l \text{erfc} \left(\frac{-z \omega_1 (\bar{\mathcal{V}} + \mathcal{V}_{22})}{\alpha_{12}} + \mathcal{V}_{22} (\omega_2 - \omega_1) \right), \tag{F.9}$$

for $l = 0, 1, 2$ with α_{12} given by equation (F.7a) and

$$\tilde{\Phi}(x) \equiv 1 - \Phi(x), \tag{F.10a}$$

$$\beta \equiv \frac{\omega_1 (\mathcal{V}_{11} + \bar{\mathcal{V}}) + \omega_2 (\mathcal{V}_{22} + \bar{\mathcal{V}})}{\sqrt{\mathcal{V}_{11} + \mathcal{V}_{22} + 2\bar{\mathcal{V}}}}, \tag{F.10b}$$

$$\Omega^2 \equiv \mathcal{V}_{11} \omega_1 + \mathcal{V}_{22} \omega_2 + 2\bar{\mathcal{V}} \omega_1 \omega_2, \tag{F.10c}$$

$$\mathcal{S}_{\mathcal{V}} \equiv \mathcal{V}_{11} + \mathcal{V}_{22} + 2\bar{\mathcal{V}}, \tag{F.10d}$$

$$\sigma_{12} \equiv \frac{\bar{\mathcal{V}}^2 - 2\bar{\mathcal{V}}_1 \bar{\mathcal{V}}_{22} - \bar{\mathcal{V}}_{22} \bar{\mathcal{V}}}{\bar{\mathcal{V}}^2}. \tag{F.10e}$$

If the label vector is $\mathbf{y} = [0, 1]^T$, one just needs to perform the following trivial changes in the equations above for $\mathbf{y} = [1, 0]^T$:

$$\mathcal{V}_{11} \rightarrow \mathcal{V}_{22}, \tag{F.11a}$$

$$\mathcal{V}_{22} \rightarrow \mathcal{V}_{11}, \tag{F.11b}$$

$$\omega_1 \rightarrow \omega_2, \tag{F.11c}$$

$$\omega_2 \rightarrow \omega_1, \tag{F.11d}$$

Observe that the mapping from appendix A has allowed us to reduce the number of integrals to be numerically computed at each AMP iteration to three, given by equation (F.6) or equation (F.9), depending on the one-hot output representation \mathbf{y} . These integrals were solved through the `integrate.quad` module from SciPy [45]. To speed up the integration, we have also used Numba [46] decorators.

F.2. Prior updates

F.2.1. Gaussian prior

Under the mapping of appendix A, the prior partition function is written as

$$\mathcal{Z}_w(\gamma, \Lambda) = \int_{\mathbb{R}^{k-1}} \frac{dw}{\sqrt{(2\pi)^{k-1} \det(\tilde{\Sigma})}} \exp \left[-\frac{1}{2} \mathbf{w}^\top (\tilde{\Sigma}^{-1} + \Lambda) \mathbf{w} + \gamma^\top \mathbf{w} \right], \quad (\text{F.12})$$

and can be analytically computed,

$$\mathcal{Z}_w(\gamma, \Lambda) = \frac{1}{\sqrt{\det(\tilde{\Sigma}) \det(\tilde{\Sigma}^{-1} + \Lambda)}} \exp \left[\frac{1}{2} \gamma^\top (\tilde{\Sigma}^{-1} + \Lambda)^{-1} \gamma \right], \quad (\text{F.13a})$$

as well as the denoising functions:

$$f_w(\gamma, \Lambda) = \partial_\gamma \log \mathcal{Z}_w(\gamma, \Lambda) = (\tilde{\Sigma}^{-1} + \Lambda)^{-1} \gamma, \quad (\text{F.13b})$$

$$\partial_\gamma f_w(\gamma, \Lambda) = (\tilde{\Sigma}^{-1} + \Lambda)^{-1}. \quad (\text{F.13c})$$

For $k = 3$, the reduced covariance matrix is given by

$$\tilde{\Sigma} = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}. \quad (\text{F.14})$$

F.2.2. Rademacher prior

Considering $k = 3$, the reduced prior given by equation (A.11) becomes

$$P_{\tilde{w}}(\tilde{w}_1, \tilde{w}_2) = \frac{1}{2^3} [2\delta(\tilde{w}_1)\delta(\tilde{w}_2) + \delta(\tilde{w}_1)\delta(\tilde{w}_2 + 2) + \delta(\tilde{w}_1 + 2)\delta(\tilde{w}_2) + \delta(\tilde{w}_1)\delta(\tilde{w}_2 - 2) + \delta(\tilde{w}_1 - 2)\delta(\tilde{w}_2) + \delta(\tilde{w}_1 + 2)\delta(\tilde{w}_2 + 2) + \delta(\tilde{w}_1 - 2)\delta(\tilde{w}_2 - 2)]. \quad (\text{F.15})$$

The denoising functions for this case are computed numerically, via Monte Carlo sampling of the distribution given equation (F.15).

For more details, see the `amp` folder on Github¹².

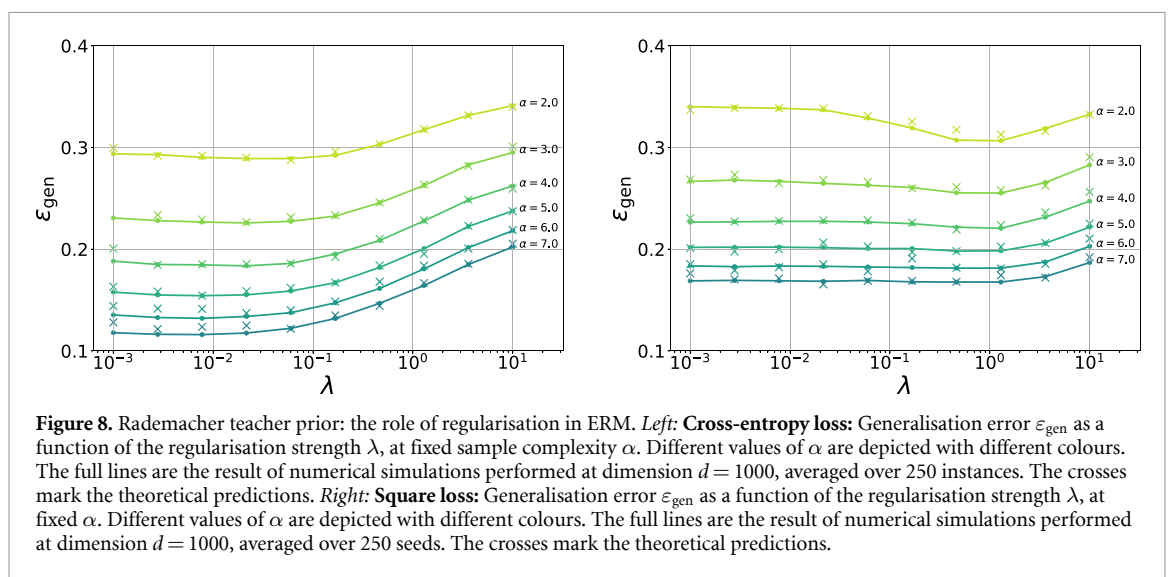
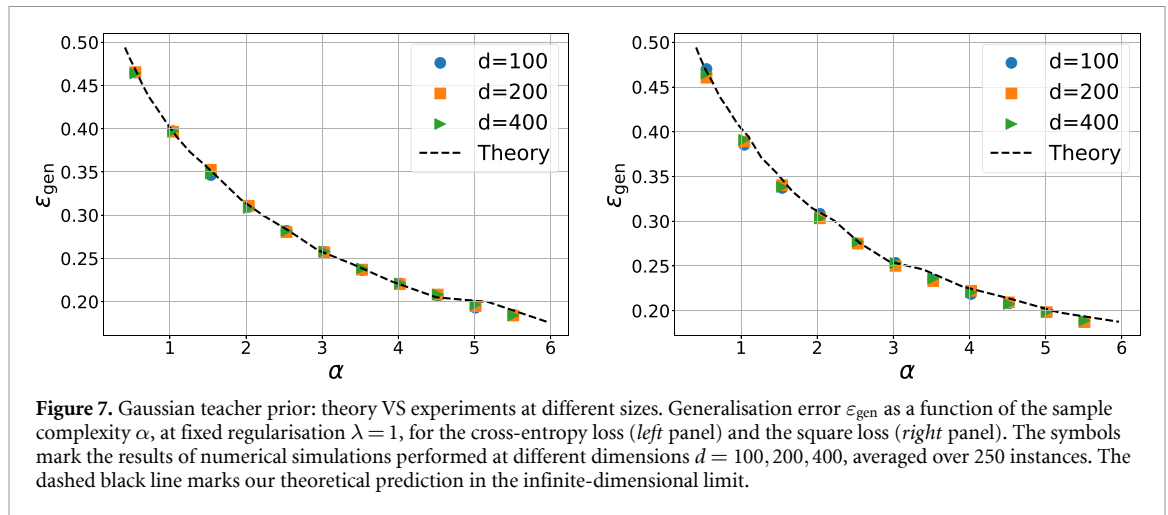
Appendix G. Details on the numerical simulations

In this section we provide some more details on the numerical simulations implemented to test our theory for the learning curves of ERM (figures 3 and 4). The solution of the convex optimisation problem defined in equation (2) can be computed by a standard gradient descent algorithm. We ran simulations using the squared loss and the cross-entropy loss. The simulations for the cross-entropy loss have been implemented using the `LogisticRegression` module of the `scikit-learn` package [33]. The solution for the square loss is analytical. The results from numerical simulations that we show in the figures of the main text are averaged over 250 instances of the problem at dimension $d = 1000$.

In figure 7, we show the comparison between the generalisation error curves derived by our theory and experiments at various system sizes ($d = 100, 200, 400$) for both the cross-entropy and the square loss at fixed regularisation $\lambda = 1$. We find that our theoretical predictions—derived in the infinite-dimensional limit—are still valid at moderately large system sizes. We notice that in our experiments we fix the dimension d and vary $n = \alpha d$ accordingly. This choice is arbitrary and the opposite procedure would have led to similar results.

In figure 8, we explore the role of regularisation for ERM on labels generated by a Rademacher teacher prior. Notice that we do not enforce any constraints on the weights other than ridge regularisation during the optimisation. We notice a qualitatively similar behaviour with respect as for the Gaussian teacher prior, reproduced in figure 4 of the main text. Also in this case, we observe a very mild dependence of the optimal regularisation on the sample complexity α . However, at variance with the Gaussian case, here we observe a clear sub-optimality with respect to the Bayes-error. Indeed, the ERM error is bounded away from zero even at large values of the sample complexity, where the Bayes-optimal AMP algorithm is able to achieve perfect classification.

¹² Code repository: https://github.com/rodsveiga/mc_perceptron.



ORCID iDs

Francesca Mignacco  <https://orcid.org/0000-0001-9944-2498>

Rodrigo Veiga  <https://orcid.org/0000-0002-6835-4871>

References

- [1] Gardner E and Derrida B 1989 Three unfinished works on the optimal storage capacity of networks *J. Phys. A: Math. Gen.* **22** 1983
- [2] Seung H S, Sompolinsky H and Tishby N 1992 Statistical mechanics of learning from examples *Phys. Rev. A* **45** 6056
- [3] Watkin T L, Rau A and Biehl M 1993 The statistical mechanics of learning a rule *Rev. Mod. Phys.* **65** 499
- [4] Engel A and Van den Broeck C 2001 *Statistical Mechanics of Learning* (Cambridge: Cambridge University Press)
- [5] Györgyi G 1990 First-order transition to perfect generalization in a neural network with binary synapses *Phys. Rev. A* **41** 7097
- [6] Sompolinsky H, Tishby N and Seung H S 1990 Learning from examples in large neural networks *Phys. Rev. Lett.* **65** 1683
- [7] Barbier J, Krzakala F, Macris N, Miolane L and Zdeborová L 2019 Optimal errors and phase transitions in high-dimensional generalized linear models *Proc. Natl Acad. Sci.* **116** 5451–60
- [8] Aubin B, Krzakala F, Lu Y and Zdeborová L 2020 Generalization error in high-dimensional perceptrons: approaching bayes error with convex optimization *Advances in Neural Information Processing Systems* vol 33 (New York: Curran Associates, Inc.) pp 12199–210
- [9] Sollich P and Ashton S 2012 Learning curves for multi-task gaussian process regression *Advances in Neural Information Processing Systems* vol 25 (New York: Curran Associates, Inc.)
- [10] Loureiro B, Sicuro G, Gerbelot C, Pacco A, Krzakala F and Zdeborová L 2021 Learning gaussian mixtures with generalized linear models: precise asymptotics in high-dimensions *Advances in Neural Information Processing Systems* vol 34 (New York: Curran Associates, Inc.) pp 10144–157
- [11] Wang K, Muthukumar V, and Thrampoulidis C 2021 Benign overfitting in multiclass classification: all roads lead to interpolation (arXiv:2106.10865)
- [12] Kini G R and Thrampoulidis C 2021 Phase transitions for one-vs-one and one-vs-all linear separability in multiclass gaussian mixtures *ICASSP 2021–2021 IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)* pp 4020–4

- [13] Thrampoulidis C 2020 Theoretical insights into multiclass classification: a high-dimensional asymptotic view *Neural Information Processing Systems (NeurIPS 2020)*
- [14] Mai X, Liao Z and Couillet R 2019 A large scale analysis of logistic regression: asymptotic performance and new insights *ICASSP 2019-2019 IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)* (IEEE) pp 3357–61
- [15] Deng Z, Kammoun A and Thrampoulidis C 2020 A model of double descent for high-dimensional logistic regression *ICASSP 2020–2020 IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)* (IEEE) pp 4267–71
- [16] Kini G R and Thrampoulidis C 2020 Analytic study of double descent in binary classification: the impact of loss *2020 IEEE Int. Symp. on Information Theory (ISIT)* (IEEE) pp 2527–32
- [17] Mignacco F, Krzakala F, Lu Y, Urbani P and Zdeborova L 2020 The role of regularization in classification of high-dimensional noisy gaussian mixture *Int. Conf. on Machine Learning* (PMLR) pp 6874–83
- [18] Aubin B, Maillard A, Barbier J, Krzakala F, Macris N and Zdeborová L 2019 The committee machine: computational to statistical gaps in learning a two-layers neural network *J. Stat. Mech.* **2019** 124023
- [19] Barbier J 2021 Overlap matrix concentration in optimal bayesian inference *Inf. Inference A* **10** 597–623
- [20] Loureiro B, Gerbelot C, Cui H, Goldt S, Krzakala F, Mezard M and Zdeborová L 2021 Learning curves of generic features maps for realistic datasets with a teacher–student model *Advances in Neural Information Processing Systems* vol 34 18137–51
- [21] Thrampoulidis C, Abbasi E and Hassibi B 2018 Precise error analysis of regularized m -estimators in high dimensions *IEEE Trans. Inf. Theory* **64** 5592–628
- [22] Loureiro B, Gerbelot C, Refinetti M, Sicuro G and Krzakala F 2022 Fluctuations, bias, variance & ensemble of learners: Exact asymptotics for convex losses in high-dimension (arXiv:2201.13383)
- [23] Bartlett P L, Long P M, Lugosi G and Tsigler A 2020 Benign overfitting in linear regression *Proc. Natl Acad. Sci.* **117** 30063–70
- [24] Goldt S, Mézard M, Krzakala F and Zdeborová L 2020 Modeling the influence of data structure on learning in neural networks: the hidden manifold model *Phys. Rev. X* **10** 041044
- [25] Jacot A, Simsek B, Spadaro E, Hongler C and Gabriel F 2020 Kernel alignment risk estimator: risk prediction from training data *Advances in Neural Information Processing Systems* vol 33 (New York: Curran Associates, Inc.) pp 15568–78
- [26] Bordelon B, Canatar A and Pehlevan C 2020 Spectrum dependent learning curves in kernel regression and wide neural networks *Int. Conf. on Machine Learning* (PMLR) pp 1024–34
- [27] Duda R, Hart P and Stork D 2012 *Pattern Classification* (New York: Wiley)
- [28] Viering T and Loog M 2021 The shape of learning curves: a review (arXiv:2103.10948)
- [29] Bayati M and Montanari A 2011 The LASSO risk for Gaussian matrices *IEEE Trans. Inf. Theory* **58** 1997–2017
- [30] Javanmard A and Montanari A 2013 State evolution for general approximate message passing algorithms, with applications to spatial coupling *Inf. Inference A* **2** 115–44
- [31] Gerbelot C and Berthier R 2021 Graph-based approximate message passing iterations (arXiv:2109.11905)
- [32] Nishimori H and Press O U 2001 *Statistical Physics of Spin Glasses and Information Processing: An Introduction (International Series of Monographs on Physics)* (Oxford: Oxford University Press)
- [33] Pedregosa F et al 2011 Scikit-learn: machine learning in Python *J. Mach. Learn. Res.* **12** 2825–30
- [34] Celentano M, Montanari A and Wu Y 2020 The estimation error of general first order methods *Conf. on Learning Theory* (PMLR) pp 1078–141
- [35] Donoho D L, Maleki A and Montanari A 2009 Message-passing algorithms for compressed sensing *Proc. Natl Acad. Sci.* **106** 18914–19
- [36] Rangan S 2011 Generalized approximate message passing for estimation with random linear mixing *2011 IEEE Int. Symp. on Information Theory Proc.* (IEEE) pp 2168–72
- [37] Aubin B 2020 Mean-field methods and algorithmic perspectives for high-dimensional machine learning *PhD dissertation* Université Paris-Saclay, 2020, thèse de doctorat dirigée par Zdeborová, Lenka Physique université Paris-Saclay
- [38] Vershynin R 2018 *High-Dimensional Probability: An Introduction With Applications in Data Science* vol 47 (Cambridge: Cambridge University Press)
- [39] Bauschke H H, Borwein J M and Combettes P L 2003 Bregman monotone optimization algorithms *SIAM J. Control Optim.* **42** 596–636
- [40] Bauschke H, Combettes P and Noll D 2006 Joint minimization with alternating bregman proximity operators *Pac. J. Optim.* **2**
- [41] Bayati M and Montanari A 2011 The dynamics of message passing on dense graphs, with applications to compressed sensing *IEEE Trans. Inf. Theory* **57** 764–85
- [42] Parikh N and Boyd S 2014 Proximal algorithms *Found. Trends Optim.* **1** 127–239
- [43] Bauschke H H and Combettes P L et al 2011 *Convex Analysis and Monotone Operator Theory in Hilbert Spaces* vol 408 (Berlin: Springer)
- [44] Zdeborová L and Krzakala F 2016 Statistical physics of inference: thresholds and algorithms *Adv. Phys.* **65** 453–552
- [45] Virtanen P et al (SciPy 1.0 Contributors) 2020 SciPy 1.0: fundamental algorithms for scientific computing in Python *Nat. Methods* **17** 261–72
- [46] Lam S K, Pitrou A and Seibert S 2015 Numba: a LLVM-based python JIT compiler *Proc. 2nd Workshop on the LLVM Compiler Infrastructure in HPC (LLVM 2015)* (New York: Association for Computing Machinery)