Learning Gaussian Mixtures with Generalised Linear Models: Precise Asymptotics in High-dimensions

Bruno Loureiro[1], Gabriele Sicuro[2], Cédric Gerbelot[3], Alessandro Pacco[1], Florent Krzakala[1], and Lenka Zdeborová[4]

[1]IdePHICS Lab. EPFL, Lausanne
[2]Department of Mathematics, King's College London
[3]Laboratoire de Physique de l'École Normale Supérieure, Université PSL, CNRS, Sorbonne Université, Université Paris-Diderot, Sorbonne Paris Cité
[4]SPOC, EPFL, Lausanne

June 8, 2021

Abstract

Generalised linear models for multi-class classification problems are one of the fundamental building blocks of modern machine learning tasks. In this manuscript, we characterise the learning of a mixture of K Gaussians with generic means and covariances via empirical risk minimisation (ERM) with any convex loss and regularisation. In particular, we prove exact asymptotics characterising the ERM estimator in high-dimensions, extending several previous results about Gaussian mixture classification in the literature. We exemplify our result in two tasks of interest in statistical learning: a) classification for a mixture with sparse means, where we study the efficiency of ℓ_1 penalty with respect to ℓ_2; b) max-margin multi-class classification, where we characterise the phase transition on the existence of the multi-class logistic maximum likelihood estimator for $K > 2$. Finally, we discuss how our theory can be applied beyond the scope of synthetic data, showing that in different cases Gaussian mixtures capture closely the learning curve of classification tasks in real data sets.

Contents

1 Introduction

A recurring observation in modern deep learning practice is that neural networks often defy the standard wisdom of classical statistical theory. For instance, deep neural networks typically achieve good generalisation performances at a regime in which it interpolates the data, a fact at odds with the intuitive bias-variance trade-off picture stemming from classical theory [1–3]. Surprisingly, many of the "exotic" behaviours encountered in deep neural networks have recently been shown to be shared by models as simple as overparametrised linear classifiers [4, 5], e.g., the aforementioned benign over-fitting [6]. Therefore, understanding the generalisation properties of simple models in high-dimensions has proven to be a fertile ground for elucidating some of the challenging statistical questions posed by modern machine learning practice [7–16].

In this manuscript, we pursue this enterprise in the context of a commonly used model for high-dimensional classification problems: the Gaussian mixture. Indeed, it has been recently argued that the features learned by deep neural networks trained on the cross-entropy loss "collapse" in a mixture of well-separated clusters, with the last layer acting as a simple linear classifier [17]. Another observation put forward in [18] is that data obtained using generative adversarial networks behave as Gaussian mixtures. Here, we derive an exact asymptotic formula characterising the performance of generalised linear classifiers trained on K Gaussian clusters with generic covariances and means. Our formula is valid for any convex loss and penalty, encompassing popular tasks in the machine learning literature such as ridge regression, basis pursuit, cross-entropy minimisation and max-margin estimation. This allow us to answer relevant questions for statistical learning, such as: what is the separability threshold for K-clustered data? How does regularisation affects estimation? Can different penalties help when the means are sparse? We

also extend the observation of [18] showing that the learning curves of binary classification tasks on *real data* are indeed well captured by our asymptotic analysis.

Model definition We consider learning from a d-dimensional mixture of K Gaussian clusters $\mathcal{C}_{k\in[K]}$. The data set is obtained by sampling n pairs $(\boldsymbol{x}^\nu, \boldsymbol{y}^\nu)_{\nu\in[n]} \in \mathbb{R}^{d+1}$ identically and independently. We adopt the one-hot encoded representation of the labels, i.e., if $\boldsymbol{x}^\nu \in \mathcal{C}_k$, then $\boldsymbol{y}^\nu = \boldsymbol{e}_k$, kth basis vector of $\mathbb{R}^K$. We will denote the matrix of concatenated samples $\boldsymbol{X} \in \mathbb{R}^{d\times n}$. The mixture density then reads:

$$P(\boldsymbol{x}, \boldsymbol{y}) = \sum_{k=1}^{K} y_k \rho_k \mathcal{N}\left(\boldsymbol{x} \,\middle|\, \boldsymbol{\mu}_k, \Sigma_k\right), \tag{1}$$

where $\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}, \Sigma)$ is the multivariate normal distribution with mean $\boldsymbol{\mu}$ and covariance matrix Σ. The matrix of concatenated means is denoted $\boldsymbol{M} \in \mathbb{R}^{d\times K}$. In Eq. (1), $\forall k$, $\rho_k = P(\boldsymbol{y} = \boldsymbol{e}_k) \geq 0$, $\boldsymbol{\mu}_k \in \mathbb{R}^d$ and $\Sigma_k \in \mathbb{R}^{d\times d}$ is positive-definite. We will consider the estimator obtained by minimising the following empirical risk:

$$\mathcal{R}(\boldsymbol{W}, \boldsymbol{b}) \equiv \sum_{\nu=1}^{n} \ell\left(\boldsymbol{y}^\nu, \frac{\boldsymbol{W}\boldsymbol{x}^\nu}{\sqrt{d}} + \boldsymbol{b}\right) + \lambda r(\boldsymbol{W}), \tag{2}$$

$$(\boldsymbol{W}^\star, \boldsymbol{b}^\star) \equiv \underset{\boldsymbol{W}\in\mathbb{R}^{K\times d},\, \boldsymbol{b}\in\mathbb{R}^K}{\operatorname{argmin}} \mathcal{R}(\boldsymbol{W}, \boldsymbol{b}), \tag{3}$$

where $\boldsymbol{W} \in \mathbb{R}^{K\times d}$ and $\boldsymbol{b} \in \mathbb{R}^K$ are the weights and bias to be learned, ℓ is a convex loss function, and r is a regularisation function whose strength is tuned by the parameter $\lambda \geq 0$. For example the loss function ℓ can represent the composition of a cross-entropy loss with a softmax thresholding on the linear part of Eq. (2). We will characterise the distribution of the estimator $(\boldsymbol{W}^\star, \boldsymbol{b}^\star)$, and we will evaluate the average training loss defined as

$$\epsilon_\ell = \frac{1}{n}\sum_{\nu=1}^{n} \ell\left(\boldsymbol{y}^\nu, \frac{\boldsymbol{W}^\star\boldsymbol{x}^\nu}{\sqrt{d}} + \boldsymbol{b}^\star\right), \tag{4}$$

as well as the average training error ϵ_t and generalisation error ϵ_g, defined as the misclassification rates:

$$\epsilon_t = \frac{1}{n}\sum_{\nu=1}^{n} \mathbb{I}\left[\boldsymbol{y}^\nu \neq \hat{\boldsymbol{y}}\left(\frac{\boldsymbol{W}^\star\boldsymbol{x}^\nu}{\sqrt{d}} + \boldsymbol{b}^\star\right)\right], \tag{5}$$

$$\epsilon_g = \mathbb{E}_{(\boldsymbol{x}^{\text{new}}, \boldsymbol{y}^{\text{new}})}\left[\mathbb{I}\left[\boldsymbol{y}^{\text{new}} \neq \hat{\boldsymbol{y}}\left(\frac{\boldsymbol{W}^\star\boldsymbol{x}^{\text{new}}}{\sqrt{d}} + \boldsymbol{b}^\star\right)\right]\right],$$

where $(\boldsymbol{x}^{\text{new}}, \boldsymbol{y}^{\text{new}})$ is a new unseen data point sampled from the distribution in Eq. (1). In the previous equations, we have used the function $\hat{\boldsymbol{y}}\colon \mathbb{R}^K \to \mathbb{R}^K$, so that $\hat{y}_k(\boldsymbol{x}) = \mathbb{I}(\max_\kappa x_\kappa = x_k)$.

The **main contributions** in this manuscript are the following:

(C1) In Sec. 2 and Appendix A we prove closed-form equations characterizing the asymptotic distribution of the matrix of weights $\boldsymbol{W}^\star \in \mathbb{R}^{K\times d}$, enabling the exact computation of key quantities such as the training and generalisation error. Our proof method solves shortcomings of previous approaches by introducing a novel approximate message-passing sequence, building on recent advances in this framework, that is of independent interest.

(C2) In Sec. 3.1 we study the problem of classifying an anisotropic mixture with sparse means, where the strong or weak directions in the data are correlated with the non-zero components of the mean as in [19]. We study how learning the sparsity with an ℓ_1 penalty improves the classification performance.

(C3) In Sec. 3.2 we study the performance of the cross-entropy estimator in the limit of vanishing regularisation $\lambda \to 0^+$ for K Gaussian clusters as a function of the sample complexity $\alpha = n/d$; we show that a phase transition takes place at a certain value $\alpha_K^\star$ between a regime of complete separability of the data and a regime in which the correct classification of almost all points in the data set is not possible. We also investigate the effect of $\lambda > 0$ regularisation on the generalisation error, comparing the $K > 2$ case with the results given in the literature for $K = 2$ [14, 20].

(C4) In Sec. 3.3 we investigate the applicability of our formula beyond the Gaussian assumption by applying it to classification tasks on *real data*. We show that for different tasks and losses, it closely captures the real learning curves, even when data is mapped through a non-linear feature map. This further shows that Gaussian mixtures are a good surrogate model for investigating real classification tasks, as put forward in [18].

Relation to previous work The analysis of Gaussian mixture models in the high-dimensional regime has been the subject of many recent works. Exact asymptotics has been derived for the binary classification case with diagonal covariances in [21–23] for the logistic loss and in [24, 25] for the square loss, both with ℓ_2 penalty. A similar analysis has been performed in [26] for the hard-margin SVM. These works were generalised to generic convex losses and ℓ_2 penalty in [14], where it has been also shown that the regularisation term can play an important role in reaching Bayes-optimal performances. Hinge regression with ℓ_1 penalty and diagonal covariance was treated in [13]. Recently, these asymptotic results were generalised to the case in which both clusters share the same covariance in [27], and finite rate bounds were given in [28, 29] in the case of sub-Gaussian mixtures. Asymptotic results for the multiclass problem with diagonal covariance were derived in [20] for the restricted case of the square loss with ℓ_2 penalty. Our result unifies all the aforementioned asymptotic formulas, and extends them to the general case of a multiclass problem with generic covariances and arbitrary convex losses and penalties.
From a technical standpoint, in [13, 14, 20, 21, 25, 27, 30] the authors use convex Gaussian comparison inequalities, see e.g. [31, 32], to prove their result. In particular, the proof given in [20] for the multiclass problem harnesses the geometry of least-squares, and it is then stressed that this method breaks down for multiclass problems in which the risk does not factorise over the K clusters (as for the cross-entropy, for example). We solve this problem using an innovative proof technique which has an interest in its own. Our approach is to capture the effect of non-linearity and generic covariances via the rigorous study of an approximate message-passing (AMP) sequence, a family of iterations that admit closed-form asymptotics at each step called *state evolution equations* [33]. Our proof relies on several refinements of AMP methods to handle the full complexity of the problem, notably spatial coupling with matrix valued variables [34–36] and non-separable update functions [37], via a multi-layer approach to AMP [38].
The sparse Gaussian mixture model analysed in Section 3.1 is closely related to the rare/weak features model introduced in [19] and widely studied in the context of sparse linear discriminant analysis [39–42]. It was recently revisited in [28, 29] in the context of ERM with max-margin classifiers. Here, we consider a correlated variation of the model and study the benefit of using a sparsity inducing ℓ_1 penalty.
The separability transition is a classical topic [43, 44] that has recently witnessed a renewal of interest thanks to its connection to overparametrization. It was studied in [16] in the context of uncorrelated Gaussian data, in [8] in the random features model and in [14, 21] for binary Gaussian mixtures.
Recently, [12, 45, 46] showed that the performance of different regression tasks on real data are well-captured by a teacher-student Gaussian model in high-dimensions for ridge regression, but this turned not to be true for non-linear problems such as logistic classification [12]. Authors of [18] showed instead that data from generative adversarial networks behave like Gaussian mixtures, motivating the modeling

of such mixture for real-data in the present paper.

2 Technical results

Our main technical result is an exact asymptotic characterization of the distribution of the estimator $\boldsymbol{W}^\star$. Informally, the estimator $\boldsymbol{W}^\star$ and the quantity $\boldsymbol{W}^\star \boldsymbol{X}/\sqrt{d}$ behave asymptotically as non-linear transforms of multivariate Gaussian distributions. These transforms are directly linked to the proximal operators [47, 48] associated to the loss and regularisation functions, summarizing the effect of the cost function landscape on the estimator. The parameters of these Gaussian distributions and proximals can then be computed from the fixed point of a self-contained set of equations. We start by presenting the most generic form of our result in a concentration of measure-like statement in Theorem 1, and discuss an intuitive interpretation of the different quantities involved. Theorem 2 then states how the training and generalisation errors can be computed. All results presented in the experiments section can be obtained from Theorem 1. In Corollary 3 we discuss a particular case where explicit simplifications can be obtained. But first, let's summarise the required assumptions for our result to hold.

(A1) The functions ℓ (as a function of its second argument) and r are proper, closed, lower semi-continuous convex functions.

(A2) The covariance matrices are positive definite and their spectral norms are bounded.

(A3) The mean vectors $\boldsymbol{\mu}_k$ are distributed according to some density $P_{\boldsymbol{\mu}}(\boldsymbol{M})$ such that the following quantity is finite

$$\forall d \qquad \mathbb{E}\left[\left\|\boldsymbol{M}^\top \boldsymbol{M}\right\|_{\mathrm{F}}\right] < +\infty, \tag{6}$$

where $\| \bullet \|_{\mathrm{F}}$ denotes the Frobenius norm.

(A4) The number of samples n and dimension d both go to infinity with fixed ratio $\alpha = n/d$, called hereafter the sample complexity. The number of clusters K is finite.

Before proceeding further, let us specify a useful notation. Suppose that the matrix $\boldsymbol{G} = (G_{ki})_{ki} \in \mathbb{R}^{K\times d}$ is given, alongside the four-index tensor $\mathbf{A} = (A_{kik'i'})_{kik'i'} \in \mathbb{R}^{K\times d} \otimes \mathbb{R}^{K\times d}$. We will use the notation $\boldsymbol{G} \odot \mathbf{A} = \sum_{ki} G_{ki} A_{kik'i'} \in \mathbb{R}^{K\times d}$. Similarly, given a four-index tensor $\mathbf{A}$, we will define $\sqrt{\mathbf{A}}$ as the tensor such that $\mathbf{A} = \sqrt{\mathbf{A}} \odot \sqrt{\mathbf{A}}$. We are now in a position to state our main result.

Theorem 1 (Concentration properties of the estimator)**.** *Let $\boldsymbol{\xi}_{k\in[K]} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_K)$ be collection of K-dimensional standard normal vectors independent of other quantities. Let also be $\{\Xi_k\}$ a set of K matrices, $\Xi_k \in \mathbb{R}^{K\times d}$, with i.i.d. standard normal entries, independent of other quantities. Under the set of assumptions (A1–A4), for any pseudo-Lispchitz functions of finite order $\phi_1 : \mathbb{R}^{K\times d} \to \mathbb{R}$, $\phi_2 : \mathbb{R}^{K\times n} \to \mathbb{R}$, the estimator $\boldsymbol{W}^\star$ and the matrix $\boldsymbol{Z}^\star = \frac{1}{\sqrt{d}} \boldsymbol{W}^\star \boldsymbol{X}$ verify:*

$$\phi_1(\boldsymbol{W}^\star) \xrightarrow[n,d\to+\infty]{P} \mathbb{E}_{\Xi}\left[\phi_1(\boldsymbol{G})\right], \qquad \phi_2(\boldsymbol{Z}^\star) \xrightarrow[n,d\to+\infty]{P} \mathbb{E}_{\boldsymbol{\xi}}\left[\phi_2(\boldsymbol{H})\right], \tag{7}$$

where we have introduced the proximal for the loss:

$$\boldsymbol{h}_k = \boldsymbol{V}_k^{1/2} \operatorname{Prox}_{\ell(\boldsymbol{e}_k, \boldsymbol{V}_k^{1/2}\bullet)}(\boldsymbol{V}_k^{-1/2}\boldsymbol{\omega}_k) \in \mathbb{R}^K, \qquad \boldsymbol{\omega}_k \equiv \boldsymbol{m}_k + \boldsymbol{b} + \boldsymbol{Q}_k^{1/2}\boldsymbol{\xi}_k, \tag{8}$$

and $\boldsymbol{H} \in \mathbb{R}^{K\times n}$ *is obtained by concatenating each* $\boldsymbol{h}_k$, $\rho_k n$ *times. We have also introduced the matrix proximal* $\boldsymbol{G} \in \mathbb{R}^{K\times d}$*:*

$$\boldsymbol{G} = \mathbf{A}^{\frac{1}{2}} \odot \mathrm{Prox}_{r(\mathbf{A}^{\frac{1}{2}}\odot\bullet)}(\mathbf{A}^{\frac{1}{2}} \odot \boldsymbol{B}), \qquad \mathbf{A}^{-1} \equiv \sum_k \hat{\boldsymbol{V}}_k \otimes \Sigma_k, \qquad \boldsymbol{B} \equiv \sum_k \left(\boldsymbol{\mu}_k \hat{\boldsymbol{m}}_k^\top + \Xi_k \odot \sqrt{\hat{\boldsymbol{Q}}_k \otimes \Sigma_k}\right). \quad (9)$$

The collection of parameters $(\boldsymbol{Q}_k, \boldsymbol{m}_k, \boldsymbol{V}_k, \hat{\boldsymbol{Q}}_k, \hat{\boldsymbol{m}}_k, \hat{\boldsymbol{V}}_k)_{k\in[K]}$ *is given by the fixed point of the following self-consistent equations:*

$$\begin{cases} \boldsymbol{Q}_k = \frac{1}{d}\mathbb{E}_\Xi[\boldsymbol{G}\Sigma_k\boldsymbol{G}^\top] \\ \boldsymbol{m}_k = \frac{1}{\sqrt{d}}\mathbb{E}_\Xi[\boldsymbol{G}\boldsymbol{\mu}_k] \\ \boldsymbol{V}_k = \frac{1}{d}\mathbb{E}_\Xi\left[\left(\boldsymbol{G} \odot \left(\hat{\boldsymbol{Q}}_k \otimes \Sigma_k\right)^{-\frac{1}{2}} \odot (\boldsymbol{I}_K \otimes \Sigma_k)\right)\Xi_k^\top\right] \end{cases} \qquad \begin{cases} \hat{\boldsymbol{Q}}_k = \alpha\rho_k \mathbb{E}_{\boldsymbol{\xi}}\left[\boldsymbol{f}_k \boldsymbol{f}_k^\top\right] \\ \hat{\boldsymbol{V}}_k = -\alpha\rho_k \boldsymbol{Q}_k^{-\frac{1}{2}}\mathbb{E}_{\boldsymbol{\xi}}\left[\boldsymbol{f}_k \boldsymbol{\xi}^\top\right] \\ \hat{\boldsymbol{m}}_k = \alpha\rho_k \mathbb{E}_{\boldsymbol{\xi}}\left[\boldsymbol{f}_k\right] \end{cases} \quad (10)$$

where $\boldsymbol{f}_k \equiv \boldsymbol{V}_k^{-1}(\boldsymbol{h}_k - \boldsymbol{\omega}_k)$*, and the vector* $\boldsymbol{b}^\star$ *is such that* $\sum_k \rho_k \mathbb{E}_{\boldsymbol{\xi}}\left[\boldsymbol{V}_k \boldsymbol{f}_k\right] = \mathbf{0}$ *holds.*

The purpose of this statement is to have an asymptotically exact description of the distribution of the estimator, where the dimensions going to infinity are effectively summarized as averages over simple, independent distributions. Those distributions are parametrised by the set of finite-size parameters $(\boldsymbol{Q}_k, \boldsymbol{m}_k, \boldsymbol{V}_k, \hat{\boldsymbol{Q}}_k, \hat{\boldsymbol{m}}_k, \hat{\boldsymbol{V}}_k)_{k\in[K]}$ that can be exactly evaluated and have a clear interpretation. Indeed, the parameters $(\boldsymbol{m}_k, \hat{\boldsymbol{m}}_k)$ and $(\boldsymbol{Q}_k, \hat{\boldsymbol{Q}}_k)$ respectively represent means and covariances of multivariate Gaussians (combined with the original $\boldsymbol{\mu}_k, \Sigma_k$), and the $(\boldsymbol{V}_k, \hat{\boldsymbol{V}}_k)$ parametrise the deformations that should be applied to these Gaussians to obtain the distribution of $\boldsymbol{W}^\star, \boldsymbol{Z}^\star$. The distribution is characterized in a weak sense with concentration of pseudo-Lipschitz (i.e. sufficiently regular) functions, whose definition is reminded in the Appendix A. From this result one can work out a number of properties of the weights $\boldsymbol{W}^\star$, e.g., training and generalisation error, but also hypothesis tests as done in [49] for the LASSO. Due to the generality of the statement, no direct simplification is possible. However, we will see that in certain specific cases all quantities can be greatly simplified. This is notably the case for diagonal covariance matrices and separable estimators and observables ϕ_1, ϕ_2, where the sums over high-dimensional Gaussians concentrate explicitly to one-dimensional expectations. For instance the results of [14, 20] can be recovered as special cases of the present work. Theorem 1 then allows to obtain the asymptotic values of the generalisation error, of the training loss and of the training error. Their explicit expression is given in the following Theorem.

Theorem 2 (generalisation error and training loss)**.** *In the hypotheses of Theorem 1, the training loss, the training error and the generalisation error are given by*

$$\epsilon_\ell = \sum_{k=1}^K \rho_k \mathbb{E}_{\boldsymbol{\xi}}[\ell(\boldsymbol{e}_k, \boldsymbol{h}_k)] \quad (11)$$

$$\epsilon_t = 1 - \sum_{k=1}^K \rho_k \mathbb{E}_{\boldsymbol{\xi}}\left[\hat{y}_k(\boldsymbol{h}_k)\right], \quad (12)$$

$$\epsilon_g = 1 - \sum_{k=1}^K \rho_k \mathbb{E}_{\boldsymbol{\xi}}\left[\hat{y}_k(\boldsymbol{\omega}_k)\right]. \quad (13)$$

The case of ridge regularisation and diagonal Σ_k The general formulas given above can be remarkably simplified under some assumptions about the choice of the regularisation and about the structure of the covariance matrices Σ_k. This is the case for instance for the ridge regularisation $r(\boldsymbol{W}) = \|\boldsymbol{W}\|_{\mathrm{F}}^2/2$ and jointly diagonalizable covariances. In this case, Theorem 1 simplifies as follows.

Corollary 3. *Under the hypotheses of Theorem 1, let us further assume that a ridge regularisation is adopted, $r(\boldsymbol{W}) = \|\boldsymbol{W}\|_{\mathrm{F}}^2/2$, and that the covariance matrices Σ_k have a common set of orthonormal eigenvectors $\{\boldsymbol{v}_i\}_{i=1}^d$, so that, for each $\Sigma_k = \sum_{i=1}^d \sigma_i^k \boldsymbol{v}_i \boldsymbol{v}_i^\top$. Let us also introduce, in the $d \to +\infty$ limit, the joint distribution for the K-dimensional vectors $\boldsymbol{\sigma} = (\sigma^1, \dots, \sigma^K)$ and $\boldsymbol{\mu} = (\mu^1, \dots, \mu^K)$,*

$$\frac{1}{d}\sum_{i=1}^{d}\prod_{k=1}^{K}\delta(\sigma^k - \sigma_i^k)\delta(\mu^k - \sqrt{d}\boldsymbol{\mu}_k^\top \boldsymbol{v}_i) \xrightarrow{d\to+\infty} p(\boldsymbol{\sigma}, \boldsymbol{\mu}), \tag{14}$$

Then, the first three saddle point equations in eqs. (10) take the form

$$\begin{cases} \boldsymbol{Q}_k = \mathbb{E}_{\boldsymbol{\sigma},\boldsymbol{\mu}}\left[\sigma^k\left(\lambda \boldsymbol{I}_K + \sum_{\kappa=1}^{K}\sigma^\kappa \hat{\boldsymbol{V}}_k\right)^{-2}\left(\sum_{\kappa\kappa'}\mu^\kappa\mu^{\kappa'}\hat{\boldsymbol{m}}_\kappa\hat{\boldsymbol{m}}_{\kappa'}^\top + \sum_{\kappa=1}^{K}\sigma^\kappa\hat{\boldsymbol{Q}}_k\right)\right], \\ \boldsymbol{m}_k = \mathbb{E}_{\boldsymbol{\sigma},\boldsymbol{\mu}}\left[\mu^k\left(\lambda \boldsymbol{I}_K + \sum_{\kappa=1}^{K}\sigma^\kappa \hat{\boldsymbol{V}}_k\right)^{-1}\sum_{\kappa=1}^{K}\mu^\kappa\hat{\boldsymbol{m}}_\kappa\right], \\ \boldsymbol{V}_k = \mathbb{E}_{\boldsymbol{\sigma},\boldsymbol{\mu}}\left[\sigma^k\left(\lambda \boldsymbol{I}_K + \sum_{\kappa=1}^{K}\sigma^\kappa \hat{\boldsymbol{V}}_k\right)^{-1}\right]. \end{cases} \tag{15}$$

Narrative of the proof The proof is detailed in Appendix A. It overcomes problems that existing methods, notably convex Gaussian comparison inequalities [20], have yet to be adapted to. The first main technical difficulty resides in the estimator of interest being a matrix learned with non-linear functions. This makes it impossible to decompose the problem on each row of the estimator, which must be characterized in a probabilistic sense directly as a matrix. The second main difficulty is brought by the mixture of arbitrary covariances. Intuitively, the covariances correlate the estimator with the individual clusters, and therefore the correlation function cannot be represented by a single quantity. In our proof, these points are handled using the AMP and related state-evolution techniques [33, 50–52]. The main idea of the proof is to express the estimator $\boldsymbol{W}^\star$ as the limit of a convergent sequence whose structure enables the decomposition of all correlations and distributions in closed form. AMP iterations can handle matrix valued variables [36, 53], correlations in block-structure [36], non-separable functions [37, 38] and compositions of the previous three, leaving a large choice of possibilities in their design. We thus reformulate the problem in a way that makes the interaction between the estimator and each cluster explicit, effectively introducing a block structure to the problem, and isolate the overlaps with the means $\{\boldsymbol{\mu}_k\}$. We then design a matrix-valued sequence that obeys the update rule of an AMP sequence, in order to benefit from its exact asymptotics, and whose fixed point condition matches the optimality condition of the ERM problem, Eq. (2). Our proof builds on the spatial coupling framework in the AMP literature [36, 54], which shows that the effect of random matrices defined with non-identically distributed blocks can be embedded in an AMP iteration while explicitly keeping the effect of each block. The non-linearities are then obtained by a block decomposition of the proximal operators defined on sets of matrices, acting on different variables of the AMP sequence and representing the effect of each cluster. The convergence analysis is made possible by the convexity of the problem: the sequence is defined with proximal operators of convex functions which are roughly contractions, and results in converging sequences when combined with the high-dimensional properties of the iteration. It is also interesting to note that the replica method, although heuristic, yet again gives the correct prediction without any hindering from the aforementioned main difficulties, as detailed in Appendix B.

Universality AMP-type proofs are amenable to both finite sample size analysis and universality proofs. For instance, in [55] it is shown that simpler instances of AMP for the LASSO exhibit exponential con-

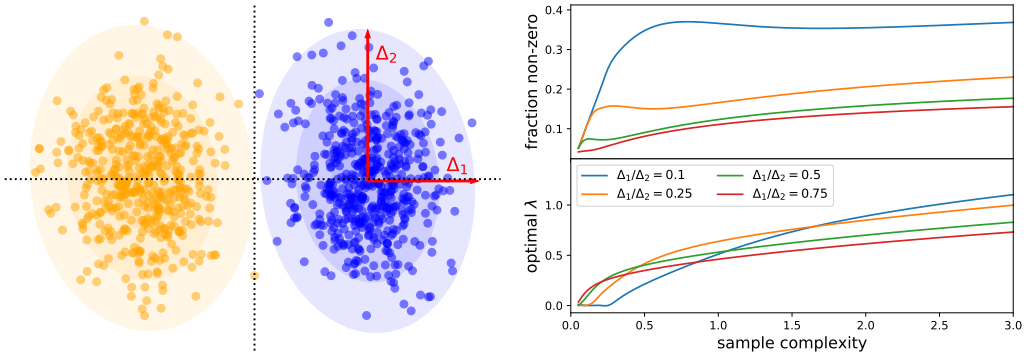

Figure 1: (**Left**) Two-dimensional projection of the Gaussian mixture introduced via Eq. (16) in which the sparse directions of the means are correlated with the weak/strong directions in the data. (**Right**) Fraction of non-zero elements of the lasso estimator (*top*) and optimal regularisation strength (*bottom*) as a function of the sample complexity $\alpha = n/d$ for different anisotropy ratios and fixed sparsity $\rho = 0.1$. Note that for $\Delta_1/\Delta_2 \lesssim 1$ and for low α the optimal error is achieved for vanishing regularisation, which corresponds to the *basis pursuit* algorithm [58].

centration in the system size, and the i.i.d. Gaussian assumption can be relaxed to independently sampled sub-Gaussian distributions, as shown in [56, 57]. Although these results do not formally encompass our case, their proof method contains most of the required technicalities, and it should be possible to prove similar results in the present setting. Indeed, recent results in [18] suggest that the formula of Theorem 1 and 2 should be universal for all mixtures of concentrated distribution in high-dimension, not only Gaussian ones. As we discuss Sec. 3.3, even real data learning curves are empirically found to follow the behavior of the mixture of Gaussians.

3 Results on synthetic and real datasets

In this section we exemplify how Theorem 1 can be employed to compute quantities of interest in different empirical risk minimisation tasks in high-dimensions.

3.1 Correlated sparse mixtures

As a first example, consider a binary classification problem in which the most relevant features live in a subspace of $\mathbb{R}^d$, and can be either weaker or stronger with respect to the irrelevant features. This problem can be modelled with a Gaussian mixture model with sparse means, and where the strong/weak directions of the covariance matrix are correlated with the non-zero components of the means. Mathematically, we consider a data set with n independent samples $(\boldsymbol{x}^\nu, y^\nu) \in \mathbb{R}^d \times \{-1, 1\}$ drawn from a Gaussian mixture $\boldsymbol{x}^\nu \sim \mathcal{N}(y^\nu \boldsymbol{\mu}, \Sigma)$ with diagonal covariance $\Sigma_{ij} = \sigma_i \delta_{ij}$ which is correlated with the sparse means:

$$P(\boldsymbol{\mu}, \boldsymbol{\sigma}) = \prod_{i=1}^{d} \left\{ \rho \mathcal{N}(\mu_i | 0, 1) \delta_{\sigma_i, \Delta_1} + (1 - \rho) \delta_{\mu_i} \delta_{\sigma_i, \Delta_2} \right\} \tag{16}$$

where $\rho > 0$ is the fraction of non-zero entries in $\boldsymbol{\mu}$. This model is closely related to the rare/weak features model introduced by Donoho and Jin in [19]. Indeed, in the case $\Delta_1 = \Delta_2 \equiv \Delta$ the signal-to-noise ratio of

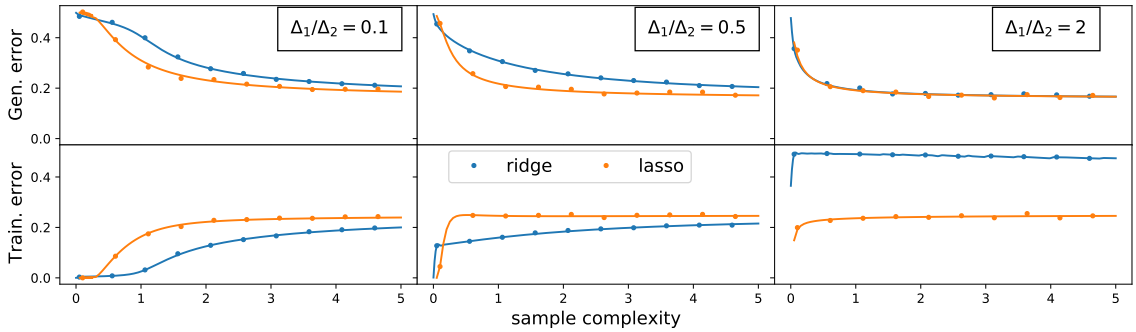

Figure 2: Learning curves for the sparse mixture model defined via Eq. (16) at fixed sparsity $\rho = 0.1$, comparing the performance of the ridge (blue) and the lasso (orange) estimators at optimal regularisation strength λ^* and for different anisotropy ratio Δ_1/Δ_2 (here $\Delta_1 = 0.1$ and we vary Δ_2). Full lines denote the theoretical prediction, and dots denote finite instance simulations with $d = 1000$ using the `ElasticNet` module in the `Scikit-learn` package [59]. Above a certain sample complexity α, we can identify two regimes: a) a $\Delta_1/\Delta_2 \lesssim 1$ regime in which the ℓ_1 penalty improves significantly over ℓ_2; b) a $\Delta_1/\Delta_2 \gtrsim 1$ regime in which the performance is similar. Interestingly, even though the generalisation error of lasso is considerably better in a), the training loss (i.e. the mse on the labels) is higher, & vice-versa in b).

the model is proportional to $\rho/\sqrt{\Delta}$, with ρ and $\Delta^{-1/2}$ playing the roles of the parameters ϵ and μ_0 setting the "rareness" and "strength" of the features in [19].

The formulas given in Theorem 1 simplify considerably for this model (see Appendix C for details), and therefore can be readily used to characterise the learning performance of different losses and penalties. For instance, one fundamental question we can address is when learning a sparse solution with the ℓ_1 regularization is advantageous over the usual ℓ_2. Figure 2 compares the learning curves computed from Theorem 1 for the lasso and ridge estimators, with optimal regularisation strength $\lambda^\star(\alpha) = \operatorname{argmin} \epsilon_g(\alpha, \lambda)$ at fixed sparsity $\rho = 0.1$. We can see that lasso performs considerably better than ridge in the regime where $\Delta_1/\Delta_2 \lesssim 1$, while it achieves a similar performance when $\Delta_1/\Delta_2 \gtrsim 1$. This is quite intuitive: the sparse directions are uninformative, and therefore learning the relevant features is better when they are stronger. Figure 1 (right) shows how the sparsity of the learned estimator $\boldsymbol{W}^\star$ and the optimal regularisation $\lambda^\star$ depends on the sample complexity $\alpha = n/d$. Interestingly, for $\Delta_1/\Delta_2 = 0.1$ or lower there is a region of small α in which basis pursuit ($\lambda = 0^+$) [58] is optimal, and the sparsity of the estimator has a curious non-monotonic behaviour with α.

3.2 Separability transition for the cross-entropy loss

We now consider the problem of classifying points of K Gaussian clusters using a cross-entropy loss

$$\ell(\boldsymbol{y}, \boldsymbol{x}) = -\sum_{k=1}^{K} y_k \ln \frac{e^{x_k}}{\sum_{\kappa=1}^{K} e^{x_\kappa}}. \tag{17}$$

Using the results of Theorem 2, we estimate the dependence of the generalisation error ϵ_g on the sample complexity α and on the regularisation λ. We assume Gaussian means $\boldsymbol{\mu}_k \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_d/d)$ and diagonal covariances $\Sigma_k \equiv \Sigma = \Delta \boldsymbol{I}_d$. Finally, we adopt a ridge penalty, $r(\boldsymbol{W}) \equiv \|\boldsymbol{W}\|_{\mathrm{F}}^2/2$, and we focus on the case of balanced clusters, i.e., $\rho_k = 1/K$ for the sake of simplicity.

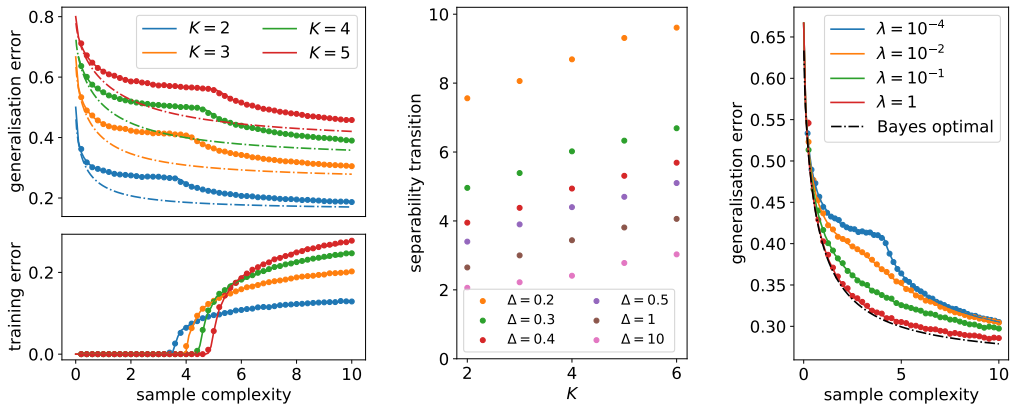

Figure 3: Classification of K Gaussian clusters in d dimensions, having Gaussian means and $\Sigma_k \equiv \Sigma = \Delta \boldsymbol{I}_d$ with $\Delta = 1/2$. In all presented cases, a quadratic regularisation has been adopted. Numerical experiments have been performed using $d = 10^3$. (**Left**) Generalisation error ϵ_g (*top*) and training error ϵ_t (*bottom*) as function of α at $\lambda = 10^{-4}$. Theoretical predictions (full lines) are compared with the results of numerical experiments (dots). Dash-dotted lines of the corresponding color represent, for comparison, the Bayes-optimal error. The results of numerical experiments are in agreement with the theoretical predictions in all cases. (**Center**) Separability transition $\alpha_K^\star$ as a function of K in the same setting for different values of Δ. (**Right**) Dependence of the generalisation error on the regularization λ for $K = 3$ and $\Delta = 1/2$ in the balanced case, $\rho_k = 1/K$.

Separability transition In Fig. 3 (left top) we plot the generalisation error ϵ_g as function of α for $2 \leq K \leq 5$ and $\lambda = 10^{-4}$. The smooth curve is obtained solving the fixed point equations in Theorem 1 and plugging the results in the formulas in Theorem 2. The results of numerical experiments are obtained averaging over 10^2 instances of the problem solved using the `LogisticRegression` module in the `Scikit-learn` package [59]. An excellent agreement is observed. For each pair (K, Δ) and for vanishing regularisation $\lambda \to 0^+$ we observe a double-descent-like behaviour in the generalisation error. Indeed, the cusp $\alpha_K^\star(\Delta)$ in the generalisation error corresponds to the point in which the cross-entropy estimator ceases to perfectly interpolate the data, revealing the existence of a separability transition of the type discussed in [16] for Gaussian i.i.d. data. As stressed therein, a phase of perfect separability of the data points corresponds to a regime in which the maximum-likelihood estimate does not exist with probability one. This is visible, in the same figure (left bottom), from the training error ϵ_t that is identically zero for $\alpha < \alpha_K^\star$, and strictly positive otherwise. Our result extends the observations in [14, 21], where an

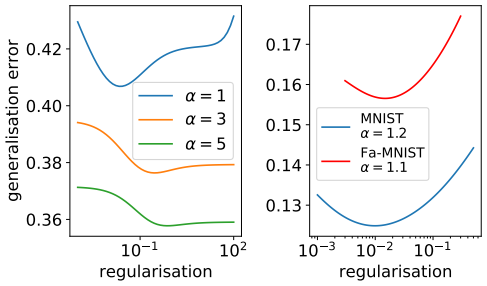

Figure 4: (**Left.**) Generalisation error obtained using ridge regression in the case of two balanced Gaussian clusters having $\Sigma_1 = \frac{1}{10}\boldsymbol{I}_d$ and $\Sigma_2 = \frac{1}{100}\boldsymbol{I}_d$ as function of λ for different values of the sample complexity α. (**Right**) Generalisation error ϵ_g as a function of λ at fixed α in the binary classification of MNIST and in the FashionMNIST via logistic regression (see Sec. 3.3 for details).

analytic expression for $\alpha_2^\star$ has been given in the case of for $K = 2$, $\boldsymbol{\mu}_1 = -\boldsymbol{\mu}_2$ Gaussian vector, generalising the classical result of Cover [43]. The separability transition point $\alpha_K^\star$ decreases with Δ and increases with K, showing that for larger K it is easier to separate the different clusters: this intuitively follows from the fact that, at fixed α and Δ, each cluster is given by $\alpha d/K$ points, i.e., fewer for increasing K and therefore easier to classify, see Fig. 3 (center).

The role of regularisation In Fig. 3 (right) we compare the performances of the cross-entropy loss with respect to the Bayes-optimal error (detailed in Appendix D) for different strength λ of the regularisation, assuming all identical diagonal covariances $\Sigma_k \equiv \Sigma = \Delta \boldsymbol{I}_d$. In the case of balanced clusters (i.e., $\rho_k = 1/K$ for all k) it is observed that the generalisation error approaches the Bayes-optimal error for $\lambda \to +\infty$. The same phenomenology has been observed in [14, 24] in the $K = 2$ case with opposite means and generic loss, and in [20] for $K > 2$ for the square loss. Using the concentration results of Section 2, we investigated the robustness of this result in the case of balanced clusters but with different covariances and various losses. First, we considered two opposite *balanced* clusters with $\Sigma_1 = \Delta_1 \boldsymbol{I}_d$ and $\Sigma_2 = \Delta_2 \boldsymbol{I}_2$, $\Delta_1 \neq \Delta_2$, and we estimated the generalisation error at fixed sample complexity as function of $\lambda \in [10^{-4}, 10^2]$ using ridge regression. As shown in Fig. 4 (left), the regularisation strength optimising the error is finite, and in particular depends on the sample complexity. This situation is closer to what is observed in real problems with balanced data analysed using logistic regression. Indeed, using the covariances from real data sets such as MNIST or Fashion-MNIST yields a similar behaviour, see Fig. 4 (right), with an optimal λ that is found to be finite.

3.3 Binary classification with real data

A recent line of work has reported that the asymptotic learning curves of simple regression tasks on real data sets can be well approximated by a surrogate Gaussian model matching the first two moments of the data [12, 45, 46]. However, this analysis was fundamentally restricted to least-squares regression, and considerable deviation from the Gaussian model was observed for classification tasks [12]. Authors of [18] have shown that realistic-looking data from trained generative adversarial networks behave like Gaussian mixtures. Here, we pursue these observations and investigate whether Theorem 2 can be used to capture the learning curves of classification tasks on two popular data sets: MNIST [60] and Fashion-MNIST [61]. Our goal is to compare the performances of some classification tasks on them with the predictions provided by the theory for the Gaussian mixture model.

Both data sets consist of $n_{\text{tot}} = 7 \times 10^4$ images $\hat{\boldsymbol{x}}^\mu \in \mathbb{R}^d$, $d = 784$. Each image $\hat{\boldsymbol{x}}^\mu$ is associated to a label $\hat{y}^\mu = \{0, 1, \dots, 9\}$ specifying the type of represented digit (in the case of MNIST) or item (in the case of Fashion-MNIST). In both cases, we divided the database into two balanced classes (even vs odd digits for MNIST, clothes vs accessories for Fashion-MNIST), relabelling the elements $\hat{\boldsymbol{x}}^\mu$ with $y^\mu \in \{-1, 1\}$ depending on their class, and we selected $n < n_{\text{tot}}$ elements to perform the training, leaving the others for the test of the performances. We adopted a logistic loss with ℓ_2 regularisation. First, we performed logistic regression on the training real data set, then we tested the learned estimators on the remaining $n_{\text{tot}} - n$ images. At the same time, for each class k of the training set, we empirically estimated the corresponding mean $\boldsymbol{\mu}_k \in \mathbb{R}^d$ and covariance matrix $\Sigma_k \in \mathbb{R}^{d \times d}$. We then assumed that the classification problem on the real database corresponds to a Gaussian mixture model of $K = 2$ clusters with means $\{\boldsymbol{\mu}_k\}_{k \in [2]}$ and covariances $\{\Sigma_k\}_{k \in [2]}$. Under this assumption, we computed the generalisation error and the training loss predicted by the theory inserting the empirical means and covariances in our general formulas. The results are given in Fig. 5, showing a good agreement between the theoretical prediction and the results obtained on MNIST and Fashion-MNIST. In Fig. 5 we also plot, as reference, the results of a classification

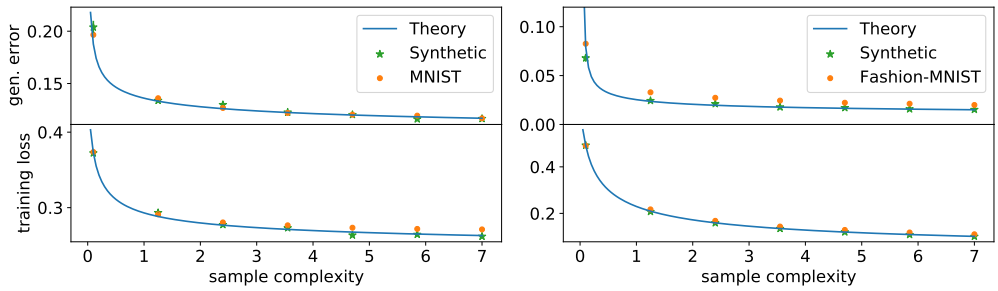

Figure 5: Generalisation error and training loss for the binary classification using the logistic loss on MNIST with $\lambda = 0.05$ (**left**) and on Fashion-MNIST with $\lambda = 1$ (**right**). The results are compared with synthetic data produced from the corresponding Gaussian mixture, and the theoretical prediction.

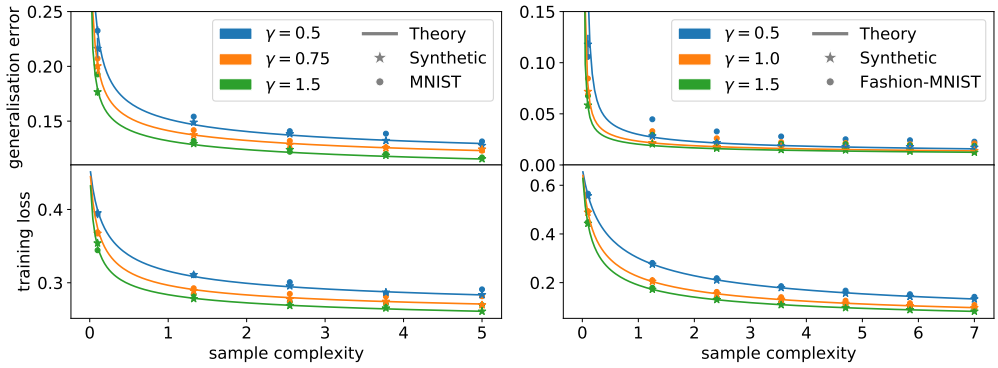

Figure 6: Generalisation error and training loss for the binary classification using the logistic on MNIST at $\lambda = 0.05$ (**left**) and on Fashion-MNIST at $\lambda = 1$ (**right**) in the random feature setting, for different values of γ, ratio between the number of parameters and the dimensionality of the data. The results are compared with synthetic data produced with the same γ, and the theoretical prediction.

task performed on synthetic data, obtained generating a genuine Gaussian mixture with the means and covariances of the real data set.

Interestingly, this construction can also be used to analyse the learning curves of classification problems with non-linear feature maps [12], e.g. random features [62]. In this case, we first apply to our data set a feature map $\boldsymbol{x}^\mu = \mathrm{erf}(\boldsymbol{F}\hat{\boldsymbol{x}}^\mu)$, where $\boldsymbol{F} \in \mathbb{R}^{p\times d}$ has i.i.d. Gaussian entries and the erf function is applied component wise. The classification task is then performed on the new data set $\{(\boldsymbol{x}^\nu, y^\nu)\}_{\nu\in[n]}$, the new data points $\boldsymbol{x}^\nu$ living in a p-dimensional space. We denote $\gamma = p/d$. We repeat the analysis described above in this new setting. Our results are in Fig. 6 for different values of γ. Once again, the generalisation error and the training loss are shown to be in a good agreement with both the theoretical prediction and the synthetic data sets obtained plugging in our formulas the real data means and the real data covariance matrices.

Acknowledgements

We thank Raphaël Berthier and Francesca Mignacco for discussions. We acknowledge funding from the ERC under the European Union's Horizon 2020 Research and Innovation Program Grant Agreement 714608-SMiLe, and from the French National Research Agency grants ANR-17-CE23-0023-01 PAIL.

Appendix

A Proof

This appendix presents the proof of the main technical result, Theorem 1. Throughout the whole proof, we assume that the set of conditions from Sec. 2 is verified.

A.1 Required background

In this Section, we give an overview of the main concepts and tools on approximate message passing algorithms which will be required for the proof.

We start with some definitions that commonly appear in the approximate message-passing literature, see e.g. [33,36,37]. 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 $n, m \in \mathbb{N}^*$, a function $\boldsymbol{\phi}\colon \mathbb{R}^{n\times K} \to \mathbb{R}^{m\times K}$ is called a* pseudo-Lipschitz of order k *if there exists a constant $L(k,K)$ such that for any $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^{n\times K}$,*

$$\frac{\|\boldsymbol{\phi}(\boldsymbol{x}) - \boldsymbol{\phi}(\boldsymbol{y})\|_{\mathrm{F}}}{\sqrt{m}} \leqslant L(k,K)\left(1 + \left(\frac{\|\boldsymbol{x}\|_{\mathrm{F}}}{\sqrt{n}}\right)^{k-1} + \left(\frac{\|\boldsymbol{y}\|_{\mathrm{F}}}{\sqrt{n}}\right)^{k-1}\right)\frac{\|\boldsymbol{x}-\boldsymbol{y}\|_{\mathrm{F}}}{\sqrt{n}} \tag{18}$$

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

For example, the function $f : \mathbb{R}^n \to \mathbb{R}, \boldsymbol{x} \mapsto \frac{1}{n}\|\boldsymbol{x}\|_2^2$ is pseudo-Lipshitz of order 2.

Moreau envelopes and proximal operators — In our proof, we will also frequently use the notions of Moreau envelopes and proximal operators, see e.g. [47,48]. These elements of convex analysis are often encountered in recent works on high-dimensional asymptotics of convex problems, and a detailed analysis of their properties can be found for example in [12,31]. For the sake of brevity, we will only sketch the main properties of such mathematical objects, referring to the cited literature for further details. In this proof, we will mainly use proximal operators acting on sets of real matrices endowed with their canonical scalar product. Furthermore, proximals will be defined with matrix valued parameters in the following way: for a given convex function $f\colon \mathbb{R}^{d\times K} \to \mathbb{R}$, a given matrix $\boldsymbol{X} \in \mathbb{R}^{d\times K}$ and a given symmetric positive definite matrix $\boldsymbol{V} \in \mathbb{R}^{K\times K}$ with bounded spectral norm, we will consider operators of the type

$$\underset{\boldsymbol{T}\in\mathbb{R}^{d\times K}}{\operatorname{argmin}}\left\{f(\boldsymbol{T}) + \frac{1}{2}\operatorname{tr}\left((\boldsymbol{T}-\boldsymbol{X})\boldsymbol{V}^{-1}(\boldsymbol{T}-\boldsymbol{X})^\top\right)\right\} \tag{19}$$

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

$$\mathrm{Prox}_{f(\bullet \boldsymbol{V}^{1/2})}(\boldsymbol{X}\boldsymbol{V}^{-1/2})\boldsymbol{V}^{1/2} \in \mathbb{R}^{d\times K} \tag{20}$$

or as a Bregman proximal operator [63] defined with the Bregman distance induced by the strictly convex, coercive function

$$\boldsymbol{X} \mapsto \frac{1}{2}\mathrm{tr}(\boldsymbol{X}\boldsymbol{V}^{-1}\boldsymbol{X}^\top) \tag{21}$$

which justifies the use of the Bregman resolvent

$$\underset{\boldsymbol{T}\in\mathbb{R}^{d\times K}}{\mathrm{argmin}} \left\{ f(\boldsymbol{T}) + \frac{1}{2}\mathrm{tr}\left((\boldsymbol{T}-\boldsymbol{X})\boldsymbol{V}^{-1}(\boldsymbol{T}-\boldsymbol{X})^\top\right)\right\} = (\mathrm{Id} + \partial f(\bullet)\boldsymbol{V})^{-1}(\boldsymbol{X}) \tag{22}$$

All the usual properties of standard proximal operators (i.e. firm non-expansiveness, link with Moreau/Bregman envelopes,…) hold for Bregman proximal operators defined with the distance (20), see e.g. [63, 64], justifying their use without any additional proof.

Gaussian concentration — Gaussian concentration properties are at the root of this proof. Such properties are reviewed in great detail, for example, in [12, 37]. We refer the interested reader to this set of works for a detailed and complete discussion.

Notations — For any set of matrices $\{\boldsymbol{A}_k \in \mathbb{R}^{n_k\times d_k}\}_{k\in[K]}$ we will use the following notation:

$$\begin{bmatrix} \boldsymbol{A}_1 & & & \\ & \boldsymbol{A}_2 & (*) & \\ & (*) & \ddots & \\ & & & \boldsymbol{A}_K \end{bmatrix} \equiv [\boldsymbol{A}_k]_{k=1}^K \in \mathbb{R}^{(\sum_{k=1}^K n_k)\times(\sum_{k=1}^K d_k)} \tag{23}$$

where the terms denoted by $(*)$ will be zero most of the time.
For a given function $\boldsymbol{\phi}\colon \mathbb{R}^{d\times K} \to \mathbb{R}^{d\times K}$, we write :

$$\boldsymbol{\phi}(\boldsymbol{X}) = \begin{bmatrix} \boldsymbol{\phi}^1(\boldsymbol{X}) \\ \vdots \\ \boldsymbol{\phi}^d(\boldsymbol{X}) \end{bmatrix} \in \mathbb{R}^{d\times K} \tag{24}$$

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

$$\frac{\partial \boldsymbol{\phi}^i}{\partial \boldsymbol{X}_j}(\boldsymbol{X}) = \begin{bmatrix} \frac{\partial \phi_1^i(\boldsymbol{X})}{\partial X_{j1}} & \cdots & \frac{\partial \phi_1^i(\boldsymbol{X})}{\partial X_{jK}} \\ \vdots & \ddots & \vdots \\ \frac{\partial \phi_K^i(\boldsymbol{X})}{\partial X_{j1}} & \cdots & \frac{\partial \phi_K^i(\boldsymbol{X})}{\partial X_{jK}} \end{bmatrix} \in \mathbb{R}^{K\times K} \tag{25}$$

For a given matrix $\boldsymbol{Q} \in \mathbb{R}^{K\times K}$, we write $\boldsymbol{Z} \in \mathbb{R}^{n\times K} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{Q}\otimes \boldsymbol{I}_n)$ to denote that the lines of $\boldsymbol{Z}$ are sampled i.i.d. from $\mathcal{N}(\boldsymbol{0}, \boldsymbol{Q})$. Note that this is equivalent to saying that $\boldsymbol{Z} = \tilde{\boldsymbol{Z}}\boldsymbol{Q}^{1/2}$ where $\tilde{\boldsymbol{Z}} \in \mathbb{R}^{n\times K}$ is an i.i.d. standard normal random matrix. The notation $\overset{\mathrm{p}}{\simeq}$ denotes convergence in probability.

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 [65]. 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 [36, 37], included in [66]. Consider a sequence Gaussian matrices $\boldsymbol{A}(n) \in \mathbb{R}^{n\times d}$ with i.i.d. Gaussian entries, $A_{ij}(n) \sim \mathcal{N}(0, 1/d)$. For each $n, d \in \mathbb{N}$, consider two sequences of pseudo-Lipschitz functions

$$\{\boldsymbol{h}_t : \mathbb{R}^{n\times K} \to \mathbb{R}^{n\times K}\}_{t\in\mathbb{N}} \qquad \{\boldsymbol{e}_t : \mathbb{R}^{d\times K} \to \mathbb{R}^{d\times K}\}_{t\in\mathbb{N}} \tag{26}$$

initialized on $\boldsymbol{u}_0 \in \mathbb{R}^{d\times K}$ in such a way that the limit

$$\lim_{d\to\infty} \frac{1}{d}\left\|\boldsymbol{e}_0(\boldsymbol{u}_0)^\top \boldsymbol{e}_0(\boldsymbol{u}_0)\right\|_{\mathrm{F}} \tag{27}$$

exists and it is finite, and recursively define:

$$\boldsymbol{u}^{t+1} = \boldsymbol{A}^\top \boldsymbol{h}_t(\boldsymbol{v}^t) - \boldsymbol{e}_t(\boldsymbol{u}^t)\langle \boldsymbol{h}_t'\rangle^\top \tag{28}$$

$$\boldsymbol{v}^t = \boldsymbol{A}\boldsymbol{e}_t(\boldsymbol{u}^t) - \boldsymbol{h}_{t-1}(\boldsymbol{v}^{t-1})\langle \boldsymbol{e}_t'\rangle^\top \tag{29}$$

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

$$\langle \boldsymbol{h}_t'\rangle = \frac{1}{d}\sum_{i=1}^{n} \frac{\partial \boldsymbol{h}_t^i}{\partial \boldsymbol{v}_i}(\boldsymbol{v}^t) \in \mathbb{R}^{K\times K} \quad \langle \boldsymbol{e}_t'\rangle = \frac{1}{d}\sum_{i=1}^{d} \frac{\partial \boldsymbol{e}_t^i}{\partial \boldsymbol{u}_i}(\boldsymbol{u}^t) \in \mathbb{R}^{K\times K} \tag{30}$$

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

$$\boldsymbol{Q}_{t+1,s} = \boldsymbol{Q}_{s,t+1} = \lim_{d\to\infty} \frac{1}{d}\mathbb{E}\left[\boldsymbol{e}_s(\hat{\boldsymbol{Z}}^s)^\top \boldsymbol{e}_{t+1}(\hat{\boldsymbol{Z}}^{t+1})\right] \in \mathbb{R}^{K\times K} \tag{31}$$

$$\hat{\boldsymbol{Q}}_{t+1,s+1} = \hat{\boldsymbol{Q}}_{s+1,t+1} = \lim_{d\to\infty} \frac{1}{d}\mathbb{E}\left[\boldsymbol{h}_s(\boldsymbol{Z}^s)^\top \boldsymbol{h}_t(\boldsymbol{Z}^t)\right] \in \mathbb{R}^{K\times K} \tag{32}$$

where $(\boldsymbol{Z}^0, \dots, \boldsymbol{Z}^{t-1}) \sim \mathcal{N}(\boldsymbol{0}, \{\boldsymbol{Q}_{r,s}\}_{0\leqslant r,s\leqslant t-1} \otimes \boldsymbol{I}_n)$, $(\hat{\boldsymbol{Z}}^1, \dots, \hat{\boldsymbol{Z}}^t) \sim \mathcal{N}(\boldsymbol{0}, \{\hat{\boldsymbol{Q}}_{r,s}\}_{1\leqslant r,s\leqslant t} \otimes \boldsymbol{I}_d)$ and $\hat{\boldsymbol{Z}}^0 = \boldsymbol{u}_0$. Then the following holds

Theorem 4. *In the setting of the previous paragraph, for any sequence of pseudo-Lipschitz functions* $\phi_n : (\mathbb{R}^{n\times K} \times \mathbb{R}^{d\times K})^t \to \mathbb{R}$, *for* $n, d \to \infty$:

$$\phi_n(\boldsymbol{u}^0, \boldsymbol{v}^0, \boldsymbol{u}^1, \boldsymbol{v}^1, \dots, \boldsymbol{v}^{t-1}, \boldsymbol{u}^t) \overset{\mathrm{P}}{\simeq} \mathbb{E}\left[\phi_n\left(\boldsymbol{u}^0, \boldsymbol{Z}^0, \hat{\boldsymbol{Z}}^1, \boldsymbol{Z}^1, \dots, \boldsymbol{Z}^{t-1}, \hat{\boldsymbol{Z}}^t\right)\right] \tag{33}$$

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

Proof. This theorem is a consequence of Theorem 1 from [66]. □

Spatial coupling As a final premise to our proof, we give the intuition on how to handle a specific form of block random matrix in an AMP sequence. Consider the iteration (28), but this time with a Gaussian matrix defined as:

$$\boldsymbol{A} = \begin{bmatrix} \boldsymbol{A}_1 & & & \\ & \boldsymbol{A}_2 & (0) & \\ & (0) & \ddots & \\ & & & \boldsymbol{A}_K \end{bmatrix} \in \mathbb{R}^{n \times Kd} \tag{34}$$

where $\boldsymbol{A}_k \in \mathbb{R}^{n_k \times d}$ and $\sum_{k=1}^{K} n_k = n$, which leads to the following form for the products between matrices and non-linearities:

$$\boldsymbol{A}^\top \boldsymbol{h}_t(\boldsymbol{v}^t) = \begin{bmatrix} \boldsymbol{A}_1^\top \boldsymbol{h}_{1,t}(\boldsymbol{v}^t) \\ \boldsymbol{A}_2^\top \boldsymbol{h}_{2,t}(\boldsymbol{v}^t) \\ \vdots \\ \boldsymbol{A}_K^\top \boldsymbol{h}_{K,t}(\boldsymbol{v}^t) \end{bmatrix} \in \mathbb{R}^{Kd \times K} \quad \boldsymbol{A}\boldsymbol{e}_t(\boldsymbol{u}_t) = \begin{bmatrix} \boldsymbol{A}_1 \boldsymbol{e}_{1,t}(\boldsymbol{u}^t) \\ \boldsymbol{A}_2 \boldsymbol{e}_{2,t}(\boldsymbol{u}^t) \\ \vdots \\ \boldsymbol{A}_K \boldsymbol{e}_{K,t}(\boldsymbol{u}^t) \end{bmatrix} \in \mathbb{R}^{n \times K} \tag{35}$$

where the blocks $\boldsymbol{h}_{k,t}(\boldsymbol{v}^t) \in \mathbb{R}^{n_k \times K}, \boldsymbol{e}_{k,t}(\boldsymbol{u}_t) \in \mathbb{R}^{d \times K}$ may depend on their full arguments or only the corresponding blocks depending on their separability. This iteration can be embedded as a subset of the iterates of a larger sequence defined with the full version of the matrix $\boldsymbol{A}$ and non-linearities defined as:

$$\begin{aligned} &\boldsymbol{e}_t : \mathbb{R}^{Kd \times K^2} \to \mathbb{R}^{Kd \times K^2} \\ &\text{generates} \quad \begin{bmatrix} \boldsymbol{e}_{1,t}(\bullet) & & & \\ & \boldsymbol{e}_{2,t}(\bullet) & (0) & \\ & (0) & \ddots & \\ & & & \boldsymbol{e}_{K,t}(\bullet) \end{bmatrix} \in \mathbb{R}^{Kd \times K^2} \end{aligned} \tag{36}$$

$$\begin{aligned} &\boldsymbol{h}_t : \mathbb{R}^{n \times K^2} \to \mathbb{R}^{n \times K^2} \\ &\text{generates} \quad \begin{bmatrix} \boldsymbol{h}_{1,t}(\bullet) & & & \\ & \boldsymbol{h}_{2,t}(\bullet) & (0) & \\ & (0) & \ddots & \\ & & & \boldsymbol{h}_{K,t}(\bullet) \end{bmatrix} \in \mathbb{R}^{n \times K^2} \end{aligned} \tag{37}$$

The original iteration is recovered on the block diagonal of the variables of the iteration. This new setting, however, introduces a richer correlation structure, since each block will be described by a different $K \times K$ covariance according to the state evolution equations. Formally, the new covariance will be a $K^2 \times K^2$ block diagonal matrix. Also, the shape of the Onsager term changes from a matrix of size $K \times K$ to one of size $K^2 \times K^2$ with a $K \times (K \times K)$ block diagonal structure.

A.2 Reformulation of the problem

We start by reformulating problem (2) in a way that can be treated efficiently using an AMP iteration. With respect to the main part of this paper, we will consider the estimator $\boldsymbol{W} \in \mathbb{R}^{d \times K}$ instead of $\mathbb{R}^{K \times d}$. The normalized (so that the cost does not diverge with the dimension) problem (2) then reads:

$$\min_{\boldsymbol{W} \in \mathbb{R}^{d \times K}, \boldsymbol{b} \in \mathbb{R}^K} \frac{1}{d} \left(L\left(\boldsymbol{Y}, \frac{1}{\sqrt{d}} \boldsymbol{X}\boldsymbol{W} + \boldsymbol{b} \right) + r(\boldsymbol{W}) \right) \tag{38}$$

where we have introduced the function $L : \mathbb{R}^{n\times K} \times \mathbb{R}^{n\times K} \to \mathbb{R}$ acting as

$$\left(Y, \frac{1}{\sqrt{d}}\boldsymbol{X}\boldsymbol{W} + \boldsymbol{b}\right) \mapsto \sum_{\nu=1}^{n} \ell\left(\boldsymbol{y}^{\nu}, \frac{\boldsymbol{W}\boldsymbol{x}^{\nu}}{\sqrt{d}} + \boldsymbol{b}\right), \tag{39}$$

the matrix $\boldsymbol{Y} \in \mathbb{R}^{n\times K}$ of concatenated one-hot encoded labels, and the matrix of concatenated means $\boldsymbol{M} \in \mathbb{R}^{K\times d}$ (in the main we took the transpose $\boldsymbol{M} \in \mathbb{R}^{d\times K}$). Until further notice, we will drop the scaling $\frac{1}{d}$ for convenience and study the problem

$$\min_{\boldsymbol{W}\in\mathbb{R}^{d\times K}, \boldsymbol{b}\in\mathbb{R}^{K}} L\left(Y, \frac{1}{\sqrt{d}}\boldsymbol{X}\boldsymbol{W} + \boldsymbol{b}\right) + r(\boldsymbol{W}) \tag{40}$$

We will write L_k the application of ℓ on each row of a sub-block in $\mathbb{R}^{n_k\times K}$. Without loss of generality, we can assume that the samples are grouped by clusters in the data matrix, giving the following form for $\boldsymbol{X} \in \mathbb{R}^{n\times d}$, separating the mean part $\boldsymbol{Y}\boldsymbol{M}$ and centered Gaussian part :

$$\boldsymbol{X} = \boldsymbol{Y}\boldsymbol{M} + \tilde{\boldsymbol{Z}}\Sigma \in \mathbb{R}^{n\times d} \tag{41}$$

where we have introduced the block-diagonal matrix $\tilde{\boldsymbol{Z}}$ and the $Kd \times d$ full-column-rank matrix Σ

$$\tilde{\boldsymbol{Z}} = \begin{bmatrix} \boldsymbol{Z}_1 & & & \\ & \boldsymbol{Z}_2 & (0) & \\ & (0) & \ddots & \\ & & & \boldsymbol{Z}_K \end{bmatrix} \in \mathbb{R}^{n\times Kd} \qquad \Sigma = \begin{bmatrix} \Sigma_1^{1/2} \\ \Sigma_2^{1/2} \\ \vdots \\ \Sigma_K^{1/2} \end{bmatrix} \in \mathbb{R}^{Kd\times d}. \tag{42}$$

Here $(\boldsymbol{Z}_1, \dots, \boldsymbol{Z}_K) \in \mathbb{R}^{n_1\times d} \times \cdots \times \mathbb{R}^{n_K\times d}$ are independent, i.i.d. standard normal matrices.

The product between the data matrix and the weights $\boldsymbol{W} \in \mathbb{R}^{d\times K}$ then reads:

$$\boldsymbol{X}\boldsymbol{W} = \boldsymbol{Y}\boldsymbol{M}\boldsymbol{W} + \tilde{\boldsymbol{Z}}\Sigma\boldsymbol{W} = \begin{bmatrix} Y_1\boldsymbol{M}\boldsymbol{W} + \boldsymbol{Z}_1\Sigma_1^{1/2}\boldsymbol{W} \\ \vdots \\ Y_K\boldsymbol{M}\boldsymbol{W} + \boldsymbol{Z}_K\Sigma_K^{1/2}\boldsymbol{W} \end{bmatrix} \in \mathbb{R}^{n\times K} \tag{43}$$

where each $\boldsymbol{Y}_k \in \mathbb{R}^{n_k\times d}$ is a n_k copy of the same label vector. Defining now $\tilde{\boldsymbol{W}} = \Sigma\boldsymbol{W}$, observe that

$$\tilde{\boldsymbol{W}} = \Sigma\boldsymbol{W} \qquad \Longrightarrow \qquad \boldsymbol{W} = \Sigma^{+}\tilde{\boldsymbol{W}}, \tag{44}$$

where

$$\Sigma^{+} \equiv \left(\sum_{k=1}^{K} \Sigma_k\right)^{-1} \Sigma^{\top} \tag{45}$$

is the pseudo-inverse of the matrix Σ. The optimization problem (2) is thus equivalent to

$$\inf_{\substack{\tilde{\boldsymbol{W}}\in\mathbb{R}^{Kd\times K} \\ \boldsymbol{b}\in\mathbb{R}^{K}}} \sum_{k=1}^{K} L_k\left(\frac{1}{\sqrt{d}}\boldsymbol{Y}_k\boldsymbol{M}\boldsymbol{W} + \frac{1}{\sqrt{d}}\boldsymbol{Z}_k\tilde{\boldsymbol{W}}_k, \boldsymbol{b}\right) + r\left(\Sigma^{+}\tilde{\boldsymbol{W}}\right) \tag{46}$$

Introducing the order parameter $\boldsymbol{m} = \frac{1}{\sqrt{d}}\boldsymbol{M}\boldsymbol{W} \in \mathbb{R}^{K\times K}$, we reformulate Eq.(46) as a constrained optimization problem :

$$\begin{aligned} \inf_{\boldsymbol{m},\tilde{\boldsymbol{W}},\boldsymbol{b}} & \sum_{k=1}^{K} L_k\left(\frac{1}{\sqrt{d}}\boldsymbol{Y}_k\boldsymbol{m} + \frac{1}{\sqrt{d}}\boldsymbol{Z}_k\tilde{\boldsymbol{W}}_k\right) + r\left(\Sigma^{+}\tilde{\boldsymbol{W}}\right) \\ \text{s.t.} \quad & \frac{1}{\sqrt{d}}\boldsymbol{M}\Sigma^{+}\tilde{\boldsymbol{W}} = \boldsymbol{m} \end{aligned} \tag{47}$$

whose Lagrangian form, with dual parameters $\hat{\boldsymbol{m}} \in \mathbb{R}^{K\times K}$, reads

$$\inf_{\boldsymbol{m},\tilde{\boldsymbol{W}},\boldsymbol{b}} \sup_{\hat{\boldsymbol{m}}} \sum_{k=1}^{K} L_k\left(\boldsymbol{Y}_k\boldsymbol{m} + \frac{1}{\sqrt{d}}\boldsymbol{Z}_k\tilde{\boldsymbol{W}}_k\right) + r\left(\Sigma^{+}\tilde{\boldsymbol{W}}\right) + \operatorname{tr}\left(\hat{\boldsymbol{m}}^{\top}\left(\boldsymbol{m} - \frac{1}{\sqrt{d}}\boldsymbol{M}\Sigma^{+}\tilde{\boldsymbol{W}}\right)\right). \tag{48}$$

This is a proper, closed, convex, strictly feasible optimization problem, thus strong duality holds and we can invert the order of the inf-sup to focus on the minimization problem in $\tilde{\boldsymbol{W}}$ for fixed $\boldsymbol{m}, \hat{\boldsymbol{m}}, \boldsymbol{b}$:

$$\inf_{\tilde{\boldsymbol{W}}\in\mathbb{R}^{Kd\times K}} \tilde{L}\left(\frac{1}{\sqrt{d}}\tilde{\boldsymbol{Z}}\tilde{\boldsymbol{W}}\right) + \tilde{r}(\tilde{\boldsymbol{W}}) \tag{49}$$

where we defined the loss term

$$\begin{aligned} &\tilde{L} : \mathbb{R}^{n\times K} \to \mathbb{R} \\ &\frac{1}{\sqrt{d}}\tilde{\boldsymbol{Z}}\tilde{\boldsymbol{W}} \mapsto \sum_{k=1}^{K} L_k\left(\boldsymbol{Y}_k\boldsymbol{m} + \frac{1}{\sqrt{d}}\boldsymbol{Z}_k\tilde{\boldsymbol{W}}_k\right) = \sum_{k=1}^{K}\sum_{i=1}^{n_k} \ell\left(\left[\boldsymbol{Y}_k\boldsymbol{m} + \frac{1}{\sqrt{d}}\boldsymbol{Z}_k\tilde{\boldsymbol{W}}_k\right]_i\right) \end{aligned} \tag{50a}$$

and the regularisation term

$$\begin{aligned} &\tilde{r} : \mathbb{R}^{Kd\times K} \to \mathbb{R} \\ &\tilde{\boldsymbol{W}} \mapsto r\left(\Sigma^{+}\tilde{\boldsymbol{W}}\right) + \operatorname{tr}\left(\hat{\boldsymbol{m}}^{\top}\left(\boldsymbol{m} - \frac{1}{\sqrt{d}}\boldsymbol{M}\Sigma^{+}\tilde{\boldsymbol{W}}\right)\right) \end{aligned} \tag{50b}$$

where $\Sigma^{\top}\tilde{\boldsymbol{W}} = \sum_{k=1}^{K}\Sigma_k^{1/2}\boldsymbol{W}_k$ and $\tilde{\boldsymbol{Z}} = [\boldsymbol{Z}_k]_{k=1}^{K} \in \mathbb{R}^{n\times Kd}$ is an i.i.d. standard normal block diagonal matrix as in Eq. (42).

A.3 Finding the AMP sequence

We now need to find an AMP iteration relating to $\tilde{\boldsymbol{W}}$ that solve the optimization problem in Eq. (49). Although this section is not written as a formal proof, all steps are rigorous. The aim is to give the reader the core intuition on how the AMP iteration is found, otherwise the solution may feel "parachuted". The reader uninterested in the underlying intuition may directly skip to the next section. In order to find the appropriate sequence two key points must be considered :

- the fixed point of the sequence has to match the optimality condition of Eq. (49);
- the update rule of the sequence should have the form Eq. (28) for the state evolution equations to hold.

These two points completely determine the form of the iteration. In the subsequent derivation, we absorb the scaling $\frac{1}{\sqrt{d}}$ in the matrix $\tilde{\boldsymbol{Z}}$, such that the $\boldsymbol{Z}_k \in \mathbb{R}^{n_k\times d}$ have i.i.d. $\mathcal{N}(0, 1/d)$ elements.

Resolvent of the loss term — Going back to problem Eq. (49), its optimality condition will look like :

$$\tilde{\boldsymbol{Z}}^\top \partial \tilde{L}(\boldsymbol{Z}\tilde{\boldsymbol{W}}) + \partial \tilde{r}(\tilde{\boldsymbol{W}}) = 0 \iff \begin{bmatrix} \boldsymbol{Z}_1^\top & & & \\ & \boldsymbol{Z}_2^\top & (0) & \\ & (0) & \ddots & \\ & & & \boldsymbol{Z}_K^\top \end{bmatrix} \begin{bmatrix} \partial \tilde{L}_1(\boldsymbol{Z}_1\tilde{\boldsymbol{W}}_1) \\ \partial \tilde{L}_2(\boldsymbol{Z}_2\tilde{\boldsymbol{W}}_2)) \\ \vdots \\ \partial \tilde{L}_K(\boldsymbol{Z}_K\tilde{\boldsymbol{W}}_K)) \end{bmatrix} + \partial \tilde{r}(\tilde{\boldsymbol{W}}) = 0 \tag{51}$$

where each $\boldsymbol{Z}_k \in \mathbb{R}^{n_k \times d}$, and the subdifferential of $\tilde{L}$ is separable across blocks of size $n_k \times d$, and $\partial \tilde{r}(\tilde{\boldsymbol{W}}) \in \mathbb{R}^{Kd \times K}$. Following the intuition of spatial coupling, we introduce the *full* matrix $\boldsymbol{Z} \in \mathbb{R}^{n \times Kd}$, with i.i.d. $\mathcal{N}(0, 1/d)$ entries. The optimality condition can then be written on the diagonal of a $Kd \times K^2$ matrix:

$$\boldsymbol{Z}^\top \begin{bmatrix} \partial \tilde{L}_1(\boldsymbol{Z}_1\tilde{\boldsymbol{W}}_1) & & & \\ & \partial \tilde{L}_2(\boldsymbol{Z}_2\tilde{\boldsymbol{W}}_2) & (0) & \\ & (0) & \ddots & \\ & & & \partial \tilde{L}_K(\boldsymbol{Z}_K\tilde{\boldsymbol{W}}_K) \end{bmatrix} + \begin{bmatrix} \partial \tilde{r}(\tilde{\boldsymbol{W}})_1 & & & \\ & \partial \tilde{r}(\tilde{\boldsymbol{W}})_2 & (0) & \\ & (0) & \ddots & \\ & & & \partial \tilde{r}(\tilde{\boldsymbol{W}})_K \end{bmatrix} = \boldsymbol{0} \tag{52}$$

where $\partial \tilde{r}(\tilde{\boldsymbol{W}})_k$ represents the k-th block of the subdifferential of $\tilde{r}$ which is non-separable across the blocks of $\tilde{\boldsymbol{W}}$. To make the resolvents/proximals appear, we add the argument of the subdifferentials on both sides weighted by a (symmetric) positive definite matrix $\boldsymbol{S}_k \in \mathbb{R}^{K \times K}$ which will be used to allow for Onsager correction while respecting the fixed point condition. Using the notation defined in section A.1

$$\begin{aligned} &\left[\boldsymbol{Z}_k^\top \partial \tilde{L}_k(\boldsymbol{Z}_k\tilde{\boldsymbol{W}}_k)\right]_{k=1}^K + \left[\partial \tilde{r}(\tilde{\boldsymbol{W}})\right]_{k=1}^K = 0 \\ &\iff \left[\boldsymbol{Z}_k^\top \partial \tilde{L}_k(\boldsymbol{Z}_k\tilde{\boldsymbol{W}}_k) + \boldsymbol{Z}_k^\top \boldsymbol{Z}_k \tilde{\boldsymbol{W}}_k \boldsymbol{S}_k^{-1}\right]_{k=1}^K + \left[\partial \tilde{r}(\tilde{\boldsymbol{W}})\right]_{k=1}^K = \left[\boldsymbol{Z}_k^\top \boldsymbol{Z}_k \tilde{\boldsymbol{W}}_k \boldsymbol{S}_k^{-1}\right]_{k=1}^K \end{aligned} \tag{53}$$

for a given set of positive definite matrices $\{\boldsymbol{S}_k\}_{k \in [K]}$. Again, the reason for introducing different $\boldsymbol{S}_k$ on each block is to match the expected structure of the Onsager term. We can introduce the resolvent, formally Bregman resolvent/proximal operator:

$$\boldsymbol{U}_k \equiv \partial \tilde{L}_k(\boldsymbol{Z}_k\tilde{\boldsymbol{W}}_k)\boldsymbol{S}_k + \boldsymbol{Z}_k\tilde{\boldsymbol{W}}_k \iff \boldsymbol{Z}_k\tilde{\boldsymbol{W}}_k = \boldsymbol{R}_{\tilde{L}_k, \boldsymbol{S}_k}(\boldsymbol{U}_k) \tag{54}$$

where

$$\begin{aligned} \boldsymbol{R}_{\tilde{L}_k, \boldsymbol{S}_k}(\boldsymbol{U}_k) &= (\mathrm{Id} + \partial \tilde{L}_k(\bullet)\boldsymbol{S}_k)^{-1}(\boldsymbol{U}_k) \\ &= \underset{\boldsymbol{T} \in \mathbb{R}^{n_k \times K}}{\operatorname{argmin}} \left\{ \tilde{L}_k(\boldsymbol{T}) + \frac{1}{2} \operatorname{tr}\left((\boldsymbol{T} - \boldsymbol{U}_k)\boldsymbol{S}_k^{-1}(\boldsymbol{T} - \boldsymbol{U}_k)^\top\right) \right\} \\ &= \underset{\boldsymbol{T} \in \mathbb{R}^{n_k \times K}}{\operatorname{argmin}} \left\{ L_k(\boldsymbol{T}) + \frac{1}{2} \operatorname{tr}\left((\boldsymbol{T} - (\boldsymbol{Y}_k\boldsymbol{m} + \boldsymbol{U}_k))\boldsymbol{S}_k^{-1}(\boldsymbol{T} - (\boldsymbol{Y}_k\boldsymbol{m} + \boldsymbol{U}_k))^\top\right) \right\} - \boldsymbol{Y}_k\boldsymbol{m}. \end{aligned} \tag{55}$$

In the previous expressions $\partial \tilde{L}_k \in \mathbb{R}^{n_k \times K}$ and $\boldsymbol{V}_k \in \mathbb{R}^{K \times K}$. The following formulation of the optimality condition is reached:

$$\begin{aligned} &\left[\boldsymbol{Z}_k^\top \boldsymbol{U}_k \boldsymbol{S}_k^{-1}\right]_{k=1}^K + \left[\partial \tilde{r}(\tilde{\boldsymbol{W}})_k\right]_{k=1}^K = \left[\boldsymbol{Z}_k^\top \boldsymbol{R}_{\tilde{L}_k, \boldsymbol{S}_k}(\boldsymbol{U}_k)\boldsymbol{S}_k^{-1}\right]_{k=1}^K \\ &\iff \left[\boldsymbol{Z}_k^\top \left(\boldsymbol{U}_k - \boldsymbol{R}_{\tilde{L}_k, \boldsymbol{S}_k}(\boldsymbol{U}_k)\right)\boldsymbol{S}_k^{-1}\right]_{k=1}^K + \left[\partial \tilde{r}(\tilde{\boldsymbol{W}})_k\right]_{k=1}^K = 0 \end{aligned} \tag{56}$$

Resolvent of the regularization term Determining the block decomposition of the subdifferential of the regularization term is less simple. We would like a block expression in the flavour of:

$$\left[\partial\tilde{r}(\tilde{\boldsymbol{W}})_k\right]_{k=1}^K + \left[\tilde{\boldsymbol{W}}_k\hat{\boldsymbol{S}}_k^{-1}\right]_{k=1}^K = \left[\tilde{\boldsymbol{W}}_k\hat{\boldsymbol{S}}_k^{-1}\right]_{k=1}^K \tag{57}$$

At this point it becomes clear that we cannot consider the resolvent as acting on $\tilde{\boldsymbol{W}} \in \mathbb{R}^{Kd\times K}$ otherwise there could be only one $\hat{\boldsymbol{S}} \in \mathbb{R}^{K\times K}$ and there would be a mismatch with the expected form of the Onsager terms. As specified by the definitions Eq.(50), the subdifferential of $\tilde{r}$ is acting on the whole block diagonal matrix $[\tilde{\boldsymbol{W}}_k]_{k=1}^K$, by way of summation due to the action of the pseudo-inverse Σ^+. We can thus consider its proximal acting on $\mathbb{R}^{d\times K^2}$ as $[\tilde{\boldsymbol{W}}_1\tilde{\boldsymbol{W}}_2...\tilde{\boldsymbol{W}}_K]$ (note that we could have also worked directly with a block diagonal matrix in $\mathbb{R}^{Kd\times K^2}$). Proceeding in this way, we can directly write our expression as an application parametrized by another set of positive definite matrices $\{\hat{\boldsymbol{S}}_k\}_{k\in[K]}$, and introduce the resolvent

$$\hat{\boldsymbol{U}} \equiv \left(\mathrm{Id} + \partial\tilde{r}(\bullet)\hat{\boldsymbol{S}}\right)(\tilde{\boldsymbol{W}}) \qquad \tilde{\boldsymbol{W}} = \boldsymbol{R}_{\tilde{r},\hat{\boldsymbol{S}}}(\hat{\boldsymbol{U}}) \tag{58}$$

where

$$\boldsymbol{R}_{\tilde{r},\hat{\boldsymbol{S}}}(\hat{\boldsymbol{U}}) = \left(\mathrm{Id} + \partial\tilde{r}(\bullet)\hat{\boldsymbol{S}}\right)^{-1}(\hat{\boldsymbol{U}}) = \underset{\boldsymbol{T}\in\mathbb{R}^{d\times K^2}}{\mathrm{argmin}}\left\{\tilde{r}(\boldsymbol{T}) + \frac{1}{2}\mathrm{tr}\left((\boldsymbol{T}-\hat{\boldsymbol{U}})\hat{\boldsymbol{S}}^{-1}(\boldsymbol{T}-\hat{\boldsymbol{U}})^\top\right)\right\} \tag{59}$$

where $\hat{\boldsymbol{S}} \in \mathbb{R}^{K^2\times K^2}$ block diagonal, and $\hat{\boldsymbol{U}} \in \mathbb{R}^{d\times K^2}$. This would lead to the equivalent optimality condition for the regularization part:

$$\hat{\boldsymbol{U}}\hat{\boldsymbol{S}}^{-1} = \boldsymbol{R}_{\tilde{r},\hat{\boldsymbol{S}}}(\hat{\boldsymbol{U}})\hat{\boldsymbol{S}}^{-1} \iff \left[\hat{\boldsymbol{U}}_k\hat{\boldsymbol{S}}_k^{-1}\right]_{k=1}^K = \left[\boldsymbol{R}_{\tilde{r},\hat{\boldsymbol{S}},k}(\hat{\boldsymbol{U}})\hat{\boldsymbol{S}}_k^{-1}\right]_{k=1}^K \tag{60}$$

We now need to figure out the block structure of this resolvent since we want to spread it across a block diagonal matrix. Let $\boldsymbol{C} = \sum_{k=1}^K \Sigma_k$, so that $\Sigma^+ = \boldsymbol{C}^{-1}\Sigma^\top$, and the blocks $\boldsymbol{T}_k \in \mathbb{R}^{d\times K}$ are the solution to the minimization problem

$$\min_{\{\boldsymbol{T}_k\}_{k\in[K]}\in(\mathbb{R}^{d\times K})^K} r(\boldsymbol{C}^{-1}\sum_{k=1}^K \Sigma_k^{1/2}\boldsymbol{T}_k) + \frac{1}{2}\mathrm{tr}\left((\boldsymbol{T}-\hat{\boldsymbol{U}})\hat{\boldsymbol{S}}^{-1}(\boldsymbol{T}-\hat{\boldsymbol{U}}^\top)\right) + \mathrm{tr}\left(\hat{\boldsymbol{m}}^\top\left(\boldsymbol{m} - \frac{1}{\sqrt{d}}\boldsymbol{M}\Sigma^+\boldsymbol{T}\right)\right) \tag{61}$$

Let $\tilde{\boldsymbol{T}} = \boldsymbol{C}^{-1}\sum_{k=1}^K \Sigma_k^{1/2}\boldsymbol{T}_k \in \mathbb{R}^{d\times K}$, and the equivalent reformulation as a constraint optimization problem:

$$\begin{aligned}\min_{\substack{\boldsymbol{T}_{k\in[K]}\in\mathbb{R}^{d\times K}\\ \tilde{\boldsymbol{T}}\in\mathbb{R}^{d\times K}}} & r(\tilde{\boldsymbol{T}}) + \frac{1}{2}\mathrm{tr}\left((\boldsymbol{T}-\hat{\boldsymbol{U}})\hat{\boldsymbol{S}}^{-1}(\boldsymbol{T}-\hat{\boldsymbol{U}}^\top)\right) + \mathrm{tr}\left(\hat{\boldsymbol{m}}^\top\left(\boldsymbol{m} - \frac{1}{\sqrt{d}}\boldsymbol{M}\tilde{\boldsymbol{T}}\right)\right) \\ \text{s.t.} \quad & \tilde{\boldsymbol{T}} = \boldsymbol{C}^{-1}\sum_{k=1}^K \Sigma_k^{1/2}\boldsymbol{T}_k\end{aligned} \tag{62}$$

This is a feasible convex problem under convex constraint with a strongly convex term, it thus has a unique solution and strong duality holds. Introducing the Lagrange multiplier $\boldsymbol{\lambda} \in \mathbb{R}^{d\times K}$, we get the equivalent representation:

$$\min_{\substack{\boldsymbol{T}_{k\in[K]}\in\mathbb{R}^{d\times K}\\ \tilde{\boldsymbol{T}}\in\mathbb{R}^{d\times K}}} \max_{\boldsymbol{\lambda}\in\mathbb{R}^{d\times K}} r(\tilde{\boldsymbol{T}}) + \sum_{k=1}^K \mathrm{tr}\left((\boldsymbol{T}_k-\hat{\boldsymbol{U}}_k)\hat{\boldsymbol{S}}_k^{-1}(\boldsymbol{T}_k-\hat{\boldsymbol{U}}_k)^\top\right)$$

$$+ \operatorname{tr}\left(\boldsymbol{\lambda}^\top \left(\tilde{\boldsymbol{T}} - \boldsymbol{C}^{-1} \sum_{k=1}^{K} \Sigma_k^{1/2} \boldsymbol{T}_k\right)\right) + \operatorname{tr}\left(\hat{\boldsymbol{m}}^\top \left(\boldsymbol{m} - \frac{1}{\sqrt{d}} \boldsymbol{M}\tilde{\boldsymbol{T}}\right)\right). \tag{63}$$

The optimality condition for this problem reads:

$$\partial_{\tilde{\boldsymbol{T}}} : \quad \partial r(\tilde{\boldsymbol{T}}) + \boldsymbol{\lambda} - \frac{1}{\sqrt{d}} \boldsymbol{M}^\top \hat{\boldsymbol{m}} = 0 \tag{64}$$

$$\partial_{\boldsymbol{T}} : \quad (\boldsymbol{T}_k - \boldsymbol{U}_k)\hat{\boldsymbol{S}}_k^{-1} = \Sigma_k^{1/2} \boldsymbol{C}^{-1} \boldsymbol{\lambda} \qquad \forall k \in [K] \tag{65}$$

$$\partial_{\boldsymbol{\lambda}} : \quad \tilde{\boldsymbol{T}} = \boldsymbol{C}^{-1} \sum_{k=1}^{K} \Sigma_k^{1/2} \boldsymbol{T}_k \tag{66}$$

Using the gradient condition on $\boldsymbol{T}$, we get

$$\sum_{k=1}^{K} \Sigma_k^{1/2} (\boldsymbol{T}_k - \hat{\boldsymbol{U}}_k)\hat{\boldsymbol{S}}_k^{-1} = \boldsymbol{\lambda} \tag{67}$$

The constraint $\tilde{\boldsymbol{T}} = \boldsymbol{C}^{-1} \sum_{k=1}^{K} \Sigma_k^{1/2} \boldsymbol{T}_k$ is solved by $\boldsymbol{T}_k = \Sigma_k^{1/2} \tilde{\boldsymbol{T}}$ which gives the solution for $\boldsymbol{\lambda}$

$$\boldsymbol{\lambda} = \sum_{k=1}^{K} \Sigma_k^{1/2} (\Sigma_k^{1/2} \tilde{\boldsymbol{T}} - \hat{\boldsymbol{U}}_k)\hat{\boldsymbol{S}}_k^{-1} = \sum_{k=1}^{K} \Sigma_k \tilde{\boldsymbol{T}} \hat{\boldsymbol{S}}_k^{-1} - \sum_{k=1}^{K} \Sigma_k^{1/2} \hat{\boldsymbol{U}}_k \hat{\boldsymbol{S}}_k^{-1} \tag{68}$$

and prescribes the following form for $\tilde{\boldsymbol{T}}$, as solution to the problem

$$\begin{aligned} &\partial r(\tilde{\boldsymbol{T}}) + \sum_{k=1}^{K} \Sigma_k \tilde{\boldsymbol{T}} \hat{\boldsymbol{S}}_k^{-1} - \sum_{k=1}^{K} \Sigma_k^{1/2} \hat{\boldsymbol{U}}_k \hat{\boldsymbol{S}}_k^{-1} - \frac{1}{\sqrt{d}} \boldsymbol{M}^\top \hat{\boldsymbol{m}} = 0 \\ &\iff \operatorname*{argmin}_{\tilde{\boldsymbol{T}}} r(\tilde{\boldsymbol{T}}) + \frac{1}{2} \sum_{k=1}^{K} \Sigma_k \tilde{\boldsymbol{T}} \hat{\boldsymbol{S}}_k^{-1} \tilde{\boldsymbol{T}} - \left(\sum_{k=1}^{K} \Sigma_k^{1/2} \hat{\boldsymbol{U}}_k \hat{\boldsymbol{S}}_k^{-1} + \frac{1}{\sqrt{d}} \boldsymbol{M}^\top \hat{\boldsymbol{m}} \right) \tilde{\boldsymbol{T}} \end{aligned} \tag{69}$$

and then recover $\boldsymbol{T}$ from $\boldsymbol{T} = \Sigma \tilde{\boldsymbol{T}}$. Thus, defining the function

$$\begin{aligned} &\boldsymbol{\eta} : \mathbb{R}^{d \times K^2} \to \mathbb{R}^{d \times K} \\ &\hat{\boldsymbol{U}} \mapsto \operatorname*{argmin}_{\tilde{\boldsymbol{T}}} r(\tilde{\boldsymbol{T}}) + \frac{1}{2} \sum_{k=1}^{K} \Sigma_k \tilde{\boldsymbol{T}} \hat{\boldsymbol{S}}_k^{-1} \tilde{\boldsymbol{T}} - \left(\sum_{k=1}^{K} \Sigma_k^{1/2} \hat{\boldsymbol{U}}_k \hat{\boldsymbol{S}}_k^{-1} + \frac{1}{\sqrt{d}} \boldsymbol{M}^\top \hat{\boldsymbol{m}} \right) \tilde{\boldsymbol{T}} \end{aligned} \tag{70}$$

the block decomposition of the resolvent for the regularizer reads:

$$\boldsymbol{R}_{\tilde{r},\hat{\boldsymbol{S}},k}(\hat{\boldsymbol{U}}) = \Sigma_k^{1/2} \boldsymbol{\eta}(\hat{\boldsymbol{U}}) \tag{71}$$

Matching the optimality condition with the AMP fixed point The global optimality condition then reads:

$$\left[\boldsymbol{Z}_k^\top \left(\boldsymbol{R}_{\tilde{L}_k, \boldsymbol{S}_k}(\boldsymbol{U}_k) - \boldsymbol{U}_k \right) \boldsymbol{S}_k^{-1} \right]_{k=1}^{K} = \left[(\hat{\boldsymbol{U}}_k - \boldsymbol{R}_{\tilde{r},\hat{\boldsymbol{S}},k}(\hat{\boldsymbol{U}}))\hat{\boldsymbol{S}}_k^{-1} \right]_{k=1}^{K} \tag{72}$$

$$\left[\boldsymbol{Z}_k \boldsymbol{R}_{\tilde{r},\hat{\boldsymbol{S}},k}(\hat{\boldsymbol{U}}) \right]_{k=1}^{K} = \left[\boldsymbol{R}_{\tilde{L}_k, \boldsymbol{S}_k}(\boldsymbol{U}_k) \right]_{k=1}^{K} \tag{73}$$

where both equations should be satisfied. We can now define update functions based on the previously obtained block decomposition. The fixed point of the matrix-valued AMP Eq.(28) reads:

$$\boldsymbol{u} + \boldsymbol{e}(\boldsymbol{u})\langle \boldsymbol{h}' \rangle^\top = \boldsymbol{Z}^\top \boldsymbol{h}(\boldsymbol{v}) \tag{74}$$

$$\boldsymbol{v} + \boldsymbol{h}(\boldsymbol{v})\langle \boldsymbol{e}' \rangle^\top = \boldsymbol{Z}\boldsymbol{e}(\boldsymbol{u}) \tag{75}$$

Matching this fixed point with the optimality condition Eq.(72) suggests the following mapping:

$$\begin{aligned} \boldsymbol{h}_k(\boldsymbol{U}_k) &= \left(\boldsymbol{R}_{\tilde{L}_k, \boldsymbol{S}_k}(\boldsymbol{U}_k) - \boldsymbol{U}_k\right) \boldsymbol{S}_k^{-1}, & \boldsymbol{S}_k &= \langle \boldsymbol{e}_k' \rangle^\top, \\ \boldsymbol{e}_k(\hat{\boldsymbol{U}}) &= \boldsymbol{R}_{\tilde{r}, \hat{\boldsymbol{S}}, k}(\hat{\boldsymbol{U}}\hat{\boldsymbol{S}}), & \hat{\boldsymbol{S}}_k &= -(\langle \boldsymbol{h}_k' \rangle^\top)^{-1}, \end{aligned} \tag{76}$$

where we redefined $\hat{\boldsymbol{U}} \equiv \hat{\boldsymbol{U}}\hat{\boldsymbol{S}}$ in (58).

A.4 Proof of Theorem 1 using the AMP sequence

Following the analysis carried out in the previous section, define the following two sequences of non-linearities, for fixed values of the parameters $\hat{\boldsymbol{m}}, \boldsymbol{m}, \boldsymbol{b}$ and any $\boldsymbol{u} \in \mathbb{R}^{d \times K^2}, \boldsymbol{v} \in \mathbb{R}^{n \times K}$:

$$\begin{aligned} &\boldsymbol{e}_t : \mathbb{R}^{Kd \times K^2} \to \mathbb{R}^{Kd \times K^2} \\ &\boldsymbol{u} \mapsto \begin{bmatrix} \boldsymbol{e}_{1,t}(\boldsymbol{u}) & & & \\ & \boldsymbol{e}_{2,t}(\boldsymbol{u}) & (0) & \\ & (0) & \ddots & \\ & & & \boldsymbol{e}_{K,t}(\boldsymbol{u}) \end{bmatrix} \in \mathbb{R}^{Kd \times K^2} \end{aligned} \tag{77}$$

$$\begin{aligned} &\boldsymbol{h}_t : \mathbb{R}^{n \times K^2} \to \mathbb{R}^{n \times K^2} \\ &\boldsymbol{v} \mapsto \begin{bmatrix} \boldsymbol{h}_{1,t}(\boldsymbol{v}_1) & & & \\ & \boldsymbol{h}_{2,t}(\boldsymbol{v}_2) & (0) & \\ & (0) & \ddots & \\ & & & \boldsymbol{h}_{K,t}(\boldsymbol{v}_K) \end{bmatrix} \in \mathbb{R}^{n \times K^2} \end{aligned} \tag{78}$$

where

$$\begin{aligned} &\boldsymbol{h}_{k,t} : \mathbb{R}^{n_k \times K} \to \mathbb{R}^{n_k \times K} \\ &\quad \boldsymbol{v}_k \mapsto \left(\boldsymbol{R}_{\tilde{L}_k, \boldsymbol{V}_k^t}(\boldsymbol{v}_k) - \boldsymbol{v}_k\right)(\boldsymbol{V}_k^t)^{-1} \\ &\quad = \left(\operatorname*{argmin}_{\boldsymbol{T} \in \mathbb{R}^{n_k \times K}} \left\{ \tilde{L}_k(\boldsymbol{T}) + \frac{1}{2}\operatorname{tr}\left((\boldsymbol{T} - \boldsymbol{v}_k)(\boldsymbol{V}_k^t)^{-1}(\boldsymbol{T} - \boldsymbol{v}_k)^\top\right) \right\} - \boldsymbol{v}_k \right)(\boldsymbol{V}_k^t)^{-1} \\ &\quad = \left(\operatorname{Prox}_{L_k(\bullet(\boldsymbol{V}_k^t)^{1/2})}((\boldsymbol{Y}_k\boldsymbol{m} + \boldsymbol{v}_k)(\boldsymbol{V}_k^t)^{-1/2})(\boldsymbol{V}_k^t)^{1/2} - (\boldsymbol{Y}_k\boldsymbol{m} + \boldsymbol{v}_k)\right)(\boldsymbol{V}_k^t)^{-1} \\ &\boldsymbol{e}_{k,t} : \mathbb{R}^{d \times K^2} \to \mathbb{R}^{d \times K} \\ &\quad \boldsymbol{u} \mapsto \Sigma_k^{1/2} \operatorname*{argmin}_{\tilde{\boldsymbol{T}} \in \mathbb{R}^{d \times K}} r(\tilde{\boldsymbol{T}}) + \frac{1}{2}\sum_{k=1}^{K} \Sigma_k \tilde{\boldsymbol{T}} \hat{\boldsymbol{V}}_k^t \tilde{\boldsymbol{T}} - \left(\sum_{k=1}^{K} \Sigma_k^{1/2}\boldsymbol{u}_k + \frac{1}{\sqrt{d}}\boldsymbol{M}^\top \hat{\boldsymbol{m}}\right)\tilde{\boldsymbol{T}} \\ &\quad = \Sigma_k^{1/2}\boldsymbol{\eta}(\boldsymbol{u}(\hat{\boldsymbol{V}}^t)^{-1}) \end{aligned} \tag{79, 80}$$

where $Y_k \in \mathbb{R}^{n_k \times K}$ and $(V^t, \hat{V}^t) \in \mathbb{R}^{K^2 \times K^2}$, are defined as the block diagonal matrices $\left[V_k^t\right]_{k\in[K]}, \left[\hat{V}_k^t\right]_{k\in[K]}$ such that

$$V_k^t = \langle (e_k^{t-1})' \rangle^\top \quad \hat{V}_k^t = -\langle (h_k^t)' \rangle^\top \tag{81}$$

using the notation from Eq. (30). Since the functions defining e_k, h_k are proximal operators, their Jacobians are positive semi-definite. Assuming they are not zero almost everywhere, the matrices obtained by averaging the Jacobians over Gaussian measures will be positive definite, justifying the validity of the choice Eq.(81) for the Onsager correction terms. Now define the following sequence, initialized with

$$u_0, h_{-1} \equiv 0, \hat{V}_0 \tag{82}$$

$$\text{such that } \lim_{d\to\infty} \frac{1}{d} \left\| e_0(u_0)^\top e_0(u_0) \right\|_{\mathrm{F}} < +\infty \text{ and } \hat{V}_0 \in \mathbb{S}_K^{++}$$

and recursively define

$$u^{t+1} = Z^\top h_t(v^t) - e_t(u^t) \langle h_t' \rangle^\top \tag{83}$$

$$v^t = Z e_t(u^t) - h_{t-1}(v^{t-1}) \langle e_t' \rangle^\top \tag{84}$$

where $Z \in \mathbb{R}^{n \times Kd}$ has i.i.d. $\mathcal{N}(0, 1/d)$ elements, and in the Jacobians defining $\hat{V}, V$, we used the notation from Eq. (25).

State evolution equations The results from section A.3 show that the functions e^t, h^t are proximals operators, and thus are Lipschitz continuous for all $t \in \mathbb{N}$, along with their block restrictions. Therefore the conditions of Theorem 4 are verified and we have the following lemma:

Lemma 5. *Consider the sequence defined by Eq.(82), for any fixed $m, \hat{m}, b$. For any sequences of pseudo-Lipschitz functions $\phi_{1,n} : \mathbb{R}^{d\times K^2} \to \mathbb{R}$, $\phi_{2,n} : \mathbb{R}^{n\times K^2} \to \mathbb{R}$, for any $t \in \mathbb{N}^*$:*

$$\phi_{1,n}(u_1^t, \dots, u_K^t) \overset{\mathrm{P}}{\simeq} \mathbb{E}\left[\phi_{1,n}(H_1(\hat{Q}_1^t)^{1/2}, \dots, H_K(\hat{Q}_K^t)^{1/2}) \right] \tag{85}$$

$$\phi_{2,n}(v_1, \dots, v_K) \overset{\mathrm{P}}{\simeq} \mathbb{E}\left[\phi_{1,n}(G_1(Q_1^t)^{1/2}, \dots, G_K(Q_K^t)^{1/2}) \right] \tag{86}$$

where the matrices $H_k \in \mathbb{R}^{d\times K}, G_k \in \mathbb{R}^{n_k \times K}$ are i.i.d. standard normal matrices, and at each time step $t \geqslant 1$

$$Q_k^t = \lim_{d\to+\infty} \frac{1}{d} \mathbb{E}\left[e_k(\{H_k(\hat{Q}_k^t)^{1/2}(\hat{V}_k^t)^{-1}\}_{k\in[K]})^\top e_k(\{H_k(\hat{Q}_k^t)^{1/2}(\hat{V}_k^t)^{-1}\}_{k\in[K]}) \right] \in \mathbb{R}^{K\times K} \tag{87}$$

$$\hat{Q}_k^t = \lim_{d\to+\infty} \frac{1}{d} \mathbb{E}\left[h_k^{t-1}(G_k(Q_k^{t-1})^{1/2})^\top h_k^{t-1}(G_k(Q_k^{t-1})^{1/2}) \right] \in \mathbb{R}^{K\times K} \tag{88}$$

$$V_k^t = \lim_{d\to+\infty} \frac{1}{d} \sum_{i=1}^{d} \frac{\partial e_k^{t-1}(\{H_k(\hat{Q}_k^{t-1})^{1/2}\}_{k\in[K]})}{\partial (H_k(\hat{Q}_k^{t-1})^{1/2})_i} \in \mathbb{R}^{K\times K} \tag{89}$$

$$\hat{V}_k^t = -\lim_{d\to+\infty} \frac{1}{d} \sum_{i=1}^{n_k} \frac{\partial h_k^t(G_k(Q_k^t)^{1/2})}{\partial (G_k(Q_k^t)^{1/2})_i} \in \mathbb{R}^{K\times K} \tag{90}$$

where the sequence is initialized with $\hat{V}_0, e_0, Q_{0,0} = \lim_{d\to\infty} \frac{1}{d} \|e_0(u_0)^\top e_0(u_0)\|_{\mathrm{F}}$.

Proof. Lemma 5 is a consequence of Theorem 4 whose assumptions have been verified in the paragraph. □

Note that the G defined here is not the same as the G in the replica computation, and that in Lemma 5, we have directly written the block decomposition of the state evolution corresponding to the iteration Eq. (82), which involves the block diagonal matrices $Q^t, \hat{Q}^t, V^t, \hat{V}^t$ which are all in $\mathbb{R}^{K^2 \times K^2}$. Using the notations introduced in section A.1

$$V = [V_k]_{k=1}^K \;\; \hat{V} = [\hat{V}_k]_{k=1}^K \;\; Q = [Q_k]_{k=1}^K \;\; \hat{Q} = [\hat{Q}_k]_{k=1}^K \tag{91}$$

Also note that we do not use the full state evolution giving the correlations across all time steps, but only use those at equal times t.

Trajectories and fixed point of the AMP sequence Now that we have a sequence with state evolution equations, the following two lemmas link the fixed points of this iteration to any optimal solution of problem Eq.(49).

Lemma 6. *Consider any fixed point $V, \hat{V}, Q, \hat{Q}$ of the state evolution equations from Lemma 5. For any fixed point u^*, v^* of iteration Eq.(82), the quantity*

$$R_{\tilde{r},\hat{V}}(u^* \hat{V}^{-1}) = \left(\mathrm{Id} + \partial \tilde{r}(\bullet)\hat{V}^{-1}\right)(u^* \hat{V}^{-1}) \tag{92}$$

is an optimal solution $\tilde{W}^\star$ of problem Eq.(49). Furthermore

$$R_{\tilde{L},V}(v^*) = (\mathrm{Id} + \partial \tilde{L}(\bullet)V)(v^*) = Z\tilde{W}^\star \tag{93}$$

where the block decompositions of each resolvents have been explicitly calculated in section A.3.

Proof. Lemma 6 is a direct consequence of the analysis carried out in section A.3. □

At this point we know the fixed points of the AMP iteration correspond to the optimal solutions of problem Eq.(49). Note that the resolvents/proximals linking the fixed point of the AMP iteration with the solutions of Eq.(49) are Lipschitz continuous, making them acceptable transforms for state evolution observables. However this does not guarantee that the optimal solution is characterized by the fixed point of the state evolution equations. Indeed, we need to show that a converging trajectory can be systematically found for any instance of the problem Eq.(49). This is the purpose of the following lemma.

Lemma 7. *Consider iteration Eq.(82), where the parameters $Q, \hat{Q}, V, \hat{V}$ are initialized at any fixed point of the state evolution equations of Lemma 5. For any sequence initialized with $\hat{V}_0 = \hat{V}$ and u_0 such that*

$$\lim_{d\to\infty} \frac{1}{d} e_0(u_0)^\top e_0(u_0) = Q \tag{94}$$

the following holds

$$\lim_{t\to\infty} \lim_{d\to\infty} \frac{1}{\sqrt{d}} \left\| u^t - u^\star \right\|_{\mathrm{F}} = 0 \quad \lim_{t\to\infty} \lim_{d\to\infty} \frac{1}{\sqrt{d}} \left\| v^t - v^\star \right\|_{\mathrm{F}} = 0 \tag{95}$$

Proof. The proof of Lemma 7 will be given in a longer version of this paper. □

Combining the lemmas 5, 6 and 7 with the pseudo-Lipschitz property, we have reached the following lemma

Lemma 8. *For any fixed $\boldsymbol{m}, \hat{\boldsymbol{m}}, \boldsymbol{b}$, consider the fixed point $(\boldsymbol{Q}, \hat{\boldsymbol{Q}}, \boldsymbol{V}, \hat{\boldsymbol{V}})$ of the state evolution equations from Lemma. 5. Then, for any sequences of pseudo-Lipschitz functions $\phi_{1,n} : \mathbb{R}^{d\times K^2} \to \mathbb{R}, \phi_{2,n} : \mathbb{R}^{n\times K} \to \mathbb{R}$, for $n, d \to \infty$*

$$\phi_{1,n}(\tilde{\boldsymbol{W}}^\star) \overset{\mathrm{P}}{\simeq} \mathbb{E}\left[\phi_{1,n}\left(R_{\tilde{r},\hat{V}}(\boldsymbol{H}\hat{\boldsymbol{Q}}^{1/2}\hat{\boldsymbol{V}}^{-1})\right)\right] \tag{96}$$

$$\phi_{2,n}(Z\tilde{\boldsymbol{W}}^\star) \overset{\mathrm{P}}{\simeq} \mathbb{E}\left[\phi_{2,n}\left(R_{\tilde{L},V}(\boldsymbol{G}\boldsymbol{Q}^{1/2})\right)\right] \tag{97}$$

where we remind that $\boldsymbol{G} = [\boldsymbol{G}_k]_{k=1}^K, \boldsymbol{H} = [\boldsymbol{H}_k]_{k=1}^K$ are block diagonal i.i.d. standard normal matrices as in Lemma 5, and $\boldsymbol{Q} = \left[\boldsymbol{Q}_k\right]_{k=1}^K$ $\hat{\boldsymbol{Q}} = \left[\hat{\boldsymbol{Q}}_k\right]_{k=1}^K$ are the $K^2 \times K^2$ block diagonal covariances.

Proof. Lemma 8 is a consequence of Lemmas 5,6,7 and the pseudo-Lipschitz property. It is also necessary to check that all iterates of the AMP iteration and any optimal solution of problem Eq.(49) have finite norm (rescaled by $\frac{1}{\sqrt{d}}$, to prove convergence results involving the pseudo-Lipschitz functions. This is ensured by the state evolution equations for the iterates, and can be justified for the optimal solutions using the feasibility of the optimization problem Eq.(49) using similar arguments as in [12] Lemma 8. Note that the composition of a pseudo-Lipschitz function and a Lipschitz function is also pseudo-Lipschitz. □

Note that the resolvents are implicitly acting on the block diagonals of their arguments. At this point we are quite close to Theorem 1(details for the exact matching will be given later), but we are missing the equations on $\boldsymbol{m}, \hat{\boldsymbol{m}}, \boldsymbol{b}$.

Fixed point equations for $\boldsymbol{m}, \hat{\boldsymbol{m}}, \boldsymbol{b}$ We drop the dependence on the bias term $\boldsymbol{b}$ as its solution is very similar to the one for $\hat{\boldsymbol{m}}$. To obtain the equations for $\boldsymbol{m}, \hat{\boldsymbol{m}}$, we go back to the complete optimization problem

$$\inf_{\boldsymbol{m},\tilde{\boldsymbol{W}},\boldsymbol{b}} \sup_{\hat{\boldsymbol{m}}} L(Y_k\boldsymbol{m} + Z_k\tilde{\boldsymbol{W}}_k) + r\left(\Sigma^+\tilde{\boldsymbol{W}}\right) + \mathrm{tr}\left(\hat{\boldsymbol{m}}^\top\left(\boldsymbol{m} - \frac{1}{\sqrt{d}}\boldsymbol{M}\Sigma^+\tilde{\boldsymbol{W}}\right)\right) \tag{98}$$

where we can use strong duality to write the equivalent form

$$\inf_{\boldsymbol{m},\boldsymbol{b}} \sup_{\hat{\boldsymbol{m}}} L(Y_k\boldsymbol{m} + Z_k\tilde{\boldsymbol{W}}_k^\star) + r\left(\Sigma^+\tilde{\boldsymbol{W}}\right) + \mathrm{tr}\left(\hat{\boldsymbol{m}}^\top\left(\boldsymbol{m} - \frac{1}{\sqrt{d}}\boldsymbol{M}\Sigma^+\tilde{\boldsymbol{W}}^\star\right)\right) \tag{99}$$

The gradients w.r.t. $\boldsymbol{m}, \hat{\boldsymbol{m}}$ then read:

$$\partial\hat{\boldsymbol{m}} = \boldsymbol{m} - \frac{1}{\sqrt{d}}\boldsymbol{M}\Sigma^+\tilde{\boldsymbol{W}}^\star \tag{100}$$

$$\partial\boldsymbol{m} = \hat{\boldsymbol{m}} + \partial_{\boldsymbol{m}}L(Y\boldsymbol{m} + Z\tilde{\boldsymbol{W}}^\star) \tag{101}$$

Uniform convergence of derivatives and conditions for the dominated convergence theorem are verified using similar arguments as in [12, Lemma 12]. We can thus invert limits and derivatives, and expectations and derivatives. To facilitate taking the derivative $\partial_{\boldsymbol{m}}$, we use Lemma 8 (assuming the normalized loss function is pseudo-Lipschitz, which is a very loose assumption verified by most machine learning losses) to obtain, reintroducing the scaling $1/d$

$$\frac{1}{d}L(Y\boldsymbol{m} + Z\tilde{\boldsymbol{W}}^\star) \xrightarrow[d\to\infty]{P} \frac{1}{d}\mathbb{E}\left[L(Y\boldsymbol{m} + \boldsymbol{R}_{\tilde{L},V}(\boldsymbol{G}\boldsymbol{Q}^{1/2}))\right] \tag{102}$$

Using the block decomposition from Eq.(55), the blocks $(\boldsymbol{R}_{\tilde{L},V}(\boldsymbol{G}\boldsymbol{Q}^{1/2}))_k \in \mathbb{R}^{n_k \times K}$ are given by:

$$\underset{\boldsymbol{T}\in\mathbb{R}^{n_k\times K}}{\operatorname{argmin}}\left\{L_k(\boldsymbol{T})+\frac{1}{2}\operatorname{tr}\left((\boldsymbol{T}-(\boldsymbol{Y}_k\boldsymbol{m}+\boldsymbol{G}_k\boldsymbol{Q}_k^{1/2}))\boldsymbol{V}_k^{-1}(\boldsymbol{T}-(\boldsymbol{Y}_k\boldsymbol{m}+\boldsymbol{G}_k\boldsymbol{Q}_k^{1/2}))^\top\right)\right\}-\boldsymbol{Y}_k\boldsymbol{m} \tag{103}$$

Using a block diagonal representation, we can write:

$$\begin{aligned}&\frac{1}{d}L(\boldsymbol{Y}\boldsymbol{m}+R_{\tilde{L},V}(\boldsymbol{G}\boldsymbol{Q}^{1/2}))=\frac{1}{d}L(R_{L,V}(\boldsymbol{Y}\boldsymbol{m}+\boldsymbol{G}\boldsymbol{Q}^{1/2}))\\ =&\frac{\mathcal{M}_{L,V}(\boldsymbol{Y}\boldsymbol{m}+\boldsymbol{G}\boldsymbol{Q}^{1/2})}{d}-\frac{1}{2d}\operatorname{tr}\left((\boldsymbol{R}_{L,V}(\boldsymbol{Y}\boldsymbol{m}+\boldsymbol{G}\boldsymbol{Q}^{1/2})-(\boldsymbol{Y}\boldsymbol{m}+\boldsymbol{G}\boldsymbol{Q}^{1/2}))\boldsymbol{V}^{-1}(\boldsymbol{R}_{L,V}(\boldsymbol{Y}\boldsymbol{m}+\boldsymbol{G}\boldsymbol{Q}^{1/2})-(\boldsymbol{Y}\boldsymbol{m}+\boldsymbol{G}\boldsymbol{Q}^{1/2}))^\top\right)\end{aligned} \tag{104}$$

where we have introduced the Bregman-envelope [64] with respect to the distance Eq. (20)

$$\mathcal{M}_{L,V}(\boldsymbol{Y}\boldsymbol{m}+\boldsymbol{G}\boldsymbol{Q}^{1/2})=\min_{\boldsymbol{T}}\left\{L(\boldsymbol{T})+\frac{1}{2}\operatorname{tr}\left((\boldsymbol{T}-(\boldsymbol{Y}\boldsymbol{m}+\boldsymbol{G}\boldsymbol{Q}^{1/2}))\boldsymbol{V}^{-1}(\boldsymbol{T}-(\boldsymbol{Y}\boldsymbol{m}+\boldsymbol{G}\boldsymbol{Q}^{1/2}))^\top\right)\right\}. \tag{105}$$

Then, using the state evolution equations from Lemma 5 and Stein's lemma, we can write:

$$\frac{1}{d}L(\boldsymbol{Y}\boldsymbol{m}+\boldsymbol{R}_{\tilde{L},V}(\boldsymbol{G}\boldsymbol{Q}^{1/2}))=\frac{1}{d}\mathcal{M}_{L,V}(\boldsymbol{Y}\boldsymbol{m}+\boldsymbol{G}\boldsymbol{Q}^{1/2})-\frac{1}{2}\operatorname{tr}(\boldsymbol{V}^\top\boldsymbol{Q}) \tag{106}$$

taking the gradient w.r.t. $\boldsymbol{m}$ using the expression for the derivative of a Bregman envelope [64], we get:

$$\partial_{\boldsymbol{m}}L(\boldsymbol{Y}\boldsymbol{m}+\boldsymbol{R}_{\tilde{L},V}(\boldsymbol{G}\boldsymbol{Q}^{1/2}))=\frac{1}{d}\boldsymbol{Y}^\top\left(\boldsymbol{Y}\boldsymbol{m}+\boldsymbol{G}\boldsymbol{Q}^{1/2}-\boldsymbol{R}_{L,V}(\boldsymbol{Y}\boldsymbol{m}+\boldsymbol{G}\boldsymbol{Q}^{1/2})\right)\boldsymbol{V}^{-1} \tag{107}$$

which prescribes, with high probability

$$\hat{\boldsymbol{m}}\overset{\mathrm{P}}{\simeq}\frac{1}{d}\boldsymbol{Y}^\top\left(\boldsymbol{R}_{L,V}(\boldsymbol{Y}\boldsymbol{m}+\boldsymbol{G}\boldsymbol{Q}^{1/2})-\boldsymbol{Y}\boldsymbol{m}+\boldsymbol{G}\boldsymbol{Q}^{1/2}\right)\boldsymbol{V}^{-1} \tag{108}$$

For $\boldsymbol{m}$, we use the block decomposition from Eq.(69), which simplifies the pseudo-inverse Σ^+ in Eq. (100) to give, with high probability

$$\boldsymbol{m}\overset{\mathrm{P}}{\simeq}\frac{1}{\sqrt{d}}\boldsymbol{M}\boldsymbol{\eta}(\boldsymbol{H}\hat{\boldsymbol{Q}}^{1/2}\hat{\boldsymbol{V}}^{-1}) \tag{109}$$

where the function $\boldsymbol{\eta}$ acts on the block diagonal and is defined by Eq.(70). Using those results and the definition of $\tilde{\boldsymbol{W}}$, the solution $\boldsymbol{W}^\star$ and the quantity $\boldsymbol{X}\boldsymbol{W}^*$ are characterized, in the pseudo-Lipschitz sense of Theorem 1, by the fixed point of the system of equations

$$\boldsymbol{Q}_k=\lim_{d\to+\infty}\frac{1}{d}\mathbb{E}\left[\boldsymbol{e}_k(\{\boldsymbol{H}_k(\hat{\boldsymbol{Q}}_k)^{1/2}\hat{\boldsymbol{V}}_k^{-1}\}_{k\in[K]})^\top\boldsymbol{e}_k(\{\boldsymbol{H}_k(\hat{\boldsymbol{Q}}_k)^{1/2}\hat{\boldsymbol{V}}_k^{-1}\}_{k\in[K]})\right]\in\mathbb{R}^{K\times K} \tag{110}$$

$$\hat{\boldsymbol{Q}}_k=\lim_{d\to+\infty}\frac{1}{d}\mathbb{E}\left[\boldsymbol{h}_k(\boldsymbol{G}_k\boldsymbol{Q}_k^{1/2})^\top\boldsymbol{h}_k(\boldsymbol{G}_k\boldsymbol{Q}_k^{1/2})\right]\in\mathbb{R}^{K\times K} \tag{111}$$

$$\boldsymbol{V}_k=\lim_{d\to+\infty}\frac{1}{d}\sum_{i=1}^{d}\mathbb{E}\left[\frac{\partial\boldsymbol{e}_k(\{\boldsymbol{H}_k(\hat{\boldsymbol{Q}}_k)^{1/2}\}_{k\in[K]})}{\partial(\boldsymbol{H}_k(\hat{\boldsymbol{Q}}_k)^{1/2})_i}\right]\in\mathbb{R}^{K\times K} \tag{112}$$

$$\hat{\boldsymbol{V}}_k=-\lim_{d\to+\infty}\frac{1}{d}\sum_{i=1}^{n_k}\mathbb{E}\left[\frac{\partial\boldsymbol{h}_k^t(\boldsymbol{G}_k(\boldsymbol{Q}_k^t)^{1/2})}{\partial(\boldsymbol{G}_k(\boldsymbol{Q}_k)^{1/2})_i}\right]\in\mathbb{R}^{K\times K} \tag{113}$$

$$\boldsymbol{m} = \frac{1}{\sqrt{d}}\mathbb{E}\left[\boldsymbol{M}\boldsymbol{\eta}(\boldsymbol{H}\hat{\boldsymbol{Q}}^{1/2}\hat{\boldsymbol{V}}^{-1})\right] \in \mathbb{R}^{K\times K} \tag{114}$$

$$\hat{\boldsymbol{m}} = \frac{1}{d}\boldsymbol{Y}^{\top}\left(\boldsymbol{R}_{L,V}(\boldsymbol{Y}\boldsymbol{m} + \boldsymbol{G}\boldsymbol{Q}^{1/2}) - \boldsymbol{Y}\boldsymbol{m} + \boldsymbol{G}\boldsymbol{Q}^{1/2}\right)\boldsymbol{V}^{-1} \in \mathbb{R}^{K\times K} \tag{115}$$

Using the explicit form of the different functions given by Eq.(79,80) and Stein's lemma for the derivatives, these equations match those of Theorem 1. Matching the update function $\boldsymbol{h}$ associated to the loss is straightforward using the separability of the loss and the definitions of the aspect ratio α and cluster probabilities ρ_k (the ratios n_k/n converge to ρ_k using the strong law of large numbers and convergence of the product with the Gaussian expectation is justified using Slutsky's lemma). To match the update function associated to the regularizer, it is useful to consider the integral defining its counterpart in the replica calculation Eq.(138). Using Laplace's method on this integral at zero temperature under the replica symmetric ansatz yields the same optimization problem as the one defining the update function $\boldsymbol{\eta}$ in Eq.(80). This completes the proof.

A.5 On the uniqueness of the solution to the fixed point equations (110)

If the optimization problem (49) is strictly convex, the provided proof is enough to characterize its unique solution. For a convex problem, e.g., LASSO with square loss, one can add a vanishing additional ridge penalty as done in [52], to obtain a strictly convex approximation of the original problem. A fully rigorous treatment of the convex (non-strictly) case would require proving that the fixed point equation have a unique solution, as done for instance in [12, 49]. Indeed, if the fixed point equations (110) have a unique solution, then any solution of the convex problem is characterized by the same parameters. It is possible to reconstruct Bregman envelopes on problem (49) for the loss and regularization as we have done for the loss in the previous section. We then can show that the fixed point equations (110) are the optimality condition of a convex-concave problem involving both Bregman envelopes and linear combinations of the order parameters. In the same spirit as [12,49], the authors are confident that this problem is asymptotically strictly convex. This is supported by the simulations presented in the experiments sections. We leave this analysis for a longer version of this paper.

B Replica computation

B.1 Setting of the problem

In this Section we give a full derivation of the results given in Theorem 1 and Theorem 2 by means of the replica approach, a standard method developed in the realm of statistical physics of disordered systems [67]. In the general computation, we will consider the classification problem of K clusters, assuming a dataset $\{(\boldsymbol{x}^{\nu}, \boldsymbol{y}^{\nu})\}_{\nu\in[n]}$ of n independent datapoints where, as in the main text, the labels $\boldsymbol{y}$ takes value in a set of K elements, $\boldsymbol{y}^{\nu} \in \{\boldsymbol{e}_k\}_k$, with $\boldsymbol{e}_k \in \mathbb{R}^L$. The elements of the dataset are independently generated by a mixture density in the form

$$P(\boldsymbol{x}, \boldsymbol{y}) = \sum_{k=1}^{K} \mathbb{I}(\boldsymbol{y} = \boldsymbol{e}_k)\rho_k \mathcal{N}\left(\boldsymbol{x}\,\middle|\,\boldsymbol{\mu}_k, \Sigma_k\right), \quad \sum_{k=1}^{K}\rho_k = 1. \tag{116}$$

We will perform our classification task searching for a set of parameters $(\boldsymbol{W}^{\star}, \boldsymbol{b}^{\star})$ that will allow us to construct an estimator. The parameters will be chosen by minimising an empirical risk function in the

form

$$\mathcal{R}(\boldsymbol{W},\boldsymbol{b}) \equiv \sum_{\nu=1}^{n} \ell\left(\boldsymbol{y}^{\nu}, \frac{\boldsymbol{W}\boldsymbol{x}^{\nu}}{\sqrt{d}} + \boldsymbol{b}\right) + \lambda r(\boldsymbol{W}), \tag{117}$$

i.e., they are given by

$$(\boldsymbol{W}^{\star},\boldsymbol{b}^{\star}) \equiv \underset{\boldsymbol{W}\in\mathbb{R}^{L\times d},\, \boldsymbol{b}\in\mathbb{R}^{L}}{\operatorname{argmin}} \mathcal{R}(\boldsymbol{W},\boldsymbol{b}). \tag{118}$$

We will say that $\boldsymbol{W} \in \mathbb{R}^{L\times d}$ and $\boldsymbol{b} \in \mathbb{R}^{L}$ are the weights and bias to be learned respectively, ℓ is a convex loss function with respect to its second argument, and r is a regularisation function whose strength is tuned by the parameter $\lambda \geq 0$. Finally, we will assume that a classifier $\boldsymbol{\varphi}\colon \mathbb{R}^{L} \to \{\boldsymbol{e}_k\}_k$ is given, such that, once $(\boldsymbol{W}^{\star},\boldsymbol{b}^{\star})$ are obtained, a new point $\boldsymbol{x}$ is assigned to the label

$$\boldsymbol{x} \mapsto \boldsymbol{\varphi}\left(\frac{\boldsymbol{W}^{\star}\boldsymbol{x}}{\sqrt{d}} + \boldsymbol{b}^{\star}\right) \in \{\boldsymbol{e}_k\}_k. \tag{119}$$

The described setting is slightly more general than the one given in Theorem 1. As a consequence of the fact that we choose L-dimensional labels, the order parameters that appear in the computation are L dimensional vectors or $L \times L$ matrices. A typical "high-dimensional encoding" is the one-hot encoding convention adopted in Theorem 1, where $\{\boldsymbol{e}_k\}_k \subset \mathbb{R}^K$ is the canonical basis of $\mathbb{R}^K$. In this case, the adopted classifier is

$$\boldsymbol{\varphi}(\boldsymbol{x}) \equiv \hat{\boldsymbol{y}}(\boldsymbol{x}), \quad \hat{y}_k(\boldsymbol{x}) = \mathbb{I}(\max_{\kappa} x = x_k). \tag{120}$$

Assuming *scalar* labels $\{e_k\}_k \in \mathbb{R}$, we deal with scalar order parameters. For example, in the case of binary classification ($K = 2$) it is common to adopt $L = 1$ and $\{e_1, e_2\} = \{+1, -1\}$. In this case $\varphi(x) = \operatorname{sign}(x)$, see also Section C.2.

B.2 Gibbs minimisation

The problem stated in Section 1 is formulated as an optimisation problem. We can tackle such optimisation problem introducing a Gibbs measure over the weights $(\boldsymbol{W},\boldsymbol{b})$, namely

$$\mu_{\beta}(\boldsymbol{W},\boldsymbol{b}) \propto e^{-\beta\mathcal{R}(\boldsymbol{W},\boldsymbol{b})} = \underbrace{e^{-\beta r(\boldsymbol{W})}}_{P_w(\boldsymbol{W})} \prod_{\nu=1}^{n} \underbrace{\exp\left[-\beta\ell\left(\boldsymbol{y}^{\nu}, \frac{\boldsymbol{W}\boldsymbol{x}^{\nu}}{\sqrt{d}} + \boldsymbol{b}\right)\right]}_{P_y(\boldsymbol{y}|\boldsymbol{W},\boldsymbol{b})}. \tag{121}$$

The parameter $\beta > 0$ is introduced for convenience: in the $\beta \to +\infty$ limit, the Gibbs measure concentrates on the values $(\boldsymbol{W}^{\star},\boldsymbol{b}^{\star})$ which minimize the empirical risk $\mathcal{R}(\boldsymbol{W},\boldsymbol{b})$ and are therefore the goal of the learning process. The functions P_y and P_w can be interpreted as a (unnormalised) likelihood and prior distribution respectively. Our analysis will go through the computation of the average free energy density associated to such Gibbs measure, i.e.,

$$f_{\beta} = -\lim_{\substack{n,d\to+\infty \\ n/d=\alpha}} \mathbb{E}_{\{(\boldsymbol{x},\boldsymbol{y})\}}\left[\frac{\ln \mathcal{Z}_{\beta}}{d\beta}\right], \tag{122}$$

where $\mathbb{E}_{\{(\boldsymbol{x},\boldsymbol{y})\}}[\bullet]$ is the average over the training dataset, and we have introduced the partition function

$$\mathcal{Z}_{\beta} \equiv \int e^{-\beta\mathcal{R}(\boldsymbol{W},\boldsymbol{b})}\mathrm{d}\boldsymbol{W} \tag{123}$$

To perform the computation of such quantity, we use the so-called replica method, i.e., we compute

$$-\lim_{\substack{n,d\to+\infty\\ n/d=\alpha}} \mathbb{E}_{\{(\boldsymbol{x},\boldsymbol{y})\}}\left[\frac{\ln \mathcal{Z}_\beta}{d\beta}\right] = \lim_{\substack{n,d\to+\infty\\ n/d=\alpha}} \lim_{s\to 0} \frac{1-\mathbb{E}_{\{(\boldsymbol{x},\boldsymbol{y})\}}[\mathcal{Z}_\beta^s]}{sd\beta}, \tag{124}$$

B.3 Replica approach

We proceed in our calculation considering the bias vector assuming no prior on $\boldsymbol{b}$, which will play a role of an extra parameter. The equations for the bias $\boldsymbol{b}$ will be derived extremising with respect to it the final result for the free energy. We need to evaluate

$$\mathbb{E}_{\boldsymbol{x},\boldsymbol{y}}[\mathcal{Z}_\beta^s] = \prod_{a=1}^{s}\int \mathrm{d}\boldsymbol{W}^a P_w(\boldsymbol{W}^a)\left(\sum_k \rho_k \mathbb{E}_{\boldsymbol{x}|\boldsymbol{y}=\boldsymbol{e}_k}\left[\prod_{a=1}^{s} P_y\left(\boldsymbol{e}_k\middle|\frac{\boldsymbol{W}^a\boldsymbol{x}}{\sqrt{d}}+\boldsymbol{b}\right)\right]\right)^n. \tag{125}$$

Let us take the inner average introducing a new variable $\boldsymbol{\eta}$,

$$\begin{aligned}\mathbb{E}_{\boldsymbol{x}|\boldsymbol{y}=\boldsymbol{e}_k}\left[\prod_{a=1}^{s} P_y\left(\boldsymbol{e}_k\middle|\frac{\boldsymbol{W}^a\boldsymbol{x}}{\sqrt{d}}+\boldsymbol{b}\right)\right] &= \prod_{a=1}^{s}\int \mathrm{d}\boldsymbol{\eta}^a P_y(\boldsymbol{e}_k|\boldsymbol{\eta}^a)\mathbb{E}_{\boldsymbol{x}}\left[\prod_{a=1}^{s}\delta\left(\boldsymbol{\eta}^a-\frac{\boldsymbol{W}^a\boldsymbol{x}}{\sqrt{d}}+\boldsymbol{b}\right)\right]\\ &= \prod_{a=1}^{s}\int \mathrm{d}\boldsymbol{\eta}^a P_y(\boldsymbol{e}_k|\boldsymbol{\eta}^a)\mathcal{N}\left(\boldsymbol{\eta}\middle|\frac{\boldsymbol{W}^a\boldsymbol{\mu}_k}{\sqrt{d}}-\boldsymbol{b};\frac{\boldsymbol{W}^a\Sigma_k\boldsymbol{W}^{b\top}}{d}\right).\end{aligned} \tag{126}$$

We can write then

$$\begin{aligned}\mathbb{E}_{\boldsymbol{x},\boldsymbol{y}}[\mathcal{Z}_\beta^s] &= \prod_{a=1}^{n}\int \mathrm{d}\boldsymbol{W}^a P_w(\boldsymbol{W}^a)\left(\sum_k \rho_k \prod_{a=1}^{s}\int \mathrm{d}\boldsymbol{\eta}^a P_y(\boldsymbol{e}_k|\boldsymbol{\eta}^a)\mathcal{N}\left(\boldsymbol{\eta};\frac{\boldsymbol{W}^a\boldsymbol{\mu}_k}{d}+\boldsymbol{b};\frac{\boldsymbol{W}^a\Sigma_k\boldsymbol{W}^{b\top}}{d}\right)\right)^n\\ &= \left(\prod_{k=1}^{K}\prod_{a\le b}\iint \frac{\mathrm{d}\boldsymbol{Q}_k^{ab}\mathrm{d}\hat{\boldsymbol{Q}}_k^{ab}}{(2\pi)^{L^2/2}}\right)\left(\prod_k\prod_a\int\frac{\mathrm{d}\boldsymbol{m}_k^a\mathrm{d}\hat{\boldsymbol{m}}_k^a}{(2\pi)^{L/2}}\right)e^{-d\beta\Phi^{(s)}}.\end{aligned} \tag{127}$$

where we introduced the *order parameters*

$$\boldsymbol{Q}_k^{ab} = \frac{\boldsymbol{W}^a\Sigma_k\boldsymbol{W}^{b\top}}{d}\in\mathbb{R}^{L\times L},\quad a,b=1,\dots,s, \tag{128}$$

$$\boldsymbol{m}_k^a = \frac{\boldsymbol{W}^a\boldsymbol{\mu}_k}{\sqrt{d}}\in\mathbb{R}^{L},\quad a=1,\dots,s, \tag{129}$$

and the replicated free-energy

$$\begin{aligned}\beta\Phi^{(s)}(\boldsymbol{Q},\boldsymbol{m},\hat{\boldsymbol{Q}},\hat{\boldsymbol{m}},\boldsymbol{b}) &= \sum_{k=1}^{K}\sum_a \hat{\boldsymbol{m}}_k^{a\top}\boldsymbol{m}_k^a + \sum_{k=1}^{K}\sum_{a\le b}\mathrm{tr}\left[\hat{\boldsymbol{Q}}_k^{ab\top}\boldsymbol{Q}_k^{ab}\right]\\ &\quad -\frac{1}{d}\ln\prod_{a=1}^{s}\int P_w(\boldsymbol{W}^a)\mathrm{d}\boldsymbol{W}^a\prod_k\left(\prod_{a\le b}e^{\mathrm{tr}\left[\hat{\boldsymbol{Q}}_k^{ab\top}\boldsymbol{W}^a\Sigma_k\boldsymbol{W}^{b\top}\right]}\prod_a e^{\sqrt{d}\hat{\boldsymbol{m}}_k^{a\top}\boldsymbol{W}^a\boldsymbol{\mu}_k}\right)\\ &\quad -\alpha\ln\sum_k\rho_k\prod_{a=1}^{s}\int\mathrm{d}\boldsymbol{\eta}^a P_y(\boldsymbol{e}_k|\boldsymbol{\eta}^a)\mathcal{N}\left(\boldsymbol{\eta}\middle|\boldsymbol{m}_k^a+\boldsymbol{b},\boldsymbol{Q}_k^{ab}\right).\end{aligned} \tag{130}$$

At this point, the free energy f_β should be computed extremisizing with respect to all the order parameters by virtue of the Laplace approximation (in addition to $\boldsymbol{b}$),

$$f_\beta = \lim_{s\to 0} \operatorname*{Extr}_{\{\boldsymbol{m},\boldsymbol{Q},\hat{\boldsymbol{m}},\hat{\boldsymbol{Q}}\},\boldsymbol{b}} \frac{\Phi^{(s)}(\boldsymbol{Q},\boldsymbol{m},\hat{\boldsymbol{Q}},\hat{\boldsymbol{m}},\boldsymbol{b})}{s}. \tag{131}$$

However, the convexity of the problem allows us to make an important simplification.

Replica symmetric ansatz — Before taking the $s \to 0$ limit we make the assumptions

$$\begin{aligned} Q_k^{aa} &= \begin{cases} R_k, & a = b \\ Q_k & a \neq b \end{cases} & \hat{Q}_k^{aa} &= \begin{cases} -\frac{1}{2}R_k, & a = b \\ \hat{Q}_k & a \neq b \end{cases} \\ m_k^a &= m_k & \hat{m}_k^a &= \hat{m}_k \quad \forall a \end{aligned} \tag{132}$$

This ansatz is justified by the fact that we are assuming ℓ and r to be convex, and $\lambda > 0$. In this case, the problem admit one solution only that, therefore, coincide with the replica symmetric solution, in which overlaps between two replicas do not depend on the chosen replicas. By means of the replica symmetric hypotesis, we can write

$$Q_k^{ab} \mapsto \mathbf{Q}_k \equiv \boldsymbol{I}_{s,s} \otimes (\boldsymbol{R}_k - \boldsymbol{Q}_k) + \mathbf{1}_s \otimes \boldsymbol{Q}_k. \tag{133}$$

The inverse matrix is therefore

$$\mathbf{Q}_k^{-1} = \mathbf{1}_s \otimes (\boldsymbol{R}_k - \boldsymbol{Q}_k)^{-1} - \boldsymbol{I}_{s,s} \otimes [(\boldsymbol{R}_k + (s-1)\boldsymbol{Q}_k)^{-1}\boldsymbol{Q}_k(\boldsymbol{R}_k - \boldsymbol{Q}_k)^{-1}], \tag{134}$$

whereas

$$\begin{aligned} \det \mathbf{Q}_k &= \det\big(\boldsymbol{R}_k - \boldsymbol{Q}_k\big)^{s-1} \det\big(\boldsymbol{R}_k + (s-1)\boldsymbol{Q}_k\big) \\ &= 1 + s \ln \det\big(\boldsymbol{R}_k - \boldsymbol{Q}_k\big) + s \operatorname{tr}\left[(\boldsymbol{R}_k - \boldsymbol{Q}_k)^{-1}\boldsymbol{Q}_k\right] + o(s). \end{aligned} \tag{135}$$

If we denote $\boldsymbol{V}_k \equiv \boldsymbol{R}_k - \boldsymbol{Q}_k$

$$\begin{aligned} &\ln \sum_k \rho_k \prod_{a=1}^{s} \int \mathrm{d}\boldsymbol{\eta}^a P_y(\boldsymbol{e}_k|\boldsymbol{\eta}^a) \mathcal{N}\left(\boldsymbol{\eta}\middle|\boldsymbol{m}_k^a + \boldsymbol{b}, \boldsymbol{Q}_k^{ab}\right) \\ &\qquad = s \sum_k \rho_k \mathbb{E}_{\boldsymbol{\xi}} \ln \left(\int \frac{\mathrm{d}\boldsymbol{\eta} P_y(\boldsymbol{e}_k|\boldsymbol{\eta})}{\sqrt{\det(2\pi \boldsymbol{V}_k)}} e^{-\frac{1}{2}(\boldsymbol{\eta} - \boldsymbol{m}_k - \boldsymbol{b} - \boldsymbol{Q}_k^{1/2}\boldsymbol{\xi})^\top \boldsymbol{V}_k^{-1}(\boldsymbol{\eta} - \boldsymbol{b} - \boldsymbol{m}_k - \boldsymbol{Q}_k^{1/2}\boldsymbol{\xi})} \right) + o(s) \\ &\qquad\qquad = s \sum_k \rho_k \mathbb{E}_{\boldsymbol{\xi}} \left[\ln Z\left(\boldsymbol{e}_k, \boldsymbol{m}_k + \boldsymbol{b} + \boldsymbol{Q}_k^{1/2}\boldsymbol{\xi}, \boldsymbol{V}_k\right) \right] + o(s), \end{aligned} \tag{136}$$

with $\boldsymbol{\xi} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_L)$ is a normally distributed vector and we have introduced the function

$$Z\left(\boldsymbol{e}_k, \boldsymbol{m}, \boldsymbol{V}\right) \equiv \int \frac{\mathrm{d}\boldsymbol{\eta} P_y(\boldsymbol{e}_k|\boldsymbol{\eta})}{\sqrt{\det(2\pi \boldsymbol{V})}} e^{-\frac{1}{2}(\boldsymbol{\eta} - \boldsymbol{m})^\top \boldsymbol{V}^{-1}(\boldsymbol{\eta} - \boldsymbol{m})} \tag{137}$$

On the other hand, denoting by $\hat{\boldsymbol{V}}_k = \hat{\boldsymbol{R}}_k + \hat{\boldsymbol{Q}}_k$,

$$\frac{1}{d}\ln\prod_{a=1}^{s}\left(\int P_w(\boldsymbol{W}^a)\mathrm{d}\boldsymbol{W}^a\prod_k e^{-\frac{1}{2}\mathrm{tr}\left[\hat{\boldsymbol{V}}_k^\top \boldsymbol{W}^a\Sigma_k(\boldsymbol{W}^a)^\top\right]+\sqrt{d}\hat{\boldsymbol{m}}_k^\top\boldsymbol{W}^a\boldsymbol{\mu}_k}\prod_{b,k}e^{\frac{1}{2}\mathrm{tr}\left[\hat{\boldsymbol{Q}}_k\boldsymbol{W}^a\Sigma_k(\boldsymbol{W}^b)^\top\right]}\right)=$$

$$=\frac{s}{d}\mathbb{E}_\Xi\ln\left[\int P_w(\boldsymbol{W})\mathrm{d}\boldsymbol{W}\prod_k\exp\left(-\frac{\mathrm{tr}\left[\hat{\boldsymbol{V}}_k^\top\boldsymbol{W}\Sigma_k\boldsymbol{W}^\top\right]}{2}+\sqrt{d}\hat{\boldsymbol{m}}_k^\top\boldsymbol{W}\boldsymbol{\mu}_k+\Xi_k\odot\sqrt{\hat{\boldsymbol{Q}}_k\otimes\Sigma_k}\odot\boldsymbol{W}\right)\right]+o(s). \quad (138)$$

In the expression above we have used the tensorial product $\hat{\boldsymbol{Q}}\otimes\Sigma=(\hat{Q}_{kk'}\Sigma_{ij})_{ki,k'j'}$. Given a matrix $\boldsymbol{B}\in\mathbb{R}^{L\times d}$ and the tensors $\mathbf{A},\mathbf{A}'\in\mathbb{R}^{L\times d}\otimes\mathbb{R}^{L\times d}$, we denote $(\boldsymbol{B}\odot\mathbf{A})_{ki}\equiv\sum_{k'i'}B_{k'i'}A_{k'i'ki}\in\mathbb{R}^{L\times d}$, $(\mathbf{A}\odot\boldsymbol{B})_{ki}\equiv\sum_{k'i'}A_{kik'i'}B_{k'i'}\in\mathbb{R}^{L\times d}$ and $(\mathbf{A}\odot\mathbf{A}')_{kik'i'}=\sum_{\kappa j}A_{ki\kappa j}A_{\kappa jk'i'}$. In this way, we define $\sqrt{\mathbf{A}}$ as the tensor such that $\mathbf{A}=\sqrt{\mathbf{A}}\odot\sqrt{\mathbf{A}}$. Finally, we have also introduced a set of k matrices $\Xi_k\in\mathbb{R}^{L\times d}$ with i.i.d. random Gaussian entries with zero mean and variance 1, and the average over them $\mathbb{E}_\Xi[\bullet]$. Therefore, the (replicated) *replica symmetric* free-energy is given by

$$\begin{aligned}\lim_{s\to0}\frac{\beta}{s}\Phi_{\mathrm{RS}}^{(s)}=&\sum_{k=1}^K\hat{\boldsymbol{m}}_k^\top\boldsymbol{m}_k+\frac{1}{2}\sum_{k=1}^K\mathrm{tr}\left[\hat{\boldsymbol{V}}_k^\top\boldsymbol{Q}_k\right]-\frac{1}{2}\sum_{k=1}^K\mathrm{tr}\left[\hat{\boldsymbol{Q}}_k^\top\boldsymbol{V}_k\right]-\frac{1}{2}\sum_{k=1}^K\mathrm{tr}\left[\hat{\boldsymbol{V}}_k^\top\boldsymbol{V}_k\right]\\&-\alpha\beta\Psi_{\mathrm{out}}(\boldsymbol{m},\boldsymbol{Q},\boldsymbol{V})-\beta\Psi_w(\hat{\boldsymbol{m}},\hat{\boldsymbol{Q}},\hat{\boldsymbol{V}})\end{aligned} \quad (139)$$

where we have defined two contributions

$$\Psi_{\mathrm{out}}(\boldsymbol{m},\boldsymbol{Q},\boldsymbol{V})\equiv\beta^{-1}\sum_k\rho_k\mathbb{E}_{\xi_k}\ln Z\left(\boldsymbol{e}_k,\boldsymbol{\omega}_k,\boldsymbol{V}_k\right) \quad (140)$$

$$\Psi_w(\hat{\boldsymbol{m}},\hat{\boldsymbol{Q}},\hat{\boldsymbol{V}})\equiv\frac{1}{\beta d}\mathbb{E}_\xi\ln\left(\int P_w(\boldsymbol{W})\mathrm{d}\boldsymbol{W}\prod_k e^{-\frac{\mathrm{tr}\left[\hat{\boldsymbol{V}}_k^\top\boldsymbol{W}\Sigma_k\boldsymbol{W}^\top\right]}{2}+\sqrt{d}\hat{\boldsymbol{m}}_k^\top\boldsymbol{W}\boldsymbol{\mu}_k+\Xi_k:\sqrt{\hat{\boldsymbol{Q}}_k\otimes\Sigma_k}\odot\boldsymbol{W}}\right) \quad (141)$$

and introduced, for future convenience,

$$\boldsymbol{\omega}_k\equiv\boldsymbol{m}_k+\boldsymbol{b}+\boldsymbol{Q}_k^{1/2}\boldsymbol{\xi}_k. \quad (142)$$

Note that we have separated the contribution coming from the chosen loss (the so-called *channel* part Ψ_{out}) from the contribution depending on the regularisation (the *prior* part Ψ_w). To write down the saddle-point equations in the $\beta\to+\infty$ limit, let us first rescale our order parameters as $\hat{\boldsymbol{m}}_k\mapsto\beta\hat{\boldsymbol{m}}_k$, $\hat{\boldsymbol{Q}}_k\mapsto\beta^2\hat{\boldsymbol{Q}}_k$, $\hat{\boldsymbol{V}}_k\mapsto\beta\hat{\boldsymbol{V}}_k$ and $\boldsymbol{V}_k\mapsto\beta^{-1}\boldsymbol{V}_k$. For $\beta\to+\infty$ the channel part is

$$\Psi_{\mathrm{out}}(\boldsymbol{m},\boldsymbol{Q},\boldsymbol{V})=-\sum_k\rho_k\mathbb{E}_{\boldsymbol{\xi}}\left[\mathcal{M}_{\ell(\boldsymbol{e}_k,\boldsymbol{V}_k^{1/2}\bullet)}\left(\boldsymbol{V}_k^{-1/2}\boldsymbol{\omega}_k\right)\right]. \quad (143)$$

Here and in the following the quantity

$$\mathcal{M}_{f(\bullet)}(\boldsymbol{u})\equiv\min_{\boldsymbol{v}\in\mathrm{domain}(\boldsymbol{v})}\left[\frac{1}{2}\|\boldsymbol{v}-\boldsymbol{u}\|_{\mathrm{F}}^2+f(\boldsymbol{v})\right] \quad (144)$$

is the Moreau envelope of $f\colon\mathrm{domain}(\boldsymbol{v})\to\mathbb{R}$, whereas $\|\bullet\|_{\mathrm{F}}$ is the Frobenius norm. We can write the contribution Ψ_{out} in terms of a proximal

$$\boldsymbol{h}_k=\boldsymbol{V}_k^{1/2}\mathrm{Prox}_{\ell(\boldsymbol{e}_k,\boldsymbol{V}^{1/2}\bullet)}(\boldsymbol{V}_k^{-1/2}\boldsymbol{\omega}_k)\equiv\boldsymbol{V}_k^{1/2}\arg\min_{\boldsymbol{u}\in\mathbb{R}^L}\left[\frac{1}{2}\|\boldsymbol{u}-\boldsymbol{V}_k^{-1/2}\boldsymbol{\omega}_k\|_{\mathrm{F}}^2+\ell(\boldsymbol{e}_k,\boldsymbol{V}^{1/2}\boldsymbol{u})\right]. \quad (145)$$

as

$$\Psi_{\text{out}}(\boldsymbol{m},\boldsymbol{Q},\boldsymbol{V}) = -\sum_k \rho_k \mathbb{E}_{\xi}\left[\frac{1}{2}\|\boldsymbol{V}^{-1/2}\boldsymbol{h}_k - \boldsymbol{V}^{-1/2}\boldsymbol{\omega}_k\|_{\text{F}}^2 + \ell(\boldsymbol{e}_k,\boldsymbol{h}_k)\right] \tag{146}$$

A similar expression can be obtained for Ψ_w. Defining

$$\mathbf{A} = \left(\sum_k \hat{\boldsymbol{V}}_k \otimes \Sigma_k\right)^{-1}, \qquad \boldsymbol{B} = \sqrt{d}\sum_k \boldsymbol{\mu}_k \hat{\boldsymbol{m}}_k^\top + \sum_k \Xi_k \odot \sqrt{\hat{\boldsymbol{Q}}_k \otimes \Sigma_k}. \tag{147}$$

Ψ_w can be written as

$$\Psi_w(\hat{\boldsymbol{m}},\hat{\boldsymbol{Q}},\hat{\boldsymbol{V}}) = \frac{1}{2d}\mathbb{E}_{\xi}\left[\boldsymbol{B}\odot\mathbf{A}\odot\boldsymbol{B}\right] + \frac{1}{\beta d}\mathbb{E}_{\xi}\ln\left[\int \mathrm{d}\boldsymbol{W}\exp\left(-\frac{\beta}{2}\|\mathbf{A}^{-1/2}\odot\boldsymbol{W} - \mathbf{A}^{1/2}\odot\boldsymbol{B}\|_{\text{F}}^2 - \beta r(\boldsymbol{W})\right)\right]. \tag{148}$$

It follows that, for $\beta \to +\infty$,

$$\Psi_w(\hat{\boldsymbol{m}},\hat{\boldsymbol{Q}},\hat{\boldsymbol{V}}) = \frac{1}{2d}\mathbb{E}_{\xi}\left[\boldsymbol{B}\odot\mathbf{A}\odot\boldsymbol{B}\right] - \frac{1}{d}\mathbb{E}_{\xi}\left[\mathcal{M}_{r(\mathbf{A}^{1/2}\odot\bullet)}(\mathbf{A}^{1/2}\odot\boldsymbol{B})\right]. \tag{149}$$

As before, let us introduce the proximal

$$\boldsymbol{G} = \mathbf{A}^{1/2}\odot\mathrm{Prox}_{r(\mathbf{A}^{1/2}\odot\bullet)}(\mathbf{A}^{1/2}\odot\boldsymbol{B}) \in \mathbb{R}^{L\times d} \tag{150}$$

We can rewrite the prior contribution Ψ_w as

$$\Psi_w(\hat{\boldsymbol{m}},\hat{\boldsymbol{Q}},\hat{\boldsymbol{V}}) = \frac{1}{2d}\mathbb{E}_{\Xi}\left[\boldsymbol{B}\odot\mathbf{A}\odot\boldsymbol{B}\right] - \frac{1}{d}\mathbb{E}_{\Xi}\left[\frac{\|\mathbf{A}^{-1/2}\odot\boldsymbol{G} - \mathbf{A}^{1/2}\odot\boldsymbol{B}\|_{\text{F}}^2}{2} + r(\boldsymbol{G})\right]. \tag{151}$$

The parallelism between the two contributions is evident, aside from the different dimensionality of the involved objects. The replica symmetric free energy in the $\beta \to +\infty$ limit is computed extremising with respect to the introduced order parameters,

$$f_{\text{RS}} = \underset{\substack{\boldsymbol{m},\boldsymbol{Q},\boldsymbol{V},\boldsymbol{b}\\ \hat{\boldsymbol{m}},\hat{\boldsymbol{Q}},\hat{\boldsymbol{V}}}}{\text{Extr}}\left[\sum_{k=1}^K \hat{\boldsymbol{m}}_k^\top\boldsymbol{m}_k + \frac{1}{2}\sum_{k=1}^K \mathrm{tr}\left[\hat{\boldsymbol{V}}_k^\top\boldsymbol{Q}_k\right] - \frac{1}{2}\sum_{k=1}^K \mathrm{tr}\left[\hat{\boldsymbol{Q}}_k^\top\boldsymbol{V}_k\right] - \frac{1}{2}\sum_{k=1}^K \mathrm{tr}\left[\hat{\boldsymbol{V}}_k^\top\boldsymbol{V}_k\right] - \alpha\Psi_{\text{out}}(\boldsymbol{m},\boldsymbol{Q},\boldsymbol{V}) - \Psi_w(\hat{\boldsymbol{m}},\hat{\boldsymbol{Q}},\hat{\boldsymbol{V}})\right]. \tag{152}$$

To do so, we have to write down a set of saddle-point equations and solve them.

Saddle-point equations — The saddle-point equations are derived straightforwardly from the obtained free energy extremising with respect to all parameters. A first set of equations is obtained from Ψ_{out} as[1]

$$\hat{\boldsymbol{Q}}_k = \alpha\rho_k\mathbb{E}_{\xi}\left[\boldsymbol{f}_k\boldsymbol{f}_k^\top\right], \tag{153a}$$

[1] To obtain the equation for $\hat{\boldsymbol{V}}$ it is convenient to use Stein's lemma, so that $\mathbb{E}[\partial_{\xi}\boldsymbol{f}_k] = \mathbb{E}[\boldsymbol{f}_k\xi^\top]$.

$$\hat{\boldsymbol{V}}_k = -\alpha\rho_k \boldsymbol{Q}_k^{-1/2}\mathbb{E}_{\boldsymbol{\xi}}\left[\boldsymbol{f}_k\boldsymbol{\xi}^\top\right], \tag{153b}$$

$$\hat{\boldsymbol{m}}_k = \alpha\rho_k \mathbb{E}_{\boldsymbol{\xi}}\left[\boldsymbol{f}_k\right], \tag{153c}$$

$$\boldsymbol{b} = \sum_k \rho_k \mathbb{E}_{\boldsymbol{\xi}}\left[\boldsymbol{h}_k - \boldsymbol{m}_k\right] \Longleftrightarrow \sum_k \rho_k \mathbb{E}_{\boldsymbol{\xi}}\left[\boldsymbol{V}_k \boldsymbol{f}_k\right] = \boldsymbol{0}. \tag{153d}$$

where for brevity we have denoted

$$\boldsymbol{f}_k \equiv \boldsymbol{V}_k^{-1}(\boldsymbol{h}_k - \boldsymbol{\omega}_k). \tag{154}$$

Similarly, the saddle-point equations from Ψ_{out} are

$$\boldsymbol{V}_k = \frac{1}{d}\mathbb{E}_{\Xi}\left[\left(\boldsymbol{G}\odot\left(\hat{\boldsymbol{Q}}_k\otimes\Sigma_k\right)^{-1/2}\odot\left(\boldsymbol{I}_k\otimes\Sigma_k\right)\right)\Xi_k^\top\right] \tag{155a}$$

$$\boldsymbol{Q}_k = \frac{1}{d}\mathbb{E}_{\boldsymbol{\xi}}\left[\boldsymbol{G}\Sigma_k\boldsymbol{G}^\top\right] \tag{155b}$$

$$\boldsymbol{m}_k = \frac{1}{\sqrt{d}}\mathbb{E}_{\boldsymbol{\xi}}\left[\boldsymbol{G}\boldsymbol{\mu}_k\right] \tag{155c}$$

To obtain the replica symmetric free energy, therefore, the given set of equation has to be solved, and the result then plugged in Eq. (152). No further simplification can be obtained in the most general setting. We will explore however some simple (but important) applications in Appendix C. Before going on, however, it is important to express the relevant quantities for learning, i.e., the training and generalization errors, in terms of the obtained order parameters.

B.4 Training and test errors

The order parameters introduced to solve the problem allow us to reach our ultimate goal of computing the average errors of the learning process. We will start from the estimation of the training loss. The complication in computing this quantity is that the order parameters found in the learning process are, of course, correlated with the dataset used for the learning itself. We need to compute

$$\epsilon_\ell \equiv \frac{1}{n}\sum_{\nu=1}^{n}\ell\left(\boldsymbol{y}^\nu, \frac{\boldsymbol{W}^\star\boldsymbol{x}^\nu}{\sqrt{d}} + \boldsymbol{b}^\star\right) \tag{156}$$

in the $n\to+\infty$ limit. Denoting for brevity $\ell_k(\boldsymbol{x}) \equiv \ell(\boldsymbol{e}_k, \boldsymbol{x})$, the best way to proceed is to observe that $\mathbb{E}_{\{(\boldsymbol{y}^\nu,\boldsymbol{x}^\nu)\}_\nu}[\mathcal{R}(\boldsymbol{W}^\star,\boldsymbol{b}^\star)] = -\lim_{\beta\to+\infty}\mathbb{E}_{\{(\boldsymbol{y}^\nu,\boldsymbol{x}^\nu)\}_\nu}[\partial_\beta \ln \mathcal{Z}_\beta] = \lambda\mathbb{E}_{\{(\boldsymbol{y}^\nu,\boldsymbol{x}^\nu)\}_\nu}[r(\boldsymbol{W}^\star)] + \epsilon_\ell$, where

$$\epsilon_\ell = -\lim_{\beta\to+\infty}\partial_\beta(\beta\Psi_{\text{out}}) = \lim_{\beta\to+\infty}\sum_k \rho_k \int \ell_k(\boldsymbol{\eta})\frac{e^{-\frac{\beta}{2}(\boldsymbol{\eta}-\boldsymbol{m}^\star)^\top \boldsymbol{V}^{\star\,-1}(\boldsymbol{\eta}-\boldsymbol{m}^\star)-\beta\ell_k(\boldsymbol{\eta})}}{\sqrt{\det\left(2\pi\beta^{-1}\boldsymbol{V}^\star\right)}Z(\boldsymbol{e}_k,\boldsymbol{\omega}_k^\star,\beta^{-1}\boldsymbol{V}_k^\star)}\mathrm{d}\boldsymbol{\eta}. \tag{157}$$

In the $\beta\to+\infty$ limit, the integral concentrates on the minimizer of the exponent, that is, by definition, the proximal $\boldsymbol{h}_k$. In conclusion, $\epsilon_\ell = \sum_k \rho_k \mathbb{E}[\ell(\boldsymbol{h}_k)]$.

By means of the same concentration result, the training error is

$$\epsilon_t = \frac{1}{n}\sum_{\nu=1}^{n}\mathbb{I}\left(\boldsymbol{\varphi}\left(\frac{\boldsymbol{W}^\star\boldsymbol{x}^\nu}{\sqrt{d}} + \boldsymbol{b}^\star\right) \neq \boldsymbol{y}^\nu\right) \xrightarrow{n\to+\infty} \sum_{k=1}^{K}\rho_k\mathbb{E}_{\boldsymbol{\xi}}\left[\mathbb{I}(\boldsymbol{\varphi}(\boldsymbol{h}_k)\neq\boldsymbol{e}_k)\right]. \tag{158}$$

The expressions above hold in general, but, as anticipated, important simplifications can occur in the set of saddle-point equations (153) and (155) depending on the choice of the loss ℓ and of the regularization function r.

The generalisation (or test) error can be written instead as

$$\epsilon_g = \mathbb{E}_{\boldsymbol{y}^{\text{new}},\boldsymbol{x}^{\text{new}}}\left[\mathbb{I}\left(\boldsymbol{\varphi}\left(\frac{\boldsymbol{W}^\star \boldsymbol{x}^{\text{new}}}{\sqrt{d}} + \boldsymbol{b}^\star\right) \neq \boldsymbol{y}^{\text{new}}\right)\right]. \tag{159}$$

This expression can be rewritten as

$$\epsilon_g = \sum_k \rho_k \int \mathbb{I}(\boldsymbol{\varphi}(\boldsymbol{\eta}) = \boldsymbol{e}_k)\mathbb{E}_{\boldsymbol{x}^{\text{new}}}\left[\delta\left(\boldsymbol{\eta} - \frac{\boldsymbol{W}^\star \boldsymbol{x}^{\text{new}}}{\sqrt{d}} - \boldsymbol{b}^\star\right)\right] \mathrm{d}\boldsymbol{\eta} \tag{160}$$

Once again, we write

$$\mathbb{E}_{\boldsymbol{x}^{\text{new}}}\left[\delta\left(\boldsymbol{\eta} - \frac{\boldsymbol{W}^\star \boldsymbol{x}^{\text{new}}}{\sqrt{d}} - \boldsymbol{b}^\star\right)\right] \xrightarrow{d\to+\infty} \mathcal{N}(\boldsymbol{\eta}|\boldsymbol{m}_k^\star + \boldsymbol{b}^\star, \boldsymbol{Q}_k^\star) \tag{161}$$

so that

$$\epsilon_g = \sum_{k=1}^{K} \rho_k \mathbb{E}_{\boldsymbol{\xi}}\left[\mathbb{I}\left(\boldsymbol{\varphi}\left(\boldsymbol{m}_k^\star + \boldsymbol{Q}_k^{\star\,1/2}\boldsymbol{\xi} + \boldsymbol{b}^\star\right) \neq \boldsymbol{e}_k\right)\right]. \tag{162}$$

This can be easily computed numerically once that the order parameters are given.

B.5 A note on the numerical integration of the saddle-point equations

To estimate ϵ_g, ϵ_t and ϵ_ℓ we first need to find the fixed-point solutions of the saddle-point equations (153) and (155). The simplest numerical strategy consists in updating, in a self-consistent way, the order parameters until their variation according to, e.g., the Frobenius norm is smaller than a given threshold value. The convergence to the the correct fixed point is guaranteed (in principle) by the convexity of the problem. This is the strategy that we followed to solve the problem. However, a few delicate aspects have to be taken into account in this update process.

1. In the most general case, the update rules given by the saddle-point equations (153) and (155) require an average over random matrices Ξ_k and vectors $\boldsymbol{\xi}_k$ with i.i.d. Gaussian entries. In our code, we tackled this problem using a Monte Carlo algorithm whenever an analytic integration was not possible.

2. The update requires the computation of the proximals $\boldsymbol{G}$ and $\boldsymbol{h}_k$. Such computations can be performed analytically in some specific cases only (for example, in the case of ridge regression). The existence of a unique solution is guaranteed by the convexity of the problem. In our study of the cross-entropy loss function, for example, we computed the proximals $\boldsymbol{h}_k$ numerically solving Eq. (178). In this problem, however, additional numerical instabilities emerged in the $\lambda \to 0$ limit, due the fact that the discontinuity in the gradient appear, see Eq. (182). We solved this issue performing an annealing in λ, i.e., solving for the proximal for decreasing values of the regularization strength.

3. The numerical solution of the saddle-point equations might suffer numerical instabilities due to the operations of inversion involved, see, e.g., the equation for $\hat{\boldsymbol{V}}_k$ in (153), which requires the inversion of $\boldsymbol{Q}_k$. It is convenient, in such cases, to rewrite the equation in an equivalent form which is numerically more stable. For example, in the aforementioned equation, we can observe that $\boldsymbol{f}_k$ satisfies the equation $\boldsymbol{f}_k + \partial_{\boldsymbol{x}}\ell_k(\boldsymbol{V}_k\boldsymbol{f}_k + \boldsymbol{\omega}_k) = \boldsymbol{0}$ so that $\partial_{\boldsymbol{\omega}_k}\boldsymbol{f}_k = -(\boldsymbol{I}_K + \partial^2_{\boldsymbol{x}}\ell_k(\boldsymbol{V}_k\boldsymbol{f}_k + \boldsymbol{\omega}_k)\boldsymbol{V}_k)^{-1}\partial^2_{\boldsymbol{x}}\ell_k(\boldsymbol{V}_k\boldsymbol{f}_k + \boldsymbol{\omega}_k)$. Using Stein's lemma,

$$\hat{\boldsymbol{V}}_k = -\alpha\rho_k\mathbb{E}_{\boldsymbol{\xi}}\left[\partial_{\boldsymbol{\xi}}\boldsymbol{f}_k\right] = \alpha\rho_k\mathbb{E}_{\boldsymbol{\xi}}\left[\left(\boldsymbol{I}_K + \partial^2_{\boldsymbol{x}}\ell_k(\boldsymbol{V}_k\boldsymbol{f}_k + \boldsymbol{\omega}_k)\boldsymbol{V}_k\right)^{-1}\partial^2_{\boldsymbol{x}}\ell_k(\boldsymbol{V}_k\boldsymbol{f}_k + \boldsymbol{\omega}_k)\right]. \tag{163}$$

We found this equation numerically more stable than the one given in (153) when dealing with the cross-entropy loss.

C Some relevant particular cases

In this Appendix, we will specify the saddle-point equations for the multiclass classification problem for different choices of the loss function ℓ and of the regularisation function r. From the analysis developed in the previous Appendices, it is clear that the choices of ℓ and r impact separately the set of equations (153) and (155) respectively. Once the order parameters are found, it is possible to estimate the training and generalisation errors as, for example, in Section B.4.

C.1 The case of ℓ_2 regularization

In this Section we consider the relevant case of quadratic regularization, $r(\boldsymbol{W}) = 1/2\|\boldsymbol{W}\|_{\mathrm{F}}^2$. In this case the computation of Ψ_w can be performed explicitly via a Gaussian integration,

$$\begin{aligned}\frac{1}{\beta}\Psi_w(\hat{\boldsymbol{m}},\hat{\boldsymbol{Q}},\hat{\boldsymbol{V}}) = &-\frac{1}{2d}\operatorname{tr}\ln\left(\lambda \boldsymbol{I}_K\otimes \boldsymbol{I}_d + \sum_\kappa \hat{\boldsymbol{V}}_\kappa\otimes\Sigma_\kappa\right) - \frac{K\ln\beta}{2\beta}\\ &+\frac{1}{2}\operatorname{tr}\left[\left(\lambda \boldsymbol{I}_K\otimes \boldsymbol{I}_d + \sum_\kappa \hat{\boldsymbol{V}}_\kappa\otimes\Sigma_\kappa\right)^{-1}\odot\left(\sum_{kk'}\hat{\boldsymbol{m}}_k\hat{\boldsymbol{m}}_{k'}^\top\otimes\boldsymbol{\mu}_k\boldsymbol{\mu}_{k'}^\top + \frac{1}{d}\sum_k\hat{\boldsymbol{Q}}_k\otimes\Sigma_k\right)\right].\end{aligned} \tag{164}$$

This form of Ψ_w allows us to write in a simpler way the set of Eqs. (155), that can be re-written as

$$\begin{aligned}\boldsymbol{Q}_k &= \operatorname{tr}_d\left[(\boldsymbol{I}_K\otimes\Sigma_k)\odot\mathbf{S}\odot\left(\sum_{kk'}\hat{\boldsymbol{m}}_k\hat{\boldsymbol{m}}_{k'}^\top\otimes\boldsymbol{\mu}_k\boldsymbol{\mu}_{k'}^\top + \frac{1}{d}\sum_\kappa\hat{\boldsymbol{Q}}_\kappa\otimes\Sigma_\kappa\right)\odot\mathbf{S}\right]\\ \boldsymbol{m}_k &= \sum_{k'}\operatorname{tr}_d\left[\mathbf{S}\odot\left(\hat{\boldsymbol{m}}_{k'}\otimes\boldsymbol{\mu}_{k'}\boldsymbol{\mu}_k^\top\right)\right]\\ \boldsymbol{V}_k &= \frac{1}{d}\operatorname{tr}_d\left[(\boldsymbol{I}_K\otimes\Sigma_k)\odot\mathbf{S}\right],\end{aligned} \tag{165}$$

where we have introduced, for notation compactness,

$$\mathbf{S}\equiv\left(\lambda \boldsymbol{I}_K\otimes \boldsymbol{I}_d + \sum_\kappa \hat{\boldsymbol{V}}_\kappa\otimes\Sigma_\kappa\right)^{-1} \tag{166}$$

In the previous equations, by tr_d we denoted the trace with respect to the components living in the d-dimensional space of the dataset.

Jointly diagonal covariances — Suppose now that $\Sigma_k = \sum_i \sigma_i^k \boldsymbol{v}_i \boldsymbol{v}_i^\top$ for all k, i.e., the covariance matrices share the same basis of eigenvectors $\{\boldsymbol{v}_i\}_i$. Then, denoting $\mu_i^k \equiv \sqrt{d}\boldsymbol{\mu}_k^\top \boldsymbol{v}_i$

$$\begin{aligned}
\boldsymbol{Q}_k &= \frac{1}{d}\sum_{i=1}^{d} \sigma_i^k \left(\lambda \boldsymbol{I}_K + \sum_\kappa \sigma_i^\kappa \hat{\boldsymbol{V}}_\kappa\right)^{-1} \left(\sum_{kk'} \mu_i^k \mu_i^{k'} \hat{\boldsymbol{m}}_k \hat{\boldsymbol{m}}_{k'}^\top + \sum_\kappa \sigma_i^\kappa \hat{\boldsymbol{Q}}_\kappa\right) \left(\lambda \boldsymbol{I}_K + \sum_\kappa \sigma_i^\kappa \hat{\boldsymbol{V}}_\kappa\right)^{-1} \\
\boldsymbol{m}_k &= \frac{1}{d}\sum_{i=1}^{d} \sum_{k'} \mu_i^k \mu_i^{k'} \left(\lambda \boldsymbol{I}_K + \sum_\kappa \sigma_i^\kappa \hat{\boldsymbol{V}}_\kappa\right)^{-1} \hat{\boldsymbol{m}}_{k'} \\
\boldsymbol{V}_k &= \frac{1}{d}\sum_{i=1}^{d} \sigma_i^k \left(\lambda \boldsymbol{I}_K + \sum_\kappa \sigma_i^\kappa \hat{\boldsymbol{V}}_\kappa\right)^{-1}.
\end{aligned} \tag{167}$$

Introducing the joint density

$$\frac{1}{d}\sum_{i=1}^{d} \prod_{\kappa=1}^{K} \delta(\sigma^\kappa - \sigma_i^\kappa)\delta(\mu^\kappa - \mu_i^\kappa) \xrightarrow{d\to+\infty} \rho(\boldsymbol{\sigma}, \boldsymbol{\mu}), \tag{168}$$

then we can write the saddle-point equations given in Corollary 3

$$\begin{aligned}
\boldsymbol{Q}_k &= \mathbb{E}_{\boldsymbol{\sigma},\boldsymbol{\mu}} \left[\sigma^k \left(\lambda \boldsymbol{I}_K + \sum_\kappa \sigma^\kappa \hat{\boldsymbol{V}}_\kappa\right)^{-1} \left(\sum_{kk'} \mu^k \mu^{k'} \hat{\boldsymbol{m}}_k \hat{\boldsymbol{m}}_{k'}^\top + \sum_\kappa \sigma^\kappa \hat{\boldsymbol{Q}}_\kappa\right) \left(\lambda \boldsymbol{I}_K + \sum_\kappa \sigma^\kappa \hat{\boldsymbol{V}}_\kappa\right)^{-1}\right] \\
\boldsymbol{m}_k &= \mathbb{E}_{\boldsymbol{\sigma},\boldsymbol{\mu}} \left[\mu^k \left(\lambda \boldsymbol{I}_K + \sum_\kappa \sigma^\kappa \hat{\boldsymbol{V}}_\kappa\right)^{-1} \sum_\kappa \mu^\kappa \hat{\boldsymbol{m}}_\kappa\right] \\
\boldsymbol{V}_k &= \mathbb{E}_{\boldsymbol{\sigma},\boldsymbol{\mu}} \left[\sigma^k \left(\lambda \boldsymbol{I}_K + \sum_\kappa \sigma^\kappa \hat{\boldsymbol{V}}_\kappa\right)^{-1}\right].
\end{aligned} \tag{169}$$

where the expectations $\mathbb{E}_{\boldsymbol{\sigma},\boldsymbol{\mu}}$ are taken with respect to the joint distribution ρ.

C.1.1 Uniform covariances

Let us consider the simpler case $\Sigma_k \equiv \Delta \boldsymbol{I}_d$, with $\Delta > 0$. In this case, the saddle-point equations can take a more compact form that is particularly suitable for a numerical solution. Moreover, for reasons of symmetry we can write

$$\boldsymbol{Q}_k \equiv \boldsymbol{Q}, \quad \boldsymbol{V}_k \equiv \boldsymbol{V}, \quad \hat{\boldsymbol{Q}}_k \equiv \frac{1}{K\Delta}\hat{\boldsymbol{Q}}_k, \quad \hat{\boldsymbol{V}}_k \equiv \frac{1}{K\Delta}\hat{\boldsymbol{V}}, \quad \forall k. \tag{170}$$

Let us define the following $K \times K$ matrices

- $\boldsymbol{M} \in \mathbb{R}^{K\times K}$ (resp. $\hat{\boldsymbol{M}} \in \mathbb{R}^{K\times K}$) is the matrix obtained concatenenating the vectors $\boldsymbol{m}_k$ (resp. $\hat{\boldsymbol{m}}_k$);
- $\boldsymbol{\Theta} = \left(\boldsymbol{\mu}_k^\top \boldsymbol{\mu}_{k'}\right)_{kk'}$ is the Gram matrix of the means;
- $\boldsymbol{F} \in \mathbb{R}^{K\times K}$ is the matrix obtained concatenenating the vectors $\boldsymbol{f}_k$;
- $\boldsymbol{H} \in \mathbb{R}^{K\times K}$ is the matrix obtained concatenenating the vectors $\boldsymbol{h}_k$;
- $\boldsymbol{\Pi} = \mathrm{diag}(\rho_k) \in \mathbb{R}^{K\times K}$ is a diagonal matrix with elements $\Pi_{kk'} = \delta_{kk'}\rho_k$.

The saddle-point equations then can be rewritten as

$$\begin{aligned}
&\boldsymbol{Q}=\Delta\Big(\lambda\boldsymbol{I}_K+\hat{\boldsymbol{V}}\Big)^{-1}\Big(\hat{\boldsymbol{Q}}+\hat{\boldsymbol{M}}\boldsymbol{\Theta}\hat{\boldsymbol{M}}^\top\Big)\Big(\lambda\boldsymbol{I}_K+\hat{\boldsymbol{V}}\Big)^{-1} \qquad && \hat{\boldsymbol{Q}}=\alpha\Delta\mathbb{E}_\Xi\big[\boldsymbol{F}\boldsymbol{\Pi}\boldsymbol{F}^\top\big]\\
&\boldsymbol{M}=\Big(\lambda\boldsymbol{I}_K+\hat{\boldsymbol{V}}\Big)^{-1}\hat{\boldsymbol{M}}\boldsymbol{\Theta} && \hat{\boldsymbol{V}}=-\alpha\Delta\boldsymbol{Q}^{-1/2}\mathbb{E}_\Xi\big[\boldsymbol{F}\boldsymbol{\Pi}\boldsymbol{\Xi}^\top\big]\\
&\boldsymbol{V}=\Delta\Big(\lambda\boldsymbol{I}_K+\hat{\boldsymbol{V}}\Big)^{-1}, && \hat{\boldsymbol{M}}=\alpha\mathbb{E}_\Xi[\boldsymbol{F}\boldsymbol{\Pi}]\\
& && \boldsymbol{b}=\mathbb{E}_\Xi[(\boldsymbol{H}-\boldsymbol{M})\boldsymbol{\Pi}\boldsymbol{1}_K].
\end{aligned} \tag{171}$$

Here and in the following $\boldsymbol{1}_K$ is the vector of K components all equal to 1. These expressions are particularly suitable for a numerical implementation, because involve matrix multiplications and inversions of K-dimensional objects only.

Quadratic loss — If we consider a quadratic loss $\ell(\boldsymbol{y},\boldsymbol{x}) = \frac{1}{2}(\boldsymbol{y}-\boldsymbol{x})^2$, then an explicit formula for the proximal can be found, namely

$$\boldsymbol{f}_k = (\boldsymbol{I}_K+\boldsymbol{V})^{-1}(\boldsymbol{e}_K-\boldsymbol{\omega}_k) \tag{172}$$

so that the second set of saddle-point equations (171) can be written as

$$\begin{aligned}
\hat{\boldsymbol{Q}} &= \alpha(\boldsymbol{I}_K+\boldsymbol{V})^{-1}\left[(\boldsymbol{I}_K-\boldsymbol{M}-\boldsymbol{b}\otimes\boldsymbol{1}_K)\boldsymbol{\Pi}(\boldsymbol{I}_K-\boldsymbol{M}-\boldsymbol{b}\otimes\boldsymbol{1}_K)^\top+\boldsymbol{Q}\right](\boldsymbol{I}_K+\boldsymbol{V})^{-1}\\
\hat{\boldsymbol{M}} &= \alpha(\boldsymbol{I}_K+\boldsymbol{V})^{-1}(\boldsymbol{I}_K-\boldsymbol{M}-\boldsymbol{b}\otimes\boldsymbol{1}_K)\boldsymbol{\Pi}\\
\hat{\boldsymbol{V}} &= \alpha\Delta(\boldsymbol{I}_K+\boldsymbol{V})^{-1}.
\end{aligned} \tag{173}$$

Observe at this point that we can explicitly solve for $\boldsymbol{V}$ using the equation for it in Eqs (171). In particular, $\boldsymbol{V}$ satisfies the equation $\lambda\boldsymbol{V}^2+(\alpha+\lambda-\Delta)\boldsymbol{V}=\Delta\boldsymbol{I}_K$. Being $\boldsymbol{V}$ positive definite, it follows that it is diagonal, $\boldsymbol{V}=V\boldsymbol{I}_K$ with diagonal element

$$V=\frac{\Delta(1-\alpha)-\lambda+\sqrt{(\Delta-\alpha\Delta-\lambda)^2+4\Delta\lambda}}{2\lambda}, \quad \hat{V}=\frac{\alpha\Delta}{1+V}, \tag{174}$$

so that

$$\begin{aligned}
&\boldsymbol{Q}=\frac{\Delta}{(\lambda+\Delta\hat{V})^2}\Big(\hat{\boldsymbol{Q}}+\hat{\boldsymbol{M}}\boldsymbol{\Theta}\hat{\boldsymbol{M}}^\top\Big) && \hat{\boldsymbol{Q}}=\frac{\alpha[(\boldsymbol{I}_K-\boldsymbol{M}-\boldsymbol{b}\otimes\boldsymbol{1}_K)\boldsymbol{\Pi}(\boldsymbol{I}_K-\boldsymbol{M}-\boldsymbol{b}\otimes\boldsymbol{1}_K)^\top+\boldsymbol{Q}]}{(1+V)^2}\\
&\boldsymbol{M}=\frac{\hat{\boldsymbol{M}}\boldsymbol{\Theta}}{\lambda+\Delta\hat{V}}, && \hat{\boldsymbol{M}}=-\frac{\alpha(\boldsymbol{I}_K-\boldsymbol{M}-\boldsymbol{b}\otimes\boldsymbol{1}_K)\boldsymbol{\Pi}}{1+V}.\\
&\boldsymbol{b}=(\boldsymbol{I}_K-\boldsymbol{M})\boldsymbol{\Pi}\boldsymbol{1}_K, &&
\end{aligned} \tag{175}$$

In the $\lambda\to 0$ limit, for $\alpha<1$ it is convenient to rescale $\hat{\boldsymbol{Q}}\mapsto\lambda^2\hat{\boldsymbol{Q}}$ and $\hat{\boldsymbol{M}}\mapsto\lambda\hat{\boldsymbol{M}}$, so that

$$\begin{aligned}
&\boldsymbol{Q}=\Delta(1-\alpha)^2\Big(\hat{\boldsymbol{Q}}+\hat{\boldsymbol{M}}\boldsymbol{\Theta}\hat{\boldsymbol{M}}^\top\Big), && \hat{\boldsymbol{Q}}=\frac{\alpha[(\boldsymbol{I}_K-\boldsymbol{M}-\boldsymbol{b}\otimes\boldsymbol{1}_K)\boldsymbol{\Pi}(\boldsymbol{I}_K-\boldsymbol{M}-\boldsymbol{b}\otimes\boldsymbol{1}_K)^\top+\boldsymbol{Q}]}{\Delta^2(1-\alpha)^2},\\
&\boldsymbol{M}=(1-\alpha)\hat{\boldsymbol{M}}\boldsymbol{\Theta}, && \hat{\boldsymbol{M}}=-\frac{\alpha(\boldsymbol{I}_K-\boldsymbol{M}-\boldsymbol{b}\otimes\boldsymbol{1}_K)\boldsymbol{\Pi}}{\Delta(1-\alpha)}.\\
&\boldsymbol{b}=(\boldsymbol{I}_K-\boldsymbol{M})\boldsymbol{\Pi}\boldsymbol{1}_K, &&
\end{aligned} \tag{176}$$

Cross-entropy loss — We consider now the relevant case of the cross entropy loss

$$\ell(\boldsymbol{y}, \boldsymbol{x}) = -\sum_{k=1}^{K} y_k \ln \frac{e^{x_k}}{\sum_{\kappa=1}^{K} e^{x_\kappa}}. \tag{177}$$

If $\boldsymbol{y} \in \{\boldsymbol{e}_k\}_{k\in[K]}$, the loss can be written in the form $\ell(\boldsymbol{y}, \boldsymbol{x}) = -\boldsymbol{y}^\top \boldsymbol{x} + \ln \sum_\kappa e^{x_\kappa}$. If we introduce the *softmax function* $\mathbf{soft}\colon \mathbb{R}^K \to \mathbb{R}^K$

$$\partial_{\boldsymbol{x}} \ell(\boldsymbol{y}, \boldsymbol{x}) = -\boldsymbol{y} + \mathbf{soft}(\boldsymbol{x}), \qquad \mathrm{soft}_k(\boldsymbol{x}) \equiv \frac{\exp(x_k)}{\sum_\kappa \exp(x_\kappa)} \tag{178}$$

the proximal equation for the cross-entropy loss is the solution of the equations:

$$\boldsymbol{V}^{-1}(\boldsymbol{h}_k - \boldsymbol{\omega}_k) - \boldsymbol{e}_k + \mathbf{soft}(\boldsymbol{h}_k) = \boldsymbol{0} \iff \boldsymbol{f}_k = \boldsymbol{e}_k - \mathbf{soft}(\boldsymbol{V}\boldsymbol{f}_k + \boldsymbol{\omega}_k) \quad \forall k \in [K], \tag{179}$$

having only one solution for which, however, there is no closed-form expression. The equation can be solved numerically, and in this way we obtained the results in Section 3.2.

The saddle-point equations can be written rescaling $\boldsymbol{Q} \mapsto \lambda^{-2}\boldsymbol{Q}$, $\boldsymbol{V} \mapsto \lambda^{-1}\boldsymbol{V}$, $\boldsymbol{M} \mapsto \lambda^{-1}\boldsymbol{M}$, $\boldsymbol{b} \mapsto \lambda^{-1}\boldsymbol{b}$, $\hat{\boldsymbol{V}} \mapsto \lambda\hat{\boldsymbol{V}}$. They become

$$\begin{aligned}
\boldsymbol{Q} &= \Delta \left(\boldsymbol{I}_K + \hat{\boldsymbol{V}}\right)^{-1} \left(\hat{\boldsymbol{Q}} + \hat{\boldsymbol{M}}\boldsymbol{\Theta}\hat{\boldsymbol{M}}^\top\right) \left(\boldsymbol{I}_K + \hat{\boldsymbol{V}}\right)^{-1}, & \hat{\boldsymbol{Q}} &= \alpha\Delta\mathbb{E}_\Xi\left[\boldsymbol{F}\boldsymbol{\Pi}\boldsymbol{F}^\top\right], \\
\boldsymbol{M} &= \left(\boldsymbol{I}_K + \hat{\boldsymbol{V}}\right)^{-1} \hat{\boldsymbol{M}}\boldsymbol{\Theta} & \hat{\boldsymbol{V}} &= -\alpha\Delta\boldsymbol{Q}^{-1/2}\mathbb{E}_\Xi\left[\boldsymbol{F}\boldsymbol{\Pi}\Xi^\top\right], \\
\boldsymbol{V} &= \Delta\left(\boldsymbol{I}_K + \hat{\boldsymbol{V}}\right)^{-1}, & \hat{\boldsymbol{M}} &= \alpha\mathbb{E}_\Xi\left[\boldsymbol{F}\boldsymbol{\Pi}\right], \\
& & \boldsymbol{b} &= \mathbb{E}_\Xi\left[(\boldsymbol{H} - \boldsymbol{M})\boldsymbol{\Pi}\right],
\end{aligned} \tag{180}$$

so that the dependence on λ disappears everywhere except in the equation for the proximal $\boldsymbol{f}_k$

$$\boldsymbol{f}_k = \arg\min_{\boldsymbol{x}} \left[\frac{1}{2}\boldsymbol{x}^\top\boldsymbol{V}\boldsymbol{x} + \lambda\ell\left(\boldsymbol{e}_k, \frac{\boldsymbol{V}\boldsymbol{x} + \boldsymbol{\omega}_k}{\lambda}\right)\right], \tag{181}$$

which, in the $\lambda \to 0$ limit, becomes

$$\boldsymbol{f}_k = \arg\min_{\boldsymbol{x}} \left[\frac{1}{2}\boldsymbol{x}^\top\boldsymbol{V}\boldsymbol{x} + \min_\mu\{(\boldsymbol{e}_\mu - \boldsymbol{e}_k)^\top(\boldsymbol{V}\boldsymbol{x} + \boldsymbol{\omega}_k)\}\right]. \tag{182}$$

Note that in this limit, minimising the cross-entropy loss yields precisely the max-margin estimator [68].

C.2 The $K = 2$ case with scalar labels

The formulas for the $K = 2$ case can be derived directly from the general analysis given above imposing $L = 1$. In particular, let us assume that the two clusters are labeled with $e_1 = +1$ and $e_2 = -1$. Using as classifier

$$\varphi(x) = \mathrm{sign}(x) \tag{183}$$

the expression of the average errors is

$$\begin{aligned}
\epsilon_g &= \sum_{k\in[2]} \rho_k \mathbb{E}_\xi \left[\theta\left((-1)^k \omega_k^\star\right)\right] = \sum_{k\in[2]} \frac{\rho_k}{2} \mathrm{erfc}\left((-1)^{k-1} \frac{m_k^\star + b^\star}{\sqrt{2q_k^\star}}\right), \\
\epsilon_t &= \sum_{k\in[2]} \rho_k \mathbb{E}_\xi \left[\theta\left((-1)^k h_k^\star\right)\right], \\
\epsilon_\ell &= \sum_{k\in[2]} \rho_k \mathbb{E}[\ell((-1)^{k-1}, h_k^\star)].
\end{aligned} \tag{184}$$

We will further explore this case, considering some special cases in the following.

C.2.1 Example: ℓ_1 regularization

In this Section we derive the saddle-point equations for the the case in which the two cluster have opposite means $\boldsymbol{\mu}_1 = -\boldsymbol{\mu}_2 \equiv \boldsymbol{\mu}$, and the same diagonal covariance matrix, $\Sigma_1 = \Sigma_2 \equiv \Sigma$, with $\Sigma_{ij} = \sigma_i \delta_{ij}$ and $\sigma_i > 0$. In this case, for symmetry reasons, the overlaps simplify and we have:

$$V_1 = V_2 \equiv V, \qquad q_1 = q_2 \equiv q, \qquad m_+ = -m_- \equiv m, \tag{185}$$

$$\hat{V}_+ = \hat{V}_- \equiv \frac{1}{2}\hat{V}, \qquad \hat{q}_+ = \hat{q}_- \equiv \frac{1}{2}\hat{q}, \qquad \hat{m}_+ = -\hat{m}_- \equiv \frac{1}{2}\hat{m}. \tag{186}$$

We define

$$\frac{1}{d}\sum_{i=1}^{d} \delta(\sigma - \sigma_i)\delta(\mu - \sqrt{d}\mu_i) \xrightarrow{d\to+\infty} p(\sigma, \mu) \tag{187}$$

joint distribution of the covariance diagonal elements and of the mean elements. We will denote $\mathbb{E}_{\mu,\sigma}[\bullet]$ the average with respect to this measure. We will focus in particular on the form of the saddle-point equations obtained from the prior contribution assuming ℓ_1 regularization, i.e., $r(\boldsymbol{w}) = \sum_i |w_i|$, and let us introduce the corresponding *soft-thresholding operator*:

$$\mathrm{Prox}_{\lambda|\cdot|}(x) = \mathrm{sign}(x)\max\{|x| - \lambda, 0\}. \tag{188}$$

Observe that $\mathrm{Prox}_{\alpha\lambda|\cdot|}(\alpha x) = \alpha \mathrm{Prox}_{\lambda|\cdot|}(x)$ for $\alpha > 0$. Its derivative given by $\mathrm{Prox}'_{\lambda|\cdot|}(x) = \theta(|x| > \lambda)$. The saddle point equations from the prior part simply read:

$$V = \frac{1}{\hat{V}}\, \mathbb{E}_{\mu,\sigma,\xi}\left[\mathrm{Prox}'_{\frac{\lambda}{\sigma\hat{V}}|\cdot|}\left(\frac{\hat{m}\mu + \sqrt{\hat{q}}\sigma\xi}{\hat{V}\sigma}\right)\right], \tag{189}$$

$$q = \mathbb{E}_{\mu,\sigma,\xi}\left[\sigma\, \left(\mathrm{Prox}_{\frac{\lambda}{\sigma\hat{V}}|\cdot|}\left(\frac{\hat{m}\mu + \sqrt{\hat{q}}\sigma\xi}{\hat{V}\sigma}\right)\right)^2\right], \tag{190}$$

$$m = \mathbb{E}_{\mu,\sigma,\xi}\left[\mu \mathrm{Prox}_{\frac{\lambda}{\sigma\hat{V}}|\cdot|}\left(\frac{\hat{m}\mu + \sqrt{\hat{q}}\sigma\xi}{\hat{V}\sigma}\right)\right]. \tag{191}$$

The averages over ξ can be performed explicitely using the simple expression of the proximal in this case. If we define the auxiliary functions

$$\begin{aligned}
\phi^0_\pm(v,u,\lambda) &\equiv \frac{1}{2}\mathrm{erfc}\left(\frac{\lambda \pm v}{\sqrt{2u}}\right)\\
\phi^1_\pm(u,v,\lambda) &= \sqrt{\frac{u}{2\pi}}e^{-\frac{(v\pm\lambda)^2}{2u}} - \frac{v\pm\lambda}{2}\mathrm{erfc}\left(\frac{\lambda \pm v}{\sqrt{2u}}\right),\\
\phi^2_\pm(v,u,\lambda) &= -\sqrt{\frac{u}{2\pi}}e^{-\frac{(\lambda\pm v)^2}{2u}}(\lambda\pm v) + \frac{u+(\lambda\pm v)^2}{2}\mathrm{erfc}\left(\frac{\lambda \pm v}{\sqrt{2u}}\right).
\end{aligned} \tag{192}$$

then

$$\begin{aligned}
V &= \frac{1}{\hat{V}}\mathbb{E}_{\mu,\sigma}\left[\phi^0_+(\mu\hat{m},\sigma\hat{q},\lambda) + \phi^0_-(\mu\hat{m},\sigma\hat{q},\lambda)\right]\\
q &= \mathbb{E}_{\mu,\sigma}\left[\frac{\phi^2_+(\mu\hat{m},\sigma\hat{q},\lambda) + \phi^2_-(\mu\hat{m},\sigma\hat{q},\lambda)}{\sigma\hat{V}^2}\right],\\
m &= \mathbb{E}_{\mu,\sigma}\left[\frac{\mu\phi^1_-(\mu\hat{m},\sigma q,\lambda) - \mu\phi^1_+(\mu\hat{m},\sigma q,\lambda)}{\sigma\hat{V}}\right].
\end{aligned} \tag{193}$$

Gaussian means, homogenous covariances — If $p(\mu,\sigma) = \mathcal{N}(\mu|0,1)\delta(\sigma-\Delta)$, i.e., the means have i.i.d. Gaussian entries and $\Sigma = \Delta \boldsymbol{I}_d$, then

$$\begin{aligned}
V &= \frac{1}{\hat{V}}\mathbb{E}_z\left[\mathrm{erfc}\left(\frac{\lambda+\hat{m}z}{\sqrt{2\Delta\hat{q}}}\right)\right],\\
q &= \frac{1}{\Delta\hat{V}^2}\left\{-\frac{e^{-\frac{1}{2}\frac{\lambda^2}{\hat{m}^2+\Delta\hat{q}}}}{\sqrt{2\pi(\hat{m}^2+\Delta\hat{q})}}\frac{2(\Delta\hat{q})^2\lambda}{\hat{m}^2+\Delta\hat{q}} + \mathbb{E}_z\left[(\lambda+\hat{m}z)^2\,\mathrm{erfc}\left(\frac{\lambda+\hat{m}z}{\sqrt{2\Delta\hat{q}}}\right)\right]\right\},\\
m &= \frac{1}{\Delta\hat{V}}\left\{\frac{e^{-\frac{1}{2}\frac{\lambda^2}{\hat{m}^2+\Delta\hat{q}}}}{\sqrt{2\pi(\hat{m}^2+\Delta\hat{q})}}\frac{2\Delta\hat{q}\hat{m}\lambda}{\hat{m}^2+\Delta\hat{q}} + \mathbb{E}_{z\sim\mathcal{N}(0,1)}\left[(\lambda+\hat{m}z)\,z\,\mathrm{erfc}\left(\frac{\lambda+\hat{m}z}{\sqrt{2\Delta\hat{q}}}\right)\right]\right\},
\end{aligned} \tag{194}$$

with $z\sim\mathcal{N}(0,1)$.

Covariance correlated with sparse means — In Section 3.1 we considered the case of sparse means correlated with the covariance matrices. In particular, we considered

$$p(\sigma,\mu) = p\mathcal{N}(\mu|0,1)\delta(\sigma-\Delta_1) + (1-p)\delta(\mu)\delta(\sigma-\Delta_0). \tag{195}$$

The saddle-point equations are therefore

$$V = \frac{1}{\hat{V}}\left[p\mathbb{E}_\mu\left[\mathrm{erfc}\left(\frac{\lambda+\hat{m}\mu}{\sqrt{2\Delta_1\hat{q}}}\right)\right] + (1-p)\mathrm{erfc}\left(\frac{\lambda}{\sqrt{2\Delta_0\hat{q}}}\right)\right] \tag{196}$$

$$q = \frac{p}{\Delta_1 \hat{V}^2}\left\{-\frac{e^{-\frac{1}{2}\frac{\lambda^2}{\hat{m}^2+\Delta_1\hat{q}}}}{\sqrt{2\pi(\hat{m}^2+\Delta_1\hat{q})}}\frac{2(\Delta_1\hat{q})^2\lambda}{\hat{m}^2+\Delta_1\hat{q}} + \mathbb{E}_z\left[(\lambda+\hat{m}z)^2\,\mathrm{erfc}\left(\frac{\lambda+\hat{m}z}{\sqrt{2\Delta_1\hat{q}}}\right)\right]\right\}$$
$$-\lambda(1-p)\sqrt{\frac{\Delta_0\hat{q}}{2\pi}}e^{-\frac{\lambda^2}{2\Delta_0\hat{q}}} + \frac{1-p}{2}(\Delta_0\hat{q}+\lambda^2)\mathrm{erfc}\left(\frac{\lambda}{\sqrt{2\Delta_0\hat{q}}}\right) \tag{197}$$

$$m = \frac{p}{\Delta_1\hat{V}}\left\{\frac{e^{-\frac{1}{2}\frac{\lambda^2}{\hat{m}^2+\Delta_1\hat{q}}}}{\sqrt{2\pi(\hat{m}^2+\Delta_1\hat{q})}}\frac{2\Delta_1\hat{q}\hat{m}\lambda}{\hat{m}^2+\Delta_1\hat{q}} + \mathbb{E}_z\left[(\lambda+\hat{m}z)\,z\,\mathrm{erfc}\left(\frac{\lambda+\hat{m}z}{\sqrt{2\Delta_1\hat{q}}}\right)\right]\right\}. \tag{198}$$

In Section 3.1 we compare the performance obtained adopting an ℓ_1 regularization with the corresponding one obtained using ℓ_2, $r(\boldsymbol{w}) = \sum_i w_i^2$. For the sake of completeness, we give here the expression of the saddle-point equations in that case as well. In this case, the prior term Ψ_w can be written explicitly after a Gaussian integration as

$$\Psi_w(\hat{m},\hat{Q},\hat{V}) = -\frac{1}{2d}\,\mathrm{tr}\ln\left(\lambda\boldsymbol{I}_d + \hat{V}\Sigma\right) + \frac{1}{2}\,\mathrm{tr}\left[\left(\lambda\boldsymbol{I}_d+\hat{V}\Sigma\right)^{-1}\left(\hat{m}_k^2\boldsymbol{\mu}\boldsymbol{\mu}^\top + \frac{\hat{q}}{d}\Sigma\right)\right]. \tag{199}$$

In the setting given by Eq (195) the saddle point equations are then

$$q = p\frac{\hat{m}^2\Delta_1+\hat{q}\Delta_1^2}{(\lambda+\hat{V}\Delta_1)^2} + \frac{(1-p)\hat{q}\Delta_0^2}{(\lambda+\hat{V}\Delta_0)^2} \tag{200a}$$

$$V = p\frac{\Delta_1}{\lambda+\hat{V}\Delta_1} + \frac{(1-p)\Delta_0}{\lambda+\hat{V}\Delta_0} \tag{200b}$$

$$m = \frac{\hat{m}p}{\lambda+\hat{V}\Delta_1}. \tag{200c}$$

D Bayes optimal error

In this Appendix, we derive a formula for the Bayes optimal classification error in the case of K clusters with the same covariance $\Sigma_k = \Delta\boldsymbol{I}_d$ in the large d limit, assuming that a dataset $\{(\boldsymbol{x}^\nu,\boldsymbol{y}^\nu)\}_{\nu\in[n]}$ of correctly labeled points is available. As usual, we will assume $n/d = \alpha$ finite. The distribution of a pair $(\boldsymbol{y},\boldsymbol{x})$ is given by

$$p(\boldsymbol{y},\boldsymbol{x}|\boldsymbol{M}) = \sum_k y_k\frac{\rho_k\exp\left(-\frac{1}{2\Delta}\left\|\boldsymbol{x}-\boldsymbol{\mu}_k\right\|^2\right)}{(2\pi\Delta)^{\frac{d}{2}}}. \tag{201}$$

where $\boldsymbol{M}\in\mathbb{R}^{d\times K}$ is the matrix of concatenated means $\boldsymbol{\mu}_k$ *estimated* from the dataset, so that

$$\begin{aligned} p(\boldsymbol{M}|\{\boldsymbol{y}^\nu,\boldsymbol{x}^\nu\}_\nu) &\propto p(\{\boldsymbol{x}^\nu\}_\nu|\boldsymbol{M},\{\boldsymbol{y}^\nu\}_\nu)P_{\boldsymbol{\mu}}(\boldsymbol{M}) \\ &\propto P_{\boldsymbol{\mu}}(\boldsymbol{M})\prod_{\nu=1}^n\sum_k y_k^\nu\exp\left(-\frac{1}{2\Delta}\left\|\boldsymbol{x}^\nu-\boldsymbol{\mu}_k\right\|^2\right). \end{aligned} \tag{202}$$

We will assume in the following the distribution

$$P_{\boldsymbol{\mu}}(\boldsymbol{M}) = \frac{\exp\left(-\frac{d}{2}\mathrm{tr}[\boldsymbol{M}\boldsymbol{\Theta}^{-1}\boldsymbol{M}^\top]\right)}{(2\pi)^{\frac{Kd}{2}}d^{-K/2}|\boldsymbol{\Theta}|^{1/2}} \tag{203}$$

where $\boldsymbol{\Theta} \in \mathbb{R}^{K\times K}$ is a given positive definite covariance matrix. In this way

$$\mathbb{E}\left[\boldsymbol{M}^\top \boldsymbol{M}\right] = \boldsymbol{\Theta}. \tag{204}$$

The conditional distribution for the label $\boldsymbol{y}^0$ of a new point $\boldsymbol{x}^0$,

$$\begin{aligned} p(\boldsymbol{y}^0|\boldsymbol{x}^0, \{\boldsymbol{y}^\nu, \boldsymbol{x}^\nu\}_\nu) &\propto \mathbb{E}_{\boldsymbol{M}|\{\boldsymbol{y}^\nu,\boldsymbol{x}^\nu\}_\nu}[p(\boldsymbol{y}, \boldsymbol{x}|\boldsymbol{M})] \\ &= \int \mathrm{d}\boldsymbol{M} P_{\boldsymbol{\mu}}(\boldsymbol{M}) \sum_k y_k^0 \rho_k \exp\left(-\frac{\left\|\boldsymbol{x}^0 - \boldsymbol{\mu}_k\right\|^2}{2\Delta}\right) \prod_{\nu=1}^n \sum_k y_k^\nu \exp\left(-\frac{\left\|\boldsymbol{x}^\nu - \boldsymbol{\mu}_k\right\|^2}{2\Delta}\right). \end{aligned} \tag{205}$$

If $\boldsymbol{n} = (n_k)_k$ is the vector of the number of examples n_k in the class k, then

$$\begin{aligned} p(\boldsymbol{y}^0|\boldsymbol{x}^0, \{\boldsymbol{y}^\nu, \boldsymbol{x}^\nu\}_\nu) &\propto \int \mathrm{d}\boldsymbol{M} P_{\boldsymbol{\mu}}(\boldsymbol{M}) \prod_{k=1}^K \left[\rho_k^{y_k^0} \exp\left(-\sum_{\nu=0}^n \frac{y_k^\nu \left\|\boldsymbol{x}^\nu - \boldsymbol{\mu}_k\right\|^2}{2\Delta}\right)\right] \\ &= \exp\left[\sum_k y_k^0 \left(\ln \rho_k - \frac{\|\boldsymbol{x}\|^2}{2\Delta}\right) - \frac{1}{2}\ln\det\left(1 + \frac{1}{d\Delta}\mathrm{diag}(\boldsymbol{n} + \boldsymbol{y}^0)\boldsymbol{\Theta}\right)\right] \\ &\quad \times \exp\left[\frac{1}{2\Delta}\mathrm{tr}\left[\left(\sum_{\nu=0}^n \boldsymbol{y}^\nu \otimes \boldsymbol{x}^\nu\right)^\top \left(d\Delta\boldsymbol{\Theta}^{-1} + \mathrm{diag}(\boldsymbol{n} + \boldsymbol{y})\right)^{-1} \left(\sum_{\nu=0}^n \boldsymbol{y}^\nu \otimes \boldsymbol{x}^\nu\right)\right]\right]. \end{aligned} \tag{206}$$

In the following we will denote by $\star$ the true label of $\boldsymbol{x}$. Let $\boldsymbol{\Pi} = \mathrm{diag}(\rho_k)$. Then we can write the previous expression as

$$\begin{aligned} p(\boldsymbol{y}^0|\boldsymbol{x}^0, \{\boldsymbol{y}^\nu, \boldsymbol{x}^\nu\}_\nu) &\propto \exp\left[\sum_k y_k \left(\ln \rho_k - \frac{\|\boldsymbol{x}^0\|^2}{2\Delta}\right) - \frac{1}{2}\ln\det\left(1 + \frac{1}{\Delta}\alpha\boldsymbol{\Pi}\boldsymbol{\Theta}\right)\right] \\ &\quad \times \exp\left[\frac{1}{2\Delta}\mathrm{tr}\left[\left(\frac{1}{d}\sum_{\nu=0}^n \boldsymbol{y}^\nu \otimes \boldsymbol{x}^\nu\right)^\top \left(\Delta\boldsymbol{\Theta}^{-1} + \alpha\boldsymbol{\Pi}\right)^{-1} \left(\sum_{\nu=0}^n \boldsymbol{y}^\nu \otimes \boldsymbol{x}^\nu\right)\right]\right] \end{aligned} \tag{207}$$

Observe now that

$$\frac{1}{d\Delta}\boldsymbol{x}^0 \sum_{\nu=1}^n y_k^\nu \boldsymbol{x}^\nu \xrightarrow{n,d\to+\infty} \alpha\rho_k \frac{\Theta_{\star,k} + \eta_k Z_k}{\Delta}, \qquad \eta_k \equiv \sqrt{\Delta\left(1 + \frac{\Delta}{\alpha\rho_k}\right)}, \quad Z_k \sim \mathcal{N}(0,1), \tag{208}$$

so that, defining the vector $\boldsymbol{a}^\star = (a_k)_{k\in[K]}$ with elements

$$a_k^\star \equiv \alpha\rho_k \frac{\Theta_{\star,k} + \eta_k Z_k}{\Delta}, \tag{209}$$

and neglecting the $\boldsymbol{y}^0$-independent contributions, the expression above can be rewritten as

$$p(\boldsymbol{y}^0|\boldsymbol{x}^0, \{\boldsymbol{y}^\nu, \boldsymbol{x}^\nu\}_\nu) \propto \exp\left[\sum_k y_k^0 \ln \rho_k + \left(\boldsymbol{a}^\star + \frac{1}{2}\boldsymbol{y}^0\right)^\top \left(\Delta\boldsymbol{\Theta}^{-1} + \alpha\boldsymbol{\Pi}\right)^{-1} \boldsymbol{y}^0\right] \tag{210}$$

where we have also used the fact that $\|\boldsymbol{x}^0\|^2 = d\Delta + O(1)$. This means that the Bayes optimal generalization error is

$$\varepsilon_g^{\mathrm{BO}} = \sum_k \rho_k \mathbb{P}\left[\arg\max_\kappa \left(\ln \rho_\kappa + \left(\boldsymbol{a}^k + \frac{1}{2}\boldsymbol{e}_\kappa\right)^\top \left(\Delta\boldsymbol{\Theta}^{-1} + \alpha\boldsymbol{\Pi}\right)^{-1} \boldsymbol{e}_\kappa\right) \neq k\right]. \tag{211}$$

If $\boldsymbol{\Theta} = \boldsymbol{I}_K$ and the clusters have same weights, $\rho_k \equiv 1/K \Leftrightarrow \boldsymbol{\Pi} = 1/K \boldsymbol{I}_K$, then $\eta_k \equiv \eta$ and

$$\varepsilon_g^{\text{BO}} = \mathbb{P}\left[\frac{1}{\eta} < \max_{\kappa \in [K-1]} Z_\kappa + Z\right], \tag{212}$$

that is the formula given in [20].

E Experiments with real data

In this Appendix we discuss the experiments of Section 3.3 with real data sets.

Numerical details — Consider a real data set $\{(\boldsymbol{x}^\nu, y^\nu)\}_{\nu=1}^{n_{\text{tot}}}$ with n_{tot} samples which we assume are independent. As a pre-processing step we center, normalise and flatten the inputs $\boldsymbol{x}^\nu$ into d-dimensional vectors. For both the MNIST [60] and Fashion-MNIST [61] data sets used in the experiments we have normalised the inputs by 255, such that components $x_i^\nu \in [0, 1]$. In what follows we focus on binary classification tasks and encode the labels as $y^\nu \in \{-1, 1\}$. For example, for the MNIST and Fashion-MNIST data sets we have $d = 784$ and $n_{\text{tot}} = 7 \times 10^4$, and we split the inputs into two classes depending on the task of interest, e.g. odd vs. even digits and clothes vs. accessories items, respectively. Define the empirical distribution over the data set:

$$\hat{P}(\boldsymbol{x}, y) = \frac{1}{n_{\text{tot}}} \sum_{\nu=1}^{n_{\text{tot}}} \delta(\boldsymbol{x} - \boldsymbol{x}^\nu)\delta(y - y^\nu) \tag{213}$$

The question we want to answer is: how well can we approximate the learning curves (ϵ_g, ϵ_t) on a given ERM classification task by approximating $\hat{P}$ with a Gaussian mixture distribution? To answer this question, we consider a Gaussian mixture distribution P_2 as defined in Eq. (1) with the same means and covariances as $\hat{P}$:

$$\hat{\boldsymbol{\mu}}_k = \frac{1}{n_{\text{tot}}} \sum_{\nu=1}^{n_{\text{tot}}} \boldsymbol{x}^\nu \, \mathbb{I}\left(\boldsymbol{x}^\nu \in \mathcal{C}_k\right), \qquad \hat{\Sigma}_k = \frac{1}{n_{\text{tot}}} \sum_{\nu=1}^{n_{\text{tot}}} (\boldsymbol{x}^\nu - \boldsymbol{\mu}_k)(\boldsymbol{x}^\nu - \boldsymbol{\mu}_k)^\top \, \mathbb{I}\left(\boldsymbol{x}^\nu \in \mathcal{C}_k\right) \tag{214}$$

for $k \in \{+, -\}$ labelling the two clusters. Similarly, the class probabilities ρ_k are also estimated from the full data set:

$$\hat{\rho}_k = \frac{1}{n_{\text{tot}}} \sum_{\nu=1}^{n_{\text{tot}}} \mathbb{I}\left(\boldsymbol{x}^\nu \in \mathcal{C}_k\right). \tag{215}$$

The parameters $(\hat{\boldsymbol{\mu}}_k, \hat{\Sigma}_k, \hat{\rho}_k)$ completely characterise the approximating Gaussian mixture distribution P_2, and together with Theorem 1 can be used to compute the theoretical learning curves (ϵ_g, ϵ_t) as in Fig. 5 of the main. Note that this discussion can be easily generalised to the case in which a non-linear feature map $\boldsymbol{\varphi} : \mathbb{R}^d \to \mathbb{R}^p$ is applied to the data prior to fitting. The only difference is that the empirical distribution $\hat{P}$ is defined over the features $\{(\boldsymbol{v}^\nu, y^\nu)\}_{\nu=1}^{n_{\text{tot}}}$ where $\boldsymbol{v}^\nu = \boldsymbol{\varphi}(\boldsymbol{x}^\nu)$, and the Gaussian mixture approximation P_2 is defined with respect to the empirical features distribution. Figure 6 of the main manuscript shows an example where a random feature map $\boldsymbol{v} = \operatorname{erf}(\boldsymbol{F}\boldsymbol{x})$ with $\boldsymbol{F} \in \mathbb{R}^{p \times d}$ a random Gaussian projection applied to MNIST and fashion MNIST before the fitting with different ratios $\gamma = p/d$.

The theoretical learning curves are then compared with two sets of finite instance simulations. First, we simulate the learning problem on synthetic data sampled from the approximating Gaussian mixture

distribution P_2, and the learning curves are computed by averaging over 10 instances of the problem. Second, we simulate the learning problem on the real data set. The real data set is split into training and test sets, and for a given sample complexity $\alpha = n/d$ we sub-sample $n = \alpha d$ points from the training set. The averaged learning curves are computed over different instances of the sub-sampling, with replacement.

Discussion — As expected, we find good agreement between theory and simulations with synthetic data drawn from the approximating Gaussian mixture distribution P_2, even for relatively small input dimensions (e.g. $d = 784$ for MNIST). Surprisingly, we have found that in many cases the Gaussian mixture is a good approximation to the real data curves, see Figs. 5 and 6 for examples of logistic regression on input space and with random features. Figure 7 shows an example where the feature map $\boldsymbol{\varphi}$ is given by removing the last layer of the following fully-connected 2-layer neural network pre-trained on the full MNIST odd vs. even data set:

```
Sequential(
  (0): Linear(in_features=784, out_features=784, bias=False)
  (1): ReLU()
  (2): Linear(in_features=784, out_features=1, bias=False)
  (3): Tanh()
)
```

with the training performed by minimising the square loss with the Adam optimiser and random initialisation. However, we have also found cases in which the approximation is not as sharp, see blue curves in Fig. 10. Understanding the factors determining the quality of the approximation in real data sets is an interesting question we expect to address in future work.

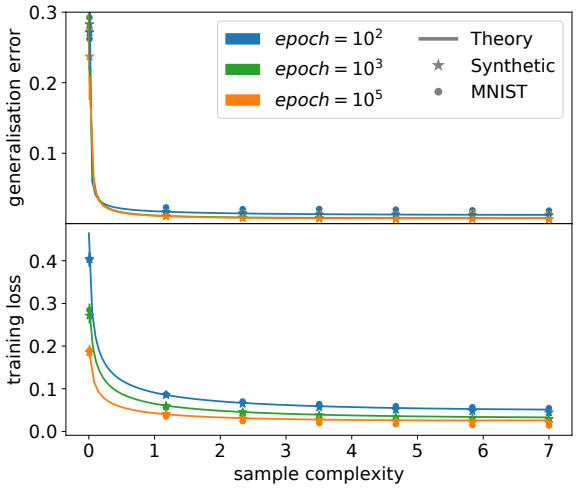

Figure 7: Generalisation error and training loss for logistic regression on MNIST with a feature map $\boldsymbol{\varphi}$ obtained by training 2-layer fully connected neural network, with ℓ_2 penalty and fixed $\lambda = 0.05$. The different curves show the performance at different stages of training.

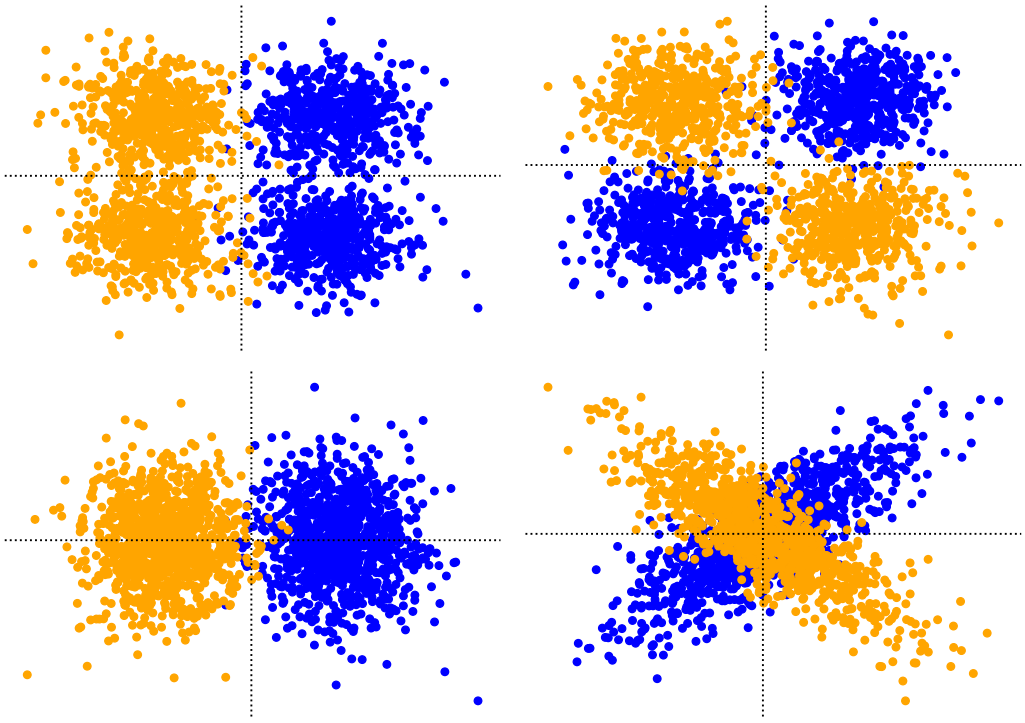

Figure 8: Two dimensional projection of the setting described in eq. (216). (**Left**) Realisable case, (**Right**) Non-realisable case (XOR function).

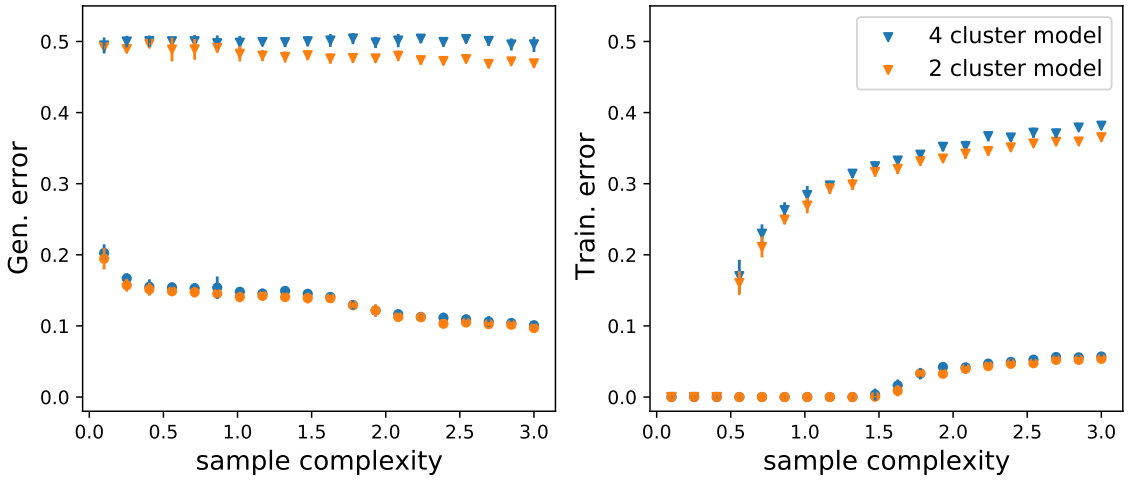

Figure 9: (**Left**) Generalisation and (**right**) training errors as a function of the sample complexity for logistic regression with ℓ_2 penalty and $\lambda = 10^{-4}$ for the four models pictured in Fig. 8. Points denote the separable model (bottom curve), and triangles denote the non-realisable xor model (top curves). We have chosen a balanced scenario with $\Delta = 0.5$.

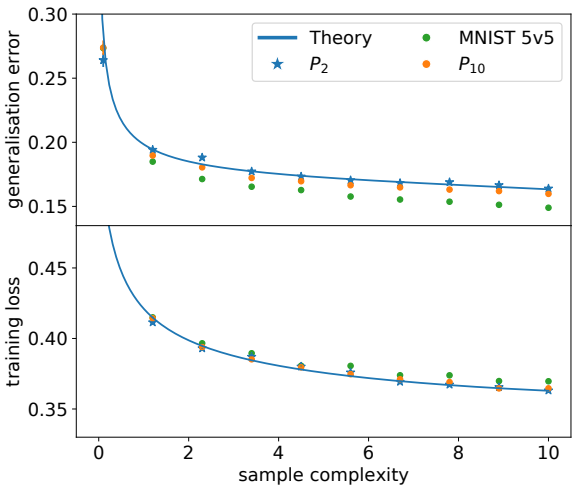

Figure 10: Generalisation error and training loss for logistic regression on the task of classifying $\{0, 1, 2, 3, 4\}$ vs $\{5, 6, 7, 8, 9\}$ digits of MNIST, as a function of the sample complexity for fixed ℓ_2 penalty $\lambda = 0.1$. The blue curves show the 2-Gaussian cluster approximation P_2 (solid for theory, points for finite size simulations), while the orange points show the 10-Gaussian cluster approximation P_{10}, which lies systematically below. The green points denote simulations on the true data set.

Multiclass vs. binary approximation – In the cases previously discussed, we have considered a $K = 2$ cluster approximation P_2 to the empirical data distribution $\hat{P}$. However, the data sets considered here (MNIST and Fashion-MNIST) are originally composed of 10 classes, and therefore we should ask the question of whether a $K = 10$ cluster approximation P_{10} where we fit the means and covariances of each original class is any different from the approximation studied above. In principle, these two approximations can have very different statistical properties. For instance, from Theorem 2 it follows that the generalisation and training errors of Gaussian mixtures only depend on the statistics of the local field $\lambda = \boldsymbol{W}\boldsymbol{x}$ conditioned on the labels, which in the binary setting considered here is $y \in \{+, -\}$. Conditioned on $y = \pm$, this local field is simply a Gaussian random variable under P_2, while it is a multi-modal random variable under P_{10}. Therefore, there is *a priori* no reason for these two approximations to give the same learning curves.

As an example, consider a $K = 4$ Gaussian mixture distribution with a common variance $\Sigma_k = \Delta \mathrm{I}_d$ and with means:

$$\boldsymbol{\mu}_1 = \boldsymbol{e}_1 + \boldsymbol{e}_2, \qquad \boldsymbol{\mu}_2 = \boldsymbol{e}_1 - \boldsymbol{e}_2, \qquad \boldsymbol{\mu}_3 = -\boldsymbol{e}_1 + \boldsymbol{e}_2, \qquad \boldsymbol{\mu}_4 = -\boldsymbol{e}_1 - \boldsymbol{e}_2 \tag{216}$$

where $\boldsymbol{e}_i \in \mathbb{R}^d$ is the canonical basis vector of $\mathbb{R}^d$, with entries $e_{ij} = \delta_{ij}$. We consider two label assignments: a) a realisable case in which clusters 1 and 2 are assigned label +1, and clusters 3 and 4 are assigned -1 and b) a non-realisable case in which clusters 1 and 4 are assigned +1 and clusters 2 and 3 are assigned -1 (XOR function), see Fig. 8 (*top*) for an illustration. Now consider a dual $K = 2$ Gaussian mixture model with means and covariances $(\boldsymbol{\mu}_\pm, \Sigma_\pm)$ chosen to match the class means and covariances of the $K = 4$ mixture, see Fig. 8 (*bottom*) for an illustration. In Fig. 9 we compare the learning curves of the $K = 4$ model with the $K = 2$ counterpart with matched class means and covariances. While in the realisable case *a*) both have identical performance under the error bars, in the non-realisable case *b*) the performance in are significantly different.

Indeed, a similar behaviour can be observed in the real data experiments. Fig. 10 compares the real learning curves of a MNIST 5v5 binary classification task (classifying five first digits vs. five last) with the two different Gaussian mixture approximations: P_{10} where we fit the means and covariances of each individual cluster and P_2, where we fit only the class-wise means and covariances. While both approximations capture the high-level behaviour of the learning curves, P_{10} is closer to the real learning curve than P_2.

Note on numerical instabilities — When dealing with means and covariance matrices estimated from real data sets, we have observed that for small regularisation strength $\lambda \ll 1$ the self-consistent equations from Theorem 1 can develop spurious fixed points corresponding to negative values of the overlap parameters $q_\pm = \boldsymbol{W}^\top \boldsymbol{\Sigma}_\pm \boldsymbol{W}$ – which is clearly not possible since $\boldsymbol{\Sigma}_\pm$ is a positive-definite matrix. This is observed across different scenarios, and is independent of the choice of loss or the particular way the equations are solved. In fact, the minimum value of λ below which the spurious fixed point develop seems to depend only on the conditioning number of the covariance matrices.

References

[1] Stuart Geman, Elie Bienenstock, and René Doursat. Neural networks and the bias/variance dilemma. *Neural Computation*, 4(1):1–58, 1992.

[2] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics. Springer New York Inc., New York, NY, USA, 2001.

[3] Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854, 2019.

[4] Trevor Hastie, Andrea Montanari, Saharon Rosset, and Ryan J. Tibshirani. Surprises in high-dimensional ridgeless least squares interpolation. *Preprint arXiv:1903.08560*, 2020.

[5] Mikhail Belkin, Siyuan Ma, and Soumik Mandal. To understand deep learning we need to understand kernel learning. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 541–549. PMLR, 10–15 Jul 2018.

[6] Peter L. Bartlett, Philip M. Long, Gábor Lugosi, and Alexander Tsigler. Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, 117(48):30063–30070, 2020.

[7] Song Mei and Andrea Montanari. The generalization error of random features regression: Precise asymptotics and double descent curve. *Communications on Pure and Applied Mathematics*, 2019. To appear, preprint arXiv:1908.05355.

[8] Federica Gerace, Bruno Loureiro, Flornet Krzakala, Marc Mézard, and Lenka Zdeborová. Generalisation error in learning with random features and the hidden manifold model. In *37th International Conference on Machine Learning*, 2020.

[9] Behrooz Ghorbani, Song Mei, Theodor Misiakiewicz, and Andrea Montanari. Limitations of lazy training of two-layers neural network. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32, pages 9111–9121, 2019.

[10] Sebastian Goldt, Marc Mézard, Florent Krzakala, and Lenka Zdeborová. Modeling the influence of data structure on learning in neural networks: The hidden manifold model. *Physical Review X*, 10(4):041044, 2020.

[11] Sebastian Goldt, Bruno Loureiro, Galen Reeves, Florent Krzakala, Marc Mézard, and Lenka Zdeborová. The Gaussian equivalence of generative models for learning with shallow neural networks. *Preprint arXiv:2006.14709*, 2020.

[12] Bruno Loureiro, Cédric Gerbelot, Hugo Cui, Sebastian Goldt, Florent Krzakala, Marc Mézard, and Lenka Zdeborová. Capturing the learning curves of generic features maps for realistic data sets with a teacher-student model. *Preprint arXiv:2102.08127*, 2021.

[13] Tengyuan Liang and Pragya Sur. A precise high-dimensional asymptotic theory for Boosting and minimum-ℓ_1-norm interpolated classifiers. *Preprint arXiv:2002.01586*, 2020.

[14] Francesca Mignacco, Florent Krzakala, Yue Lu, Pierfrancesco Urbani, and Lenka Zdeborova. The role of regularization in classification of high-dimensional noisy Gaussian mixture. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 6874–6883. PMLR, 13–18 Jul 2020.

[15] Maria Refinetti, Sebastian Goldt, Florent Krzakala, and Lenka Zdeborová. Classifying high-dimensional gaussian mixtures: Where kernel methods fail and neural networks succeed. *Preprint arXiv:2102.11742*, 2021.

[16] Emmanuel J Candès, Pragya Sur, et al. The phase transition for the existence of the maximum likelihood estimate in high-dimensional logistic regression. *The Annals of Statistics*, 48(1):27–42, 2020.

[17] Vardan Papyan, X. Y. Han, and David L. Donoho. Prevalence of neural collapse during the terminal phase of deep learning training. *Proceedings of the National Academy of Sciences*, 117(40):24652–24663, 2020.

[18] Mohamed El Amine Seddik, Cosme Louart, Mohamed Tamaazousti, and Romain Couillet. Random matrix theory proves that deep learning representations of GAN-data behave as Gaussian mixtures. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 8573–8582. PMLR, 13–18 Jul 2020.

[19] David Donoho and Jiashun Jin. Higher criticism thresholding: Optimal feature selection when useful features are rare and weak. *Proceedings of the National Academy of Sciences*, 105(39):14790–14795, 2008.

[20] Christos Thrampoulidis, Samet Oymak, and Mahdi Soltanolkotabi. Theoretical insights into multiclass classification: A high-dimensional asymptotic view. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 8907–8920. Curran Associates, Inc., 2020.

[21] Zeyu Deng, Abla Kammoun, and Christos Thrampoulidis. A model of double descent for high-dimensional binary linear classification. *Preprint arXiv:1911.05822*, 2020.

[22] Xiaoyi Mai, Zhenyu Liao, and Romain Couillet. A large scale analysis of logistic regression: Asymptotic performance and new insights. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3357–3361, 2019.

[23] Xiaoyi Mai and Zhenyu Liao. High dimensional classification via regularized and unregularized empirical risk minimization: Precise error and optimal loss. *Preprint arXiv:1905.13742*, 2020.

[24] Edgar Dobriban and Stefan Wager. High-dimensional asymptotics of prediction: Ridge regression and classification. *The Annals of Statistics*, 46(1):247 – 279, 2018.

[25] Ganesh Kini and Christos Thrampoulidis. Analytic study of double descent in binary classification: The impact of loss. *Preprint arXiv:2001.11572*, 2020.

[26] Houssem Sifaou, Abla Kammoun, and Mohamed-Slim Alouini. Phase transition in the hard-margin support vector machines. In *2019 IEEE 8th International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*, pages 415–419, 2019.

[27] Ke Wang and Christos Thrampoulidis. Binary classification of gaussian mixtures: Abundance of support vectors, benign overfitting and regularization. 2021.

[28] Niladri S. Chatterji and Philip M. Long. Finite-sample analysis of interpolating linear classifiers in the overparameterized regime. *Preprint arXiv:2004.12019*, 2021.

[29] Yuan Cao, Quanquan Gu, and Mikhail Belkin. Risk bounds for over-parameterized maximum margin classification on sub-gaussian mixtures. *Preprint arXiv:2104.13628*, 2021.

[30] Fariborz Salehi, Ehsan Abbasi, and Babak Hassibi. The impact of regularization on high-dimensional logistic regression. *Preprint arXiv:1906.03761*, 2019.

[31] Christos Thrampoulidis, Ehsan Abbasi, and Babak Hassibi. Precise error analysis of regularized m-estimators in high dimensions. *IEEE Transactions on Information Theory*, 64(8):5592–5628, 2018.

[32] Mihailo Stojnic. A framework to characterize performance of lasso algorithms. *Preprint arXiv:1303.7291*, 2013.

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

[34] Florent Krzakala, Marc Mézard, Francois Sausset, Yifan Sun, and Lenka Zdeborová. Probabilistic reconstruction in compressed sensing: algorithms, phase diagrams, and threshold achieving matrices. *Journal of Statistical Mechanics: Theory and Experiment*, 2012(08):P08009, 2012.

[35] David L Donoho, Adel Javanmard, and Andrea Montanari. Information-theoretically optimal compressed sensing via spatial coupling and approximate message passing. *IEEE transactions on information theory*, 59(11):7434–7464, 2013.

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

[37] Raphael Berthier, Andrea Montanari, and Phan-Minh Nguyen. State evolution for approximate message passing with non-separable functions. *Information and Inference: A Journal of the IMA*, 9(1):33–79, 2020.

[38] Andre Manoel, Florent Krzakala, Marc Mézard, and Lenka Zdeborová. Multi-layer generalized linear estimation. In *2017 IEEE International Symposium on Information Theory (ISIT)*, pages 2098–2102. IEEE, 2017.

[39] Jiashun Jin. Impossibility of successful classification when useful features are rare and weak. *Proceedings of the National Academy of Sciences*, 106(22):8859–8864, 2009.

[40] Jun Shao, Yazhen Wang, Xinwei Deng, and Sijian Wang. Sparse linear discriminant analysis by thresholding for high dimensional data. *The Annals of Statistics*, 39(2):1241 – 1265, 2011.

[41] Qing Mai, Hui Zou, and Ming Yuan. A direct approach to sparse discriminant analysis in ultra-high dimensions. *Biometrika*, 99(1):29–42, 12 2012.

[42] Yanfang Li and Jinzhu Jia. L1 least squares for sparse high-dimensional LDA. *Electronic Journal of Statistics*, 11(1):2499 – 2518, 2017.

[43] Thomas M Cover. Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition. *IEEE transactions on electronic computers*, (3):326–334, 1965.

[44] Elizabeth Gardner. The space of interactions in neural network models. *Journal of physics A: Mathematical and general*, 21(1):257, 1988.

[45] Arthur Jacot, Berfin Şimşek, Francesco Spadaro, Clément Hongler, and Franck Gabriel. Kernel alignment risk estimator: Risk prediction from training data. *Preprint arXiv:2006.09796*, 2020.

[46] Blake Bordelon, Abdulkadir Canatar, and Cengiz Pehlevan. Spectrum dependent learning curves in kernel regression and wide neural networks. In *International Conference on Machine Learning*, pages 1024–1034. PMLR, 2020.

[47] Neal Parikh and Stephen Boyd. Proximal algorithms. *Foundations and Trends in optimization*, 1(3):127–239, 2014.

[48] Heinz H Bauschke, Patrick L Combettes, et al. *Convex analysis and monotone operator theory in Hilbert spaces*, volume 408. Springer, 2011.

[49] Michael Celentano, Andrea Montanari, and Yuting Wei. The Lasso with general Gaussian designs with applications to hypothesis testing. *Preprint arXiv:2007.13716*, 2020.

[50] Erwin Bolthausen. An iterative construction of solutions of the TAP equations for the Sherrington–Kirkpatrick model. *Communications in Mathematical Physics*, 325(1):333–366, 2014.

[51] Mohsen Bayati and Andrea Montanari. The LASSO risk for Gaussian matrices. *IEEE Transactions on Information Theory*, 58(4):1997–2017, 2011.

[52] Cedric Gerbelot, Alia Abbara, and Florent Krzakala. Asymptotic Errors for Teacher-Student Convex Generalized Linear Models (or: How to Prove Kabashima's Replica Formula). *Preprint arXiv:2006.06581*, 2020.

[53] Benjamin Aubin, Antoine Maillard, Jean Barbier, Florent Krzakala, Nicolas Macris, and Lenka Zdeborová. The committee machine: Computational to statistical gaps in learning a two-layers neural network. *Journal of Statistical Mechanics: Theory and Experiment*, 2019(12):124023, 2019.

[54] Florent Krzakala, Marc Mézard, François Sausset, YF Sun, and Lenka Zdeborová. Statistical-physics-based reconstruction in compressed sensing. *Physical Review X*, 2(2):021005, 2012.

[55] Cynthia Rush and Ramji Venkataramanan. Finite sample analysis of approximate message passing algorithms. *IEEE Transactions on Information Theory*, 64(11):7264–7286, 2018.

[56] Mohsen Bayati, Marc Lelarge, Andrea Montanari, et al. Universality in polytope phase transitions and message passing algorithms. *Annals of Applied Probability*, 25(2):753–822, 2015.

[57] Wei-Kuo Chen and Wai-Kit Lam. Universality of approximate message passing algorithms. *Electronic Journal of Probability*, 26:1–44, 2021.

[58] Scott Shaobing Chen, David L. Donoho, and Michael A. Saunders. Atomic decomposition by basis pursuit. *SIAM Journal on Scientific Computing*, 20(1):33–61, 1998.

[59] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *The Journal of Machine Learning Research*, 12:2825–2830, 2011.

[60] Yann LeCun and Corinna Cortes. *ATT Labs [Online]*, 2010. Database released under CC BY-SA 3.0 license at http://yann.lecun.com/exdb/mnist/.

[61] Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-MNIST: a novel image dataset for benchmarking machine learning algorithms. *Preprint arXiv:1708.07747*, 2017. Database released under MIT licence at https://github.com/zalandoresearch/fashion-mnist.

[62] Ali Rahimi and Benjamin Recht. Random Features for Large-Scale Kernel Machines. In *NIPS*, pages 1177–1184, 2007.

[63] Heinz H Bauschke, Jonathan M Borwein, and Patrick L Combettes. Bregman monotone optimization algorithms. *SIAM Journal on control and optimization*, 42(2):596–636, 2003.

[64] Heinz H Bauschke, Minh N Dao, and Scott B Lindstrom. Regularizing with bregman–moreau envelopes. *SIAM Journal on Optimization*, 28(4):3208–3228, 2018.

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

[66] Cédric Gerbelot and Raphaël Berthier. Graph-based approximate message passing iterations. *To appear*, 2021.

[67] Marc Mézard, Giorgio Parisi, and Miguel Virasoro. *Spin glass theory and beyond: An Introduction to the Replica Method and Its Applications*, volume 9. World Scientific Publishing Company, 1987.

[68] Saharon Rosset, Ji Zhu, and Trevor Hastie. Margin maximizing loss functions. In *NIPS*, pages 1237–1244, 2003.