Supplementary information for

Learning Gaussian Mixtures with Generalised Linear Models: Precise Asymptotics in High-dimensions
Bruno Loureiro, Gabriele Sicuro, Cédric Gerbelot, Alessandro Pacco, Florent Krzakala, Lenka Zdeborová

A Proof of the main results

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 \( \phi: \mathbb{R}^{n \times K} \rightarrow \mathbb{R}^{n \times K} \) is called a pseudo-Lipschitz of order \( k \) if there exists a constant \( L(k, K) \) such that for any \( x, y \in \mathbb{R}^{n \times K} \),

\[
\frac{\| \phi(x) - \phi(y) \|_F}{\sqrt{n}} \leq L(k, K) \left( 1 + \left( \frac{\| x \|_F}{\sqrt{n}} \right)^{k-1} + \left( \frac{\| y \|_F}{\sqrt{n}} \right)^{k-1} \right) \frac{\| x - y \|_F}{\sqrt{n}}
\]

(14)

where \( \| \cdot \|_F \) denotes the Frobenius norm. Since \( K \) will be kept finite, it can be absorbed in any of the constants.

For example, the function \( f: \mathbb{R}^n \rightarrow \mathbb{R}, x \mapsto \frac{1}{n} \| x \|_2^2 \) is pseudo-Lipschitz of order 2.

Moreau envelopes and Bregman proximal operators — In our proof, we will also frequently use the notions of Moreau envelopes and proximal operators, see e.g. [47, 48]. These elements of convex analysis are often encountered in recent works on high-dimensional asymptotics of convex problems, and more detailed analysis of their properties can be found for example in [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 \( \phi: \mathbb{R}^{d \times K} \rightarrow \mathbb{R} \), a given matrix \( X \in \mathbb{R}^{d \times K} \) and a given symmetric positive definite matrix \( V \in \mathbb{R}^{K \times K} \) with bounded spectral norm, we will consider operators of the type

\[
\arg\min_{T \in \mathbb{R}^{d \times K}} \left\{ f(T) + \frac{1}{2} \text{tr} \left( (T - X)V^{-1}(T - X)^\top \right) \right\}
\]

(15)

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

\[
\text{PROX}_{f(\cdot V^{1/2})}(XV^{-1/2})V^{1/2} \in \mathbb{R}^{d \times K}
\]

(16)

or as a Bregman proximal operator [64] defined with the Bregman distance induced by the strictly convex, coercive function (for positive definite \( V \))

\[
X \mapsto \frac{1}{2} \text{tr}(XV^{-1}X^\top)
\]

(17)

which justifies the use of the Bregman resolvent

\[
\arg\min_{T \in \mathbb{R}^{d \times K}} \left\{ f(T) + \frac{1}{2} \text{tr} \left( (T - X)V^{-1}(T - X)^\top \right) \right\} = (\text{Id} + \partial f(\cdot)V)^{-1}(X)
\]

(18)
Many of the usual or similar properties to that of standard proximal operators (i.e. firm non-expansiveness, link with Moreau/Bregman envelopes,…) hold for Bregman proximal operators defined with the function (17), see e.g. [64, 65]. In particular, we will be using the equivalent notion to firmly nonexpansive operators for Bregman proximity operators, called $D$-firm operators. Consider the Bregman proximal defined with a differentiable, strictly convex, coercive function $g : \mathcal{X} \rightarrow \mathbb{R}$, where $\mathcal{X}$ is a given input Hilbert space. Let $T$ be the associated Bregman proximal of a given convex function $f : \mathcal{X} \rightarrow \mathbb{R}$, i.e., for any $x \in \mathcal{X}$

$$T(x) = \text{argmin}_{y \in \mathcal{X}} \{f(x) + D_g(x, y)\}$$  \hfill (19)

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

$$\langle Tx - Ty, \nabla g(Tx) - \nabla g(Ty) \rangle \leq \langle Tx - Ty, \nabla g(x) - \nabla g(y) \rangle$$  \hfill (20)

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

**Gaussian concentration** — Gaussian concentration properties are at the root of this proof. Such properties are reviewed in more 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 $\{A_k \in \mathbb{R}^{n_k \times d_k}\}_{k \in [K]}$ we will use the following notation:

$$\begin{bmatrix}
A_1 \\
A_2 \\
\vdots \\
A_K
\end{bmatrix}
\equiv
\begin{bmatrix}
A_{k_1}^{K} \\
A_{k_2}^{K} \\
\vdots \\
A_{k_K}^{K}
\end{bmatrix}
\in\mathbb{R}^{(\sum_{k=1}^{K}n_k) \times (\sum_{k=1}^{K}d_k)} \hfill (21)$$

where the terms denoted by $(\ast)$ will be zero most of the time.

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

$$\phi(X) = 
\begin{bmatrix}
\phi^1(X) \\
\phi^2(X) \\
\vdots \\
\phi^K(X)
\end{bmatrix}
\in \mathbb{R}^{d \times K} \hfill (22)$$

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

$$\frac{\partial \phi^i}{\partial X^j}(X) = 
\begin{bmatrix}
\frac{\partial \phi^i_1(X)}{\partial X_{j1}} & \cdots & \frac{\partial \phi^i_1(X)}{\partial X_{jK}} \\
\vdots & \ddots & \vdots \\
\frac{\partial \phi^i_K(X)}{\partial X_{j1}} & \cdots & \frac{\partial \phi^i_K(X)}{\partial X_{jK}}
\end{bmatrix}
\in \mathbb{R}^{K \times K} \hfill (23)$$

For a given matrix $Q \in \mathbb{R}^{K \times K}$, we write $Z \in \mathbb{R}^{n \times K} \sim \mathcal{N}(0, Q \otimes I_n)$ to denote that the lines of $Z$ are sampled i.i.d. from $\mathcal{N}(0, Q)$. Note that this is equivalent to saying that $Z = \tilde{Z}Q^{1/2}$ where $\tilde{Z} \in \mathbb{R}^{n \times K}$ is an i.i.d. standard normal random matrix. The notation $\overset{D}{\rightarrow}$ 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 [66]. 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 [67]. Consider a sequence Gaussian matrices $A(n) \in \mathbb{R}^{n \times d}$ 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

$$\{h_t : \mathbb{R}^{n \times K} \rightarrow \mathbb{R}^{n \times K}\}_{t \in \mathbb{N}} \quad \{e_t : \mathbb{R}^{d \times K} \rightarrow \mathbb{R}^{d \times K}\}_{t \in \mathbb{N}} \hfill (24)$$

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

$$\lim_{d \rightarrow \infty} \frac{1}{d} \|e_0(u^0)^\top e_0(u^0)\|_F \hfill (25)$$
exists and it is finite, and recursively define:

\[ u^{t+1} = A^\top h_t(v^t) - e_t(u^t)(h_t)^\top \]  
\[ v^t = A e_t(u^t) - h_{t-1}(v^{t-1})(e_t)^\top \]  (26)  (27)

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

\[ \langle h_t \rangle = \frac{1}{d} \sum_{i=1}^d \frac{\partial h_t}{\partial v_i}(v^t) \in \mathbb{R}^{K \times K} \quad \langle e_t \rangle = \frac{1}{d} \sum_{i=1}^d \frac{\partial e_t}{\partial u_i}(u^t) \in \mathbb{R}^{K \times K} \]  (28)

We define now the state evolution recursion on two sequences of matrices \( \{Q_{r,s}\}_{s,r \geq 0} \) and \( \{Q_{r,s}\}_{s,r \geq 1} \) initialized with \( Q_{0,0} = \lim_{d \to \infty} \frac{1}{d} e_0(u^0)^\top e_0(u^0) \):

\[ Q_{t+1,s} = Q_{s,t+1} = \lim_{d \to \infty} \frac{1}{d} E \left[ e_s(\hat{Z}^t) e_t(\hat{Z}^{t+1}) \right] \in \mathbb{R}^{K \times K} \]  (29)  \[ \hat{Q}_{t+1,s+1} = \hat{Q}_{s+1,t+1} = \lim_{d \to \infty} \frac{1}{d} E \left[ h_s(\hat{Z}^t) h_t(\hat{Z}^{t+1}) \right] \in \mathbb{R}^{K \times K} \]  (30)

where \( \{Z^0, \ldots, Z^{t-1}\} \sim N(0, \{Q_{r,s}\}_{0 \leq r \leq t-1 \in \mathbb{I}_n}) \), \( \{\hat{Z}^1, \ldots, \hat{Z}^t\} \sim N(0, \{\hat{Q}_{r,s}\}_{1 \leq r \leq s \in \mathbb{I}_n}) \). Then the following holds

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

\[ \phi_n(u^0, v^0, u^1, v^1, \ldots, u^t, v^t) \sim E \left[ \phi_n \left( u^0, Z^0, \hat{Z}^1, \ldots, Z^{t-1}, \hat{Z}^t \right) \right] \]  (31)

**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 (26), but this time with a Gaussian matrix defined as:

\[ A = \begin{bmatrix} A_1 & A_2 & \cdots & A_K \end{bmatrix} \in \mathbb{R}^{n \times Kd} \]  (32)

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

\[ A^\top h_t(v^t) = \begin{bmatrix} A_1^\top h_{1,t}(v^t) \\ A_2^\top h_{2,t}(v^t) \\ \vdots \\ A_K^\top h_{K,t}(v^t) \end{bmatrix} \in \mathbb{R}^{Kd \times K} \quad A e_t(u^t) = \begin{bmatrix} A_1 e_{1,t}(u^t) \\ A_2 e_{2,t}(u^t) \\ \vdots \\ A_K e_{K,t}(u^t) \end{bmatrix} \in \mathbb{R}^{n \times K} \]  (33)

where the blocks \( h_{k,t}(v^t) \in \mathbb{R}^{n_k \times K} \), \( e_{k,t}(u^t) \in \mathbb{R}^{d \times K} \) 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 \( A \) and non-linearities defined as:

\[ e_t : \mathbb{R}^{Kd \times K^2} \to \mathbb{R}^{Kd \times K^2} \]

\[ h_t : \mathbb{R}^{n \times K} \to \mathbb{R}^{n \times K} \]

\[ \begin{bmatrix} e_{1,t}(\bullet) \\ e_{2,t}(\bullet) \\ \vdots \\ e_{K,t}(\bullet) \end{bmatrix} \in \mathbb{R}^{Kd \times K^2} \]  (34)  \[ \begin{bmatrix} h_{1,t}(\bullet) \\ h_{2,t}(\bullet) \\ \vdots \\ h_{K,t}(\bullet) \end{bmatrix} \in \mathbb{R}^{n \times K} \]  (35)
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 $W \in \mathbb{R}^{d \times K}$ instead of $\mathbb{R}^{K \times d}$. The normalized (so that the cost does not diverge with the dimension) problem (2) then reads:

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

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

\[
\left( Y, \frac{1}{\sqrt{d}} X W + b \right) \mapsto \sum_{\nu=1}^n \ell \left( y^{\nu}, \frac{W \tilde{x}^{\nu} + b}{\sqrt{d}} \right),
\]

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

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

We will write $L_k$ the application of $\ell$ 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 $X \in \mathbb{R}^{n \times d}$, separating the mean part $Y \tilde{M}$ and centered Gaussian part :

\[
X = Y M + \tilde{Z} \Sigma \in \mathbb{R}^{n \times d}
\]

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

\[
\tilde{Z} = \begin{bmatrix} Z_1 \\ Z_2 \\ \vdots \\ Z_K \end{bmatrix} \in \mathbb{R}^{n \times Kd} \quad \Sigma = \begin{bmatrix} \Sigma_{1,1}^{1/2} \\ \Sigma_{2,2}^{1/2} \\ \vdots \\ \Sigma_{K,K}^{1/2} \end{bmatrix} \in \mathbb{R}^{K d \times d}.
\]

Here $(Z_1, \ldots, Z_K) \in \mathbb{R}^{n_1 \times d} \times \cdots \times \mathbb{R}^{n_K \times d}$ are independent, i.i.d. standard normal matrices.

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

\[
X W = Y M W + \tilde{Z} \Sigma W = \begin{bmatrix} Y_1 M W + Z_1 \Sigma_1^{1/2} W \\ \vdots \\ Y_K M W + Z_K \Sigma_K^{1/2} W \end{bmatrix} \in \mathbb{R}^{n \times K}
\]

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

\[
\tilde{W} = \Sigma W \quad \Rightarrow \quad W = \Sigma^+ \tilde{W},
\]

where

\[
\Sigma^+ \equiv \left( \sum_{k=1}^K \Sigma_k \right)^{-1} \Sigma^T
\]

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

\[
\inf_{\tilde{W} \in \mathbb{R}^{n \times K}} \sum_{k=1}^K L_k \left( \frac{1}{\sqrt{d}} Y_k M W + \frac{1}{\sqrt{d}} Z_k \tilde{W}_k, b \right) + r \left( \Sigma^+ \tilde{W} \right)
\]
Introducing the order parameter $m = \frac{1}{\sqrt{d}}MW \in \mathbb{R}^{K \times K}$, we reformulate Eq. (44) as a constrained optimization problem:

$$\inf_{m, \tilde{W}, b} \sum_{k=1}^{K} L_k \left( \frac{1}{\sqrt{d}} Y_k m + \frac{1}{\sqrt{d}} Z_k \tilde{W}_k \right) + r \left( \Sigma^+ \tilde{W} \right)$$

$$\text{s.t.} \quad \frac{1}{\sqrt{d}} M \Sigma^+ \tilde{W} = m$$

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

$$\inf_{\tilde{W} \in \mathbb{R}^{K \times K}} \sup_{m, \tilde{b}} \sum_{k=1}^{K} L_k \left( Y_k m + \frac{1}{\sqrt{d}} Z_k \tilde{W}_k \right) + r \left( \Sigma^+ \tilde{W} \right) + \text{tr} \left( m^T \left( m - \frac{1}{\sqrt{d}} M \Sigma^+ \tilde{W} \right) \right).$$

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 $\tilde{W}$ for fixed $m, \tilde{b}$:

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

where we defined the loss term

$$\tilde{L} : \mathbb{R}^{n \times K} \rightarrow \mathbb{R}$$

$$\frac{1}{\sqrt{d}} \tilde{Z} \tilde{W} \mapsto \sum_{k=1}^{K} \left( Y_k m + \frac{1}{\sqrt{d}} Z_k \tilde{W}_k \right) = \sum_{k=1}^{K} \sum_{i=1}^{n_k} \left( \left[ Y_k m + \frac{1}{\sqrt{d}} Z_k \tilde{W}_k \right] \right)_i$$

and the regularisation term

$$\tilde{r} : \mathbb{R}^{K \times K} \rightarrow \mathbb{R}$$

$$\tilde{W} \mapsto r \left( \Sigma^+ \tilde{W} \right) + \text{tr} \left( m^T \left( m - \frac{1}{\sqrt{d}} M \Sigma^+ \tilde{W} \right) \right)$$

where $\Sigma^+ \tilde{W} = \sum_{k=1}^{K} \Sigma_k^{1/2} \tilde{W}_k$ and $\tilde{Z} = [Z_k]_{k=1}^{K} \in \mathbb{R}^{n \times K}$ is an i.i.d. standard normal block diagonal matrix as in Eq. (40).

### A.3 Finding the AMP sequence

We now need to find an AMP iteration relating to $\tilde{W}$ that solve the optimization problem in Eq. (47). 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. (47);
- the update rule of the sequence should have the form Eq. (26) 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 $\tilde{Z}$, such that the $Z_k \in \mathbb{R}^{n_k \times d}$ have i.i.d. $N(0, 1/d)$ elements.

**Resolvent of the loss term** — Going back to problem Eq. (47), its optimality condition will look like:

$$\tilde{Z}^T \partial \tilde{L}(\tilde{Z} \tilde{W}) + \partial \tilde{r}(\tilde{W}) = 0 \iff \begin{bmatrix} Z_1^T & Z_2^T & \cdots & Z_K^T \end{bmatrix} \begin{bmatrix} \partial \tilde{L}_1(Z_1 \tilde{W}_1) \\ \partial \tilde{L}_2(Z_2 \tilde{W}_2) \\ \vdots \\ \partial \tilde{L}_K(Z_K \tilde{W}_K) \end{bmatrix} + \partial \tilde{r}(\tilde{W}) = 0$$

(49)
where each $Z_k \in \mathbb{R}^{n_k \times d}$, and the subdifferential of $\tilde{L}$ is separable across blocks of size $n_k \times d$, and $\partial \tilde{r}(\mathbf{W}) \in \mathbb{R}^{Kd \times K}$. Following the intuition of spatial coupling, we introduce the full matrix $Z \in \mathbb{R}^{n \times Kd}$, with i.i.d. $\mathcal{N}(0, d^{-1})$ entries. The optimality condition can then be written on the diagonal of a $Kd \times K^2$ matrix:

\[
Z^\top \begin{bmatrix}
\partial \tilde{L}_1(Z_1 \tilde{W}_1) & \partial \tilde{L}_2(Z_2 \tilde{W}_2) & \cdots & \partial \tilde{L}_K(Z_K \tilde{W}_K) \\
0 & 0 & \cdots & 0 \\
0 & 0 & \cdots & 0 \\
0 & 0 & \cdots & 0 \\
\end{bmatrix} = 0
\]  

(50)

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

\[
\begin{align*}
&\left( Z_k^\top \partial \tilde{L}_k(Z_k \tilde{W}_k) \right)_{k=1}^K + \left[ \partial \tilde{r}(\mathbf{W}) \right]_{k=1}^K = 0 \\
\iff&\left( Z_k^\top \partial \tilde{L}_k(Z_k \tilde{W}_k) + Z_k^\top Z_k \tilde{W}_k S_k^{-1} \right)_{k=1}^K + \left[ \partial \tilde{r}(\mathbf{W}) \right]_{k=1}^K = \left[ Z_k^\top Z_k \tilde{W}_k S_k^{-1} \right]_{k=1}^K 
\end{align*}
\]  

(51)

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

\[
U_k \equiv \partial \tilde{L}_k(Z_k \tilde{W}_k) S_k + Z_k \tilde{W}_k \iff Z_k \tilde{W}_k = R_{L_k, S_k}(U_k)
\]  

(52)

where

\[
R_{L_k, S_k}(U_k) = (\text{Id} + \partial \tilde{L}_k(\mathbf{\bullet}) S_k)^{-1}(U_k)
\]

\[
= \arg\min_{\tilde{T} \in \mathbb{R}^{n \times K}} \left\{ \tilde{L}_k(\tilde{T}) + \frac{1}{2} \text{tr} \left( (\tilde{T} - U_k) S_k^{-1} (\tilde{T} - U_k)^\top \right) \right\}
\]

\[
= \arg\min_{\tilde{T} \in \mathbb{R}^{n \times K}} \left\{ L_k(T) + \frac{1}{2} \text{tr} \left( (T - (Y_k m + U_k)) S_k^{-1} (T - (Y_k m + U_k))^\top \right) \right\} - Y_k m.
\]  

(53)

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

\[
\begin{align*}
&\left( Z_k^\top U_k S_k^{-1} \right)_{k=1}^K + \left[ \partial \tilde{r}(\mathbf{W})_k \right]_{k=1}^K = \left[ Z_k^\top R_{L_k, S_k}(U_k) S_k^{-1} \right]_{k=1}^K \\
\iff&\left[ Z_k^\top \left( U_k - R_{L_k, S_k}(U_k) \right) S_k^{-1} \right]_{k=1}^K + \left[ \partial \tilde{r}(\mathbf{W})_k \right]_{k=1}^K = 0
\end{align*}
\]  

(54)

**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:

\[
\left[ \partial \tilde{r}(\mathbf{W})_k \right]_{k=1}^K + \left[ \tilde{W}_k S_k^{-1} \right]_{k=1}^K = \left[ \tilde{W}_k S_k^{-1} \right]_{k=1}^K
\]  

(55)

At this point it becomes clear that we cannot consider the resolvent as acting on $\tilde{W} \in \mathbb{R}^{Kd \times K}$ otherwise there could be only one $\tilde{S} \in \mathbb{R}^{K \times K}$ and there would be a mismatch with the expected form of the Onsager terms. As specified by the definitions Eq.(48), the subdifferential of $\tilde{r}$ is acting on the whole block diagonal matrix $[\tilde{W}_k]_{k=1}^K$, by way of summation due to the action of the pseudo-inverse.
\( \Sigma^+ \). We can thus consider its proximal acting on \( \mathbb{R}^{d \times K^2} \) as \( [\hat{W}_1, \hat{W}_2, ..., \hat{W}_K] \) (note that we could have also worked directly with a block diagonal matrix in \( \mathbb{R}^{K d \times K^2} \)). Proceeding in this way, we can directly write our expression as an application parametrized by another set of positive definite matrices \( \{ \hat{S}_k \}_{k \in [K]} \):

\[
\hat{U} = \left( \text{Id} + \partial \hat{r}(\bullet) \hat{S} \right) (\hat{W}) \quad \hat{W} = R_{\hat{f}, \hat{S}}(\hat{U})
\]

(56)

where

\[
R_{\hat{f}, \hat{S}}(\hat{U}) = \left( \text{Id} + \partial \hat{r}(\bullet) \hat{S} \right)^{-1}(\hat{U}) = \arg\min_{T \in \mathbb{R}^{d \times K^2}} \left\{ \hat{r}(T) + \frac{1}{2} \text{tr} \left( (T - \hat{U}) \hat{S}^{-1} (T - \hat{U})^\top \right) \right\}
\]

(57)

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

\[
\hat{U} \hat{S}^{-1} = R_{\hat{f}, \hat{S}}(\hat{U}) \hat{S}^{-1} \iff \left[ \hat{U}_k \hat{S}_k^{-1} \right]_{k=1}^K = \left[ R_{\hat{f}, \hat{S}_k}(\hat{U}) \hat{S}_k^{-1} \right]_{k=1}^K
\]

(58)

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

\[
\min_{\{T_k\}_{k \in [K]} \in [\mathbb{R}^{d \times K}]^K} \tau(C^{-1} \sum_{k=1}^K \Sigma_k^{1/2} T_k) + \frac{1}{2} \text{tr} \left( (T - \hat{U}) \hat{S}^{-1} (T - \hat{U})^\top \right) + \text{tr} \left( \hat{m}^\top \left( m - \frac{1}{\sqrt{d}} M \Sigma^+ T \right) \right)
\]

(59)

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

\[
\min_{\hat{T} \in [\mathbb{R}^{d \times K}]} \tau(\hat{T}) + \frac{1}{2} \text{tr} \left( (T - \hat{U}) \hat{S}^{-1} (T - \hat{U})^\top \right) + \text{tr} \left( \hat{m}^\top \left( m - \frac{1}{\sqrt{d}} M \hat{T} \right) \right)
\]

(60)

s.t. \( \hat{T} = C^{-1} \sum_{k=1}^K \Sigma_k^{1/2} T_k \)

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 \( \lambda \in \mathbb{R}^{d \times K} \), we get the equivalent representation:

\[
\min_{\{T_k\}_{k \in [K]} \in [\mathbb{R}^{d \times K}]^K} \max_{\lambda \in \mathbb{R}^{d \times K}} \tau(\hat{T}) + \sum_{k=1}^K \text{tr} \left( (T_k - \hat{U}_k) \hat{S}_k^{-1} (T_k - \hat{U}_k)^\top \right)
\]

\[
+ \text{tr} \left( \lambda^\top \left( \hat{T} - C^{-1} \sum_{k=1}^K \Sigma_k^{1/2} T_k \right) \right) + \text{tr} \left( \hat{m}^\top \left( m - \frac{1}{\sqrt{d}} M \hat{T} \right) \right).
\]

(61)

The optimality condition for this problem reads:

\[
\partial T : \quad \partial \tau(\hat{T}) + \lambda - \frac{1}{\sqrt{d}} M^\top \hat{m} = 0
\]

(62)

\[
\partial T : \quad (T_k - U_k) \hat{S}_k^{-1} = \Sigma_k^{1/2} C^{-1} \lambda \quad \forall k \in [K]
\]

(63)

\[
\partial \lambda : \quad \hat{T} = C^{-1} \sum_{k=1}^K \Sigma_k^{1/2} T_k
\]

(64)
Using the gradient condition on $T$, we get
\begin{equation}
\sum_{k=1}^{K} \Sigma_k^{1/2} (T_k - \hat{U}_k) \tilde{S}_k^{-1} = \lambda
\end{equation}
(65)

The constraint $\hat{T} = C^{-1} \sum_{k=1}^{K} \Sigma_k^{1/2} T_k$ is solved by $T_k = \Sigma_k^{1/2} \hat{T}$ which gives the solution for $\lambda$
\begin{equation}
\lambda = \sum_{k=1}^{K} \Sigma_k^{1/2} (\Sigma_k^{1/2} \hat{T} - \hat{U}_k) \tilde{S}_k^{-1} = \sum_{k=1}^{K} \Sigma_k \hat{T} \tilde{S}_k^{-1} - \sum_{k=1}^{K} \Sigma_k^{1/2} \hat{U}_k \tilde{S}_k^{-1}
\end{equation}
(66)

and prescribes the following form for $\hat{T}$, as solution to the problem
\begin{equation}
\partial r(\hat{T}) + \sum_{k=1}^{K} \Sigma_k \hat{T} \tilde{S}_k^{-1} - \sum_{k=1}^{K} \Sigma_k^{1/2} \hat{U}_k \tilde{S}_k^{-1} - \frac{1}{\sqrt{d}} M^T \hat{m} = 0
\end{equation}
\begin{equation}
\iff \text{argmin}_{\hat{T}} r(\hat{T}) + \frac{1}{2} \sum_{k=1}^{K} \Sigma_k \hat{T} \tilde{S}_k^{-1} \hat{T} - \left( \sum_{k=1}^{K} \Sigma_k^{1/2} \hat{U}_k \tilde{S}_k^{-1} + \frac{1}{\sqrt{d}} M^T \hat{m} \right) \hat{T}
\end{equation}
(67)

We then recover $T$ from $T = \Sigma \hat{T}$. Thus, defining the function
\begin{equation}
\eta : \mathbb{R}^{d \times K^2} \rightarrow \mathbb{R}^{d \times K}
\end{equation}

$$U \mapsto \text{argmin}_{\hat{T}} r(\hat{T}) + \frac{1}{2} \sum_{k=1}^{K} \Sigma_k \hat{T} \tilde{S}_k^{-1} \hat{T} - \left( \sum_{k=1}^{K} \Sigma_k^{1/2} \hat{U}_k \tilde{S}_k^{-1} + \frac{1}{\sqrt{d}} M^T \hat{m} \right) \hat{T}
\end{equation}
(68)

the block decomposition of the resolvent for the regularizer reads:
\begin{equation}
R_{\hat{r}, \hat{S}, k}(\hat{U}) = \Sigma_k^{1/2} \eta(\hat{U})
\end{equation}
(69)

**Matching the optimality condition with the AMP fixed point** The global optimality condition then reads:
\begin{equation}
\left[ Z_k^T \left( R_{L_k, s_k}(U_k) - U_k \right) S_k^{-1} \right]_{k=1}^{K} = \left[ (\hat{U}_k - R_{\hat{r}, \hat{S}, k}(\hat{U})) \hat{S}_k^{-1} \right]_{k=1}^{K}
\end{equation}
(70)

\begin{equation}
\left[ Z_k R_{\hat{r}, \hat{S}, k}(\hat{U}) \right]_{k=1}^{K} = \left[ R_{L_k, s_k}(U_k) \right]_{k=1}^{K}
\end{equation}
(71)

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.(26) reads:
\begin{equation}
\text{Id} + e(u) (h')^T = Z^T h(v)
\end{equation}
(72)
\begin{equation}
\text{Id} + h(v) (e')^T = Ze(u)
\end{equation}
(73)

Matching this fixed point with the optimality condition Eq.(70) suggests the following mapping:
\begin{equation}
h_k(U_k) = \left( R_{L_k, s_k}(U_k) - U_k \right) S_k^{-1}, \quad S_k = (e'_k), \quad \hat{S}_k = -(h'_k)^{-1},
\end{equation}
\begin{equation}
e_k(\hat{U}) = R_{\hat{r}, \hat{S}, k}(\hat{U} \hat{S}),
\end{equation}
(74)

where we redefined $\hat{U} \equiv \hat{U} \hat{S}$ in (56), and the subscripts on the non-linearities are block indexes.
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 \( \hat{m}, m, b \) and any \( u \in \mathbb{R}^{d \times K^2}, v \in \mathbb{R}^{n \times K} \):

\[
\begin{align*}
e_t : \mathbb{R}^{K_d \times K^2} &\to \mathbb{R}^{K_d \times K^2} \\
u &\mapsto \left[ \begin{array}{c} e_{1,t}(u) \\ e_{2,t}(u) \\ \vdots \\ e_{K,t}(u) \end{array} \right] \in \mathbb{R}^{K_d \times K^2}
\end{align*}
\]

(75)

\[
\begin{align*}
h_t : \mathbb{R}^{n \times K} &\to \mathbb{R}^{n \times K^2} \\
v &\mapsto \left[ \begin{array}{c} h_{1,t}(v_1) \\ h_{2,t}(v_2) \\ \vdots \\ h_{K,t,t}(v_K) \end{array} \right] \in \mathbb{R}^{n \times K^2}
\end{align*}
\]

(76)

where \( Y_k \in \mathbb{R}^{n_k \times K} \) and

\[
h_{k,t} : \mathbb{R}^{n_k \times K} \to \mathbb{R}^{n_k \times K}
\]

\[
v_k \mapsto \left( R_{L_k, V_k, t}(v_k) - v_k \right) (V_{k,t})^{-1}
\]

\[
= \left( \underset{T \in \mathbb{R}^{d \times K}}{\arg\min} \left\{ L_k(T) + \frac{1}{2} \text{tr} \left( (T - v_k)(V_{k,t})^{-1}(T - v_k)^\top \right) \right\} - v_k \right) (V_{k,t})^{-1}
\]

\[
= \left( \text{Prox}_{L_k(u,v_k^{1/2})}(Y_k \hat{m} + v_k)(V_{k,t})^{-1/2}(V_{k,t})^{1/2} - (Y_k \hat{m} + v_k)(V_{k,t})^{-1} \right)
\]

(77)

\[
e_{k,t} : \mathbb{R}^{d \times K^2} \to \mathbb{R}^{d \times K}
\]

\[
u \mapsto \Sigma_k^{1/2} \arg\min_{T \in \mathbb{R}^{d \times K}} \left| |r(T) + \frac{1}{2} \sum_{k=1}^{K} \Sigma_k \tilde{T}V_{k,t} \tilde{T} - \left( \sum_{k=1}^{K} \Sigma_k^{1/2} u_k + \frac{1}{\sqrt{d}} \hat{m} \right) \right|^{-\frac{1}{2}}
\]

(78)

where \((V_t, \hat{V}_t) \in \mathbb{R}^{K^2 \times K^2}\), are defined as the block diagonal matrices \([V_{k,t}]_{k \in [K]} \cdot [\hat{V}_{k,t}]_{k \in [K]}\) such that

\[
V_{k,t} = ((e_{k,t-1})^\top) \quad \hat{V}_{k,t} = -((h_{k,t})^\top)
\]

(79)

using the notation from Eq. (28). Now define the following sequence, initialized with

\[
u^0, h^{-1} = 0, \hat{V}_0
\]

(80)

such that \( \lim_{d \to \infty} \frac{1}{d} \left\| e_0(u^0)^\top e_0(u^0) \right\|_F < +\infty \) and \( \hat{V}_0 \in S_+^{n \times n} \)

and recursively define

\[
u_t^{t+1} = Z^\top h_t(v^t) - e_t(u^t)(h^t)^\top
\]

(81)

\[
v^t = Ze_t(u^t) - h_{t-1}(v^{t-1})(e^t_{t-1})^\top
\]

(82)

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. (23).

State evolution equations The results from section A.3 show that the functions \( e^t, h^t \) are proximal 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.\((81)\), for any fixed \(m, \tilde{m}, b\). 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 \(t \in \mathbb{N}^*\):
\[
\begin{align*}
\phi_{1,n}(u_1^t, \ldots, u_K^t) & \sim \mathbb{E} \left[ \phi_{1,n}(H_1(\tilde{Q}_1, t)^{1/2}, \ldots, H_K(\tilde{Q}_K, t)^{1/2}) \right] \quad (83) \\
\phi_{2,n}(v_1^t, \ldots, v_{K}^t) & \sim \mathbb{E} \left[ \phi_{1,n}(G_1(Q_1, t)^{1/2}, \ldots, G_K(Q_K, t)^{1/2}) \right] \\
\end{align*}
\]
where the matrices \(H_k \in \mathbb{R}^{d \times K}, G_k \in \mathbb{R}^{n \times K}\) are independent with i.i.d. standard normal elements, and at each time step \(t \geq 1\)
\[
Q_{k,t} = \lim_{d \to +\infty} \frac{1}{d} \mathbb{E} \left[ \mathcal{E}_{k,t} \left( \{ H_k(\tilde{Q}_{k,t}^{1/2}) (\tilde{V}_{k,t})^{-1} \}_{k \in [K]} \right) \right] \\
\end{equation}
\]
\end{align*}
\]}
\[
\begin{align*}
&Q_{k,t} = \lim_{d \to +\infty} \frac{1}{d} \mathbb{E} \left[ \mathcal{E}_{k,t} \left( \{ H_k(\tilde{Q}_{k,t}^{1/2}) (\tilde{V}_{k,t})^{-1} \}_{k \in [K]} \right) \right] \\
&V_{k,t} = \lim_{d \to +\infty} \frac{1}{d} \sum_{i=1}^{d} \partial \mathcal{E}_{k,t} \left( \{ H_k(\tilde{Q}_{k,t}^{1/2}) (\tilde{V}_{k,t})^{-1} \}_{k \in [K]} \right) \\
&\end{align*}
\]
where the sequence is initialized with \(\tilde{V}_0, e_0, Q_{0,0} = \lim_{d \to +\infty} \frac{1}{d} \| e_0 (u_0^0)^{-} e_0 (u_0^0) \|_F\).

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

Note that in Lemma 5, we have directly written the block decomposition of the state evolution corresponding to the iteration Eq.\((81)\), which involves the block diagonal matrices \(Q, Q, V, V\) which are all in \(\mathbb{R}^{K^2 \times K^2}\). Using the notations introduced in section A.1
\[
V = [V_k]_{k=1}^{K} \quad \hat{V} = [\hat{V}_k]_{k=1}^{K} \quad Q = [Q_k]_{k=1}^{K} \quad \hat{Q} = [\hat{Q}_k]_{k=1}^{K} 
\]
Also note that we do not use the full state evolution giving the correlations across all time steps, but only 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.\((47)\).

Lemma 6. Consider any fixed point \(V, \hat{V}, Q, \hat{Q}\) of the state evolution equations from Lemma 5. For any fixed point \(u^*, \nu^*\) of iteration Eq.\((81)\), the quantity
\[
R_{V, \hat{V}}(u^*, \nu^*) = \left( 1 + \partial R(\nu) \hat{V}^{-1} \right) \left( u^* \hat{V}^{-1} \right) 
\]
is an optimal solution \(\hat{W}^*\) of problem Eq.\((47)\). Furthermore
\[
R_{\hat{V}, V}(\nu^*) = \left( 1 + \partial \hat{L}(\nu) V \right) \nu^* = Z \hat{W}^* 
\]
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.\((47)\). Note that the resolvents/proximalsm linking the fixed point of the AMP iteration with the solutions of Eq.\((47)\) 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.\((47)\). This is the purpose of the following lemma.
Lemma 7. Consider iteration Eq.(81), 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 $V_0 = \hat{V}$ and $u^0$ such that

$$\lim_{t \to \infty} \frac{1}{d} e_0(u^0)^\top e_0(u^0) = Q$$

the following holds

$$\lim_{t \to \infty} \lim_{d \to \infty} \frac{1}{\sqrt{d}} \|u^t - u^*\|_F = 0 \quad \lim_{t \to \infty} \lim_{d \to \infty} \frac{1}{\sqrt{d}} \|u^t - v^*\|_F = 0$$

Proof. The proof of Lemma 7 is deferred to subsection A.7.

Note that the $G$ defined here is not the same as the $G$ in the replica computation. Combining the lemmas 5, 6 and 7 with the pseudo-Lipschitz property, we have reached the following lemma

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

$$\phi_{1,n}(W^*) \sim \mathbb{E} \left[ \phi_{1,n} \left( R_{\hat{V}} \mathbb{V} (HQ^{1/2} \hat{V}^{-1}) \right) \right]$$

$$\phi_{2,n}(Z W^*) \sim \mathbb{E} \left[ \phi_{2,n} \left( R_{\hat{V}} \mathbb{V} (GQ^{1/2}) \right) \right]$$

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

Proof. Lemma 8 is a consequence of Lemmas 5, 6, 7 and applying the pseudo-Lipschitz property along with the fact that the iterates of the AMP have bounded norm using the state evolution and that the estimator also has bounded norm (feasibility assumption). Note that, for a generically non-strictly convex problem, being close to the zero gradient condition does not guarantee being close to the estimator. This is further discussed in Appendix A.5.

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 $m, \hat{m}, b$.

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

$$\inf_{m, \hat{m}} \sup_{W, b} L(Y_k m + Z_k \hat{W}_k) + r \left( \Sigma^+ \hat{W} \right)$$

$$+ \text{tr} \left( \hat{m}^\top \left( m - \frac{1}{\sqrt{d}} M \Sigma^+ \hat{W} \right) \right)$$

where we can use strong duality to write the equivalent form

$$\inf_{m, \hat{m}} \sup_{W} L(Y_k m + Z_k W_k^*) + r \left( \Sigma^+ \hat{W} \right)$$

$$+ \text{tr} \left( \hat{m}^\top \left( m - \frac{1}{\sqrt{d}} M \Sigma^+ \hat{W} \right) \right)$$

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

$$\partial \hat{m} = \hat{m} - \frac{1}{\sqrt{d}} M \Sigma^+ \hat{W}$$

$$\partial m = \hat{m} + \partial_m L(Y m + Z \hat{W})$$

\[\text{(98)}\]

\[\text{(99)}\]
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 $\partial_m$, 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 $1/d$

$$
\frac{1}{d} L(Y m + Z W^*) \overset{p}{\longrightarrow} d \mathbb{E} \left[ L(Y m + R_{L,V}(GQ^{1/2})) \right] \tag{100}
$$

Using the block decomposition from Eq.(53), the blocks $(R_{L,V}(GQ^{1/2}))_k \in \mathbb{R}^{n_k \times K}$ are given by:

$$
\arg\min_{T \in \mathbb{R}^{n_k \times K}} \left\{ L_k(T) + \frac{1}{2} \text{tr} \left( (T - (Y_k m + G_k Q_k^{1/2})) V_k^{-1} (T - (Y_k m + G_k Q_k^{1/2}))^\top \right) \right\} - Y_k m \tag{101}
$$

Using a block diagonal representation, we can write:

$$
\frac{1}{d} L(Y m + R_{L,V}(GQ^{1/2})) = \frac{1}{d} L(R_{L,V}(Y m + GQ^{1/2}))
= \frac{1}{d} M_{L,V}(Y m + GQ^{1/2}) - \frac{1}{2d} \text{tr} \left( (R_{L,V}(Y m + GQ^{1/2}) - (Y m + GQ^{1/2})) V^{-1} (R_{L,V}(Y m + GQ^{1/2}) - (Y m + GQ^{1/2}))^\top \right) \tag{102}
$$

where we have introduced the Bregman-envelope [65] with respect to the distance Eq. (17)

$$
\mathcal{M}_{L,V}(Y m + GQ^{1/2}) = \min_T \left\{ L(T) + \frac{1}{2} \text{tr} \left( (T - (Y m + GQ^{1/2})) V^{-1} (T - (Y m + GQ^{1/2}))^\top \right) \right\} \tag{103}
$$

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

$$
\frac{1}{d} L(Y m + R_{L,V}(GQ^{1/2})) = \frac{1}{d} M_{L,V}(Y m + GQ^{1/2}) - \frac{1}{2} \text{tr}(V^\top Q) \tag{104}
$$

Taking the gradient w.r.t. $m$ using the expression for the derivative of a Bregman envelope [65], we get:

$$
\partial_m L(Y m + R_{L,V}(GQ^{1/2})) = \frac{1}{d} Y^\top \left( Y m + GQ^{1/2} - R_{L,V}(Y m + GQ^{1/2}) \right) V^{-1} \tag{105}
$$

which prescribes, using Lemma 8

$$
\dot{m} \doteq \frac{1}{d} Y^\top \left( R_{L,V}(Y m + GQ^{1/2}) - Y m + GQ^{1/2} \right) V^{-1} \tag{106}
$$

For $m$, we use the block decomposition from Eq.(67), which simplifies the pseudo-inverse $\Sigma^+$ in Eq. (98) to give, using Lemma 8 again

$$
m \doteq \frac{1}{\sqrt{d}} \text{M} \eta(HQ^{1/2} V^{-1}) \tag{107}
$$

where the function $\eta$ acts on the block diagonal and is defined by Eq.(68). Using those results and the definition of $W$, the solution $W^*$ and the quantity $X W^*$ are characterized, in the pseudo-Lipschitz sense of Theorem 1, by the fixed point of the system of equations (the first four equations are meant
for all $1 \leq k \leq K$:

$$Q_k = \lim_{d \to +\infty} \frac{1}{d} \mathbb{E} \left[ e_k(\{H_k(Q_k)^{1/2}v_k\})^{\top} e_k(\{H_k(Q_k)^{1/2}v_k\}) \right] \in \mathbb{R}^{K \times K} \quad (108)$$

$$\dot{Q}_k = \lim_{d \to +\infty} \frac{1}{d} \mathbb{E} \left[ h_k(G_k^{1/2})^{\top} h_k(G_k^{1/2}) \right] \in \mathbb{R}^{K \times K} \quad (109)$$

$$V_k = \lim_{d \to +\infty} \frac{1}{d} \sum_{i=1}^{d} \mathbb{E} \left[ \frac{\partial e_k(\{H_k(Q_k)^{1/2}\})}{\partial (H_k(Q_k)^{1/2})} \right] \in \mathbb{R}^{K \times K} \quad (110)$$

$$\dot{V}_k = -\lim_{d \to +\infty} \frac{1}{d} \sum_{i=1}^{m_k} \mathbb{E} \left[ \frac{\partial h_k(G_k^{1/2})}{\partial (G_k^{1/2})} \right] \in \mathbb{R}^{K \times K} \quad (111)$$

$$m = \frac{1}{\sqrt{d}} \mathbb{E} \left[ M\eta(HQ^{1/2}) \right] \in \mathbb{R}^{K \times K} \quad (112)$$

$$\dot{m} = \frac{1}{\sqrt{d}} \mathbb{E} \left[ M\eta(HQ^{1/2}) \right] \in \mathbb{R}^{K \times K} \quad (113)$$

Using the explicit form of the different functions given in section A.3 and Stein’s lemma for the derivatives, these equations match those of Theorem 1. This completes the proof.

A.5 On the strict convexity assumption

If the optimization problem defining $W^*$ is strictly convex, there is only one minimizer and the provided proof is enough. Additionally, it is shown in [68] that for any loss function that is strictly convex in its argument and penalized with the $\ell_1$ norm, provided the data is sampled from a continuous distribution, the solution is unique with probability one regardless of the rank of the design matrix. Thus finding a point verifying the optimality condition of (47) is also enough to complete the proof. For generic convex (non-strictly) problems a more careful analysis could be performed in the same spirit as the one of [31]. Empirically the result still holds.

A.6 On the uniqueness of the solution to the fixed point equations (108)

It is possible to reconstruct Bregman envelopes on problem (47) for the loss and regularization as we have done for the loss in the previous section. We can then show that the fixed point equations (108) 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], this problem should be asymptotically strictly convex. This is supported by the simulations presented in the experiments sections but left as an assumption in the main paper.

A.7 Proof of Lemma 7

This proof follows a similar argument to the one used to control the trajectory of the AMP studied in [50]. Note that, because of the way the AMP is initialized using the fixed point of the state evolution equations, for any $t \geq 1$ the following holds:

$$\lim_{d \to +\infty} \frac{1}{d} \mathbb{E} \left[ e(u')^{\top} e(u') \right] = Q \in \mathbb{R}^{K \times K} \quad (114)$$

$$\lim_{d \to +\infty} \frac{1}{d} \mathbb{E} \left[ h(v')^{\top} h(v') \right] = Q \in \mathbb{R}^{K \times K} \quad (115)$$

where

$$e(u') = (Id + \partial \bar{r}(\bullet)\dot{V})^{-1}(u'\dot{V}) \quad h(v') = \left( (Id + \partial \bar{L}(\bullet)\dot{V})^{-1}(v') - v' \right) \dot{V}^{-1} \quad (116)$$

then the limit we are looking for reads:

$$\lim_{d \to +\infty} \frac{1}{d} \|u' - u'^{-1}\|^2_F = \lim_{d \to +\infty} 2(\dot{Q} - \frac{1}{d} \text{tr}(u')^{\top} u'^{-1}) \quad (117)$$
We thus need to study the correlation between successive iterates. At each time step, denote \(( \hat{C}_t, C_t)\) in \(\mathbb{R}^{K \times K}\) the correlation matrices between iterates at times \(t, t-1\) describing the Gaussian fields respectively associated to \(u^t, v^t\) i.e.,
\[
\lim_{d \to \infty} \frac{1}{d} \text{tr}(u^t \cdot u^{t-1}) = \hat{C}_t \quad \lim_{d \to \infty} \frac{1}{d} \text{tr}(v^t \cdot v^{t-1}) = C_t
\]
we can then write the block diagonal Gaussian form of \(\hat{C}_t\) as:
\[
\hat{C}_t = \begin{pmatrix} \hat{C}_t^{(1)} & 0 \\ 0 & \hat{C}_t^{(2)} \end{pmatrix}
\]
and similarly for \(C_t\):
\[
C_t = \begin{pmatrix} C_t^{(1)} & 0 \\ 0 & C_t^{(2)} \end{pmatrix}
\]
where the matrices \(\hat{C}_t, \hat{C}_t^2, \hat{C}_t'^2\) are in \(\mathbb{R}^{K \times K}\), \(G, G', G''\) are in \(\mathbb{R}^{n \times K}\) and all have i.i.d. standard normal elements. The recursion describing the evolution of these correlations then reads:
\[
\hat{C}_{t+1} = \frac{1}{2} \mathbb{E} \left[ e(\hat{H}^{1/2} \hat{C}_t^{1/2} + \hat{H}'(\hat{Q} - \hat{C}_t)) e(\hat{H}^{1/2} \hat{C}_t^{1/2} + \hat{H}'(\hat{Q} - \hat{C}_t)) \right]
\]
\[
\hat{C}_t = \frac{1}{2} \mathbb{E} \left[ h(G^{1/2} \hat{C}_t^{1/2} + G'(C_t - \hat{C}_t)) h(G^{1/2} \hat{C}_t^{1/2} + G'(C_t - \hat{C}_t)) \right]
\]
Integrating out the independent \(\hat{H}, \hat{H}'\) first, we get:
\[
C_{t+1} = \int_{\mathbb{R}^d} d\mu(H) I(H)^{-1} I(H)
\]
where \(I(H) = \int_{\mathbb{R}^d} d\mu(H') e(\hat{H}^{1/2} \hat{C}_t^{1/2} + \hat{H}'(\hat{Q} - \hat{C}_t)) \). So \(C_t\) is symmetric positive definite, assuming the resolvents aren’t trivial. The same argument applied to \(C_t\) shows it is also symmetric positive definite. From [64], the operators
\[
(Id + \partial \mathcal{F}(\bullet) \hat{V}^{-1})^{-1}(\bullet) \quad (Id + \partial \mathcal{L}(\bullet) \hat{V})^{-1}(\bullet)
\]
are \(D\)-firm w.r.t. the Bregman distances induced by the differentiable, strictly convex functions \(\frac{1}{2} \text{tr}(XVX^\top)\) and \(\frac{1}{2} \text{tr}(XV^{-1}X^\top)\) respectively. Recall
\[
e(u^t) = (Id + \partial \mathcal{F}(\bullet) \hat{V}^{-1})^{-1}(u^t \hat{V}^{-1}) \quad h(u^t) = \left( (Id + \partial \mathcal{L}(\bullet) \hat{V})^{-1}(u^t) - u^t \right) \hat{V}^{-1}
\]
Then, using the definition of \(D\)-firm
\[
\langle e(\hat{Z}^t) - e(\hat{Z}^{t-1}), (e(\hat{Z}^t) - e(\hat{Z}^{t-1})) \hat{V} \rangle \leq \langle e(\hat{Z}^t) - e(\hat{Z}^{t-1}), (\hat{Z}^t - \hat{Z}^{t-1}) \hat{V}^{-1} \hat{V} \rangle
\]
then, adding the normalization by \(\frac{1}{2}\), using the representation Eq.(119-122), taking expectations and applying the matrix form of Stein’s lemma, see for example [87] Lemma 12, we get:
\[
\text{tr}((Q - C_{t+1}) \hat{V}) \leq \text{tr}((Q - \hat{C}_t) \hat{V})
\]
Using a similar argument on \(h\), we get
\[
\text{tr}((Q - \hat{C}_t) \hat{V}) \leq \text{tr}((Q - C_t) \hat{V})
\]
and
\[
\text{tr}(C_{t+1} \hat{V}) \geq \text{tr}(C_t \hat{V})
\]
thus the sequence \(\text{tr}(C_{t+1} \hat{V})\) is a bounded (above) monotone (increasing) sequence, and therefore converges. Since \(\hat{V}\) is positive definite and given the iteration defining \(C_{t+1}\) from \(C_t\), any fixed point of this iteration is a fixed point of \(\text{tr}(C_t \hat{V})\). Assuming there is only one fixed point to the set of self-consistent equations Eq.(8) (see previous section), the proof is complete. (A similar argument can be carried out on \(\hat{C}_t\).)
B Replica computation

B.1 Setting of the problem

In this Section we give a full derivation of the results 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 [69]. In the general computation, we will consider the classification problem of \( K \) clusters, assuming a dataset \( \{ (x^{\nu}, y^{\nu}) \}_{\nu \in [n]} \) of \( n \) independent datapoints where, as in the main text, the labels \( y \) take values in a set of \( K \) elements, \( y^{\nu} \in \{ e_k \}_k \), with \( e_k \in \mathbb{R}^L \). The elements of the dataset are independently generated by a mixture density in the form

\[
P(x, y) = \sum_{k=1}^{K} \mathbb{I}(y = e_k) \rho_k \mathcal{N}(x | \mu_k, \Sigma_k), \quad \sum_{k=1}^{K} \rho_k = 1. \tag{132}
\]

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

\[
\mathcal{R}(W, b) \equiv \sum_{\nu=1}^{n} \ell \left( y^{\nu}, \frac{Wx^{\nu}}{\sqrt{d}} + b \right) + \lambda r(W), \tag{133}
\]

i.e., they are given by

\[
(W^*, b^*) = \arg\min_{W \in \mathbb{R}^{L \times d}, b \in \mathbb{R}^L} \mathcal{R}(W, b). \tag{134}
\]

We will say that \( W \in \mathbb{R}^{L \times d} \) and \( b \in \mathbb{R}^L \) are the weights and bias to be learned respectively, \( \ell \) 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 \( \varphi : \mathbb{R}^L \rightarrow \{ e_k \}_k \) is given, such that, once \((W^*, b^*)\) are obtained, a new point \( x \) is assigned to the label

\[
x \mapsto \varphi \left( \frac{W^* x}{\sqrt{d}} + b^* \right) \in \{ e_k \}_k. \tag{135}
\]

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 \( L \times 1 \) matrices. A typical “high-dimensional encoding” is the one-hot encoding convention adopted in Theorem 1, where \( L = K \) and \( \{ e_k \}_k \subset \mathbb{R}^K \) is the canonical basis of \( \mathbb{R}^K \). In this case, the adopted classifier is

\[
\varphi(x) \equiv \hat{y}(x), \quad \hat{y}_k(x) = \mathbb{I}(\max_{\kappa} x_\kappa = x_k). \tag{136}
\]

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 \( \varphi(x) = \text{sign}(x) \), 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 \((W, b)\), namely

\[
\mu_\beta(W, b) \propto e^{-\beta \mathcal{R}(W, b)} = e^{-\beta r(W)} \prod_{\nu=1}^{n} \left[ \exp \left[ -\beta \ell \left( y^{\nu}, \frac{Wx^{\nu}}{\sqrt{d}} + b \right) \right] \right]. \tag{137}
\]

The parameter \( \beta > 0 \) is introduced for convenience: in the \( \beta \to +\infty \) limit, the Gibbs measure concentrates on the values \((W^*, b^*)\) which minimize the empirical risk \( \mathcal{R}(W, b) \) and are therefore the goal of the learning process. The functions \( P_\nu \) and \( P_\nu \) 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.,

\[
f_\beta = -\lim_{n,d \to +\infty} \mathbb{E}_{(x, y)} \left[ \frac{\ln Z_\beta}{d/\beta} \right]. \tag{138}
\]
where $E_{(x,y)}[\bullet]$ is the average over the training dataset, and we have introduced the partition function

$$Z_\beta \equiv \int e^{-\beta R(W,b)}dW$$

(139)

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

$$-\lim_{n,d\to+\infty} \frac{\ln Z_\beta}{\beta} = \lim_{n,d\to+\infty} \frac{1 - E_{(x,y)}[Z^\beta_\beta]}{s\beta}.$$  

(140)

### B.3 Replica approach

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

$$E_{(x,y)}[Z^\beta_\beta] = \prod_{a=1}^n \int dW^a P_y(W^a) \left( \sum_k \rho_k \mathbb{E}_{x|y=e_k} \left[ \prod_{a=1}^s P_y(e_k | W^a x + b) \right] \right)^n.$$  

(141)

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

$$\mathbb{E}_{x|y=e_k} \left[ \prod_{a=1}^s P_y(e_k | \eta^a) \right] = \prod_{a=1}^s \int d\eta^a P_y(e_k | \eta^a) \mathcal{N} \left( \eta^a \frac{W^a \mu_k}{\sqrt{d}} + b; \frac{W^a \Sigma_k W^b_T}{d} \right).$$  

(142)

We can write then

$$E_{(x,y)}[Z^\beta_\beta] =$$

$$\prod_{a=1}^n \int dW^a P_y(W^a) \left( \sum_k \frac{\rho_k}{n} \prod_{a=1}^s \int d\eta^a P_y(e_k | \eta^a) \mathcal{N} \left( \eta^a \frac{W^a \mu_k}{\sqrt{d}} + b; \frac{W^a \Sigma_k W^b_T}{d} \right) \right)^n$$

$$\left( \prod_{k=1}^K \prod_{a \leq b} \left( \frac{dQ^a_k d\eta^a}{(2\pi)^{L^2/2}} \right) \right)^n \int \prod_{k} \prod_{a} \frac{dm^a_k}{(2\pi)^{L^2/2}} e^{-d\Phi^{(s)}}.$$  

(143)

where we introduced the order parameters

$$Q^a_k = \frac{W^a \Sigma_k W^b_T}{d} \in \mathbb{R}^{L \times L}, \quad a, b = 1, \ldots, s,$$  

(144)

$$m^a_k = \frac{W^a \mu_k}{\sqrt{d}} \in \mathbb{R}^L, \quad a = 1, \ldots, s,$$  

(145)

and the replicated free-energy

$$\beta \Phi^{(s)}(Q, m, Q, m, b) = \sum_{k=1}^K \sum_{a} \sum_{a \leq b} m^a_k Q^a_k + \frac{1}{d} \prod_{k=1}^K \prod_{a \leq b} \text{tr} \left[ Q^a_k Q^a_k \right]$$

$$- \frac{1}{d} \ln \prod_{a=1}^s P_y(W^a) \prod_{k=1}^K \prod_{a \leq b} e^{\text{tr} \left[ Q^a_k W^a \Sigma_k W^b_T \right]} \prod_{a} e^{\text{tr} \left[ \sqrt{d} m^a_k W^a \mu_k \right]}$$

$$- \alpha \ln \sum_{k} \rho_k \prod_{a=1}^s \int d\eta^a P_y(e_k | \eta^a) \mathcal{N} \left( \eta^a m^a_k + b; Q^a_k \right).$$  

(146)

At this point, the free energy $f_\beta$ should be computed extremising with respect to all the order parameters by virtue of the Laplace approximation (in addition to $b$),

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

(147)

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

\[
Q_k^a = \begin{cases} 
R_k & a = b \\
Q_k & a \neq b
\end{cases} \\
\tilde{Q}_k^a = \begin{cases} 
-\frac{1}{r}R_k & a = b \\
\tilde{Q} & a \neq b
\end{cases} \\
m_k^a = m_k \\
\tilde{m}_k^a = \tilde{m}_k \\
(148)
\]

This ansatz is justified by the fact that we are assuming $\ell$ 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 hypothesis, we can write

\[
Q_k^{ab} \mapsto Q_k \equiv I_{s,k} \otimes (R_k - Q_k) + 1_s \otimes Q_k.
\]

The inverse matrix is therefore

\[
Q_k^{-1} = I_s \otimes (R_k - Q_k)^{-1} - I_s \otimes [(R_k + (s-1)Q_k)^{-1}Q_k(R_k - Q_k)^{-1}],
\]

whereas

\[
det Q_k = det(R_k - Q_k)^{s-1}det(R_k + (s-1)Q_k) \\
= 1 + s \ln det(R_k - Q_k) + s \text{tr} [(R_k - Q_k)^{-1}Q_k] + o(s).
\]

If we denote $V_k \equiv R_k - Q_k$

\[
\ln \sum_k \rho_k \prod_{a=1}^s \int d\eta^a P_\eta(e_k|\eta^a)N(\eta|m_k^a + b, Q_k^{ab}) \\
= s \sum_k \rho_k E_{\xi} \ln \left( \int \frac{d\eta P_\eta(e_k|\eta)}{\sqrt{\det(2\pi V_k)}} e^{-\frac{1}{2}(\eta-m_k+b-Q_k^{1/2}\xi)^T V_k^{-1}(\eta-m_k+b-Q_k^{1/2}\xi)} \right) + o(s) \\
= s \sum_k \rho_k E_{\xi} \left[ \ln Z \left( e_k, m_k + b + Q_k^{1/2}\xi, V_k \right) \right] + o(s),
\]

with $\xi \sim N(0, I_e)$ is a normally distributed vector and we have introduced the function

\[
Z(e_k, m, V) \equiv \int \frac{d\eta P_\eta(e_k|\eta)}{\sqrt{\det(2\pi V)}} e^{-\frac{1}{2}(\eta-m)^T V^{-1}(\eta-m)}
\]

On the other hand, denoting by $\tilde{V}_k = R_k + Q_k$

\[
\frac{1}{d} \ln \prod_{a=1}^s \left( \int P_w(W^a)dW^a \right) e^{-\frac{1}{2}tr(V_k^a W^a \Sigma_k(W^a)^\top) + \sqrt{\det m_k^a} \int W^a \mu_k \prod_k e^{\frac{1}{2}tr(Q_k W^a \Sigma_k(W^a)^\top)}} = \\
= \frac{s}{d} E_{\xi} \ln \left[ \int P_w(W)dW \prod_k \exp \left( -\frac{tr(V_k^a W \Sigma_k W^\top)}{2} + \sqrt{\det m_k^a} W \mu_k + \xi \otimes \sqrt{\tilde{Q}_k \otimes \Sigma_k \otimes \Sigma_k} \right) \right] + o(s).
\]

In the expression above we have used the tensorial product $\tilde{Q} \otimes \Sigma = (\tilde{Q}_{kk''} \Sigma_{ij})_{kk''ij}$. Given a matrix $B \in \mathbb{R}^{L \times d}$ and the tensors $A, A' \in \mathbb{R}^{L \times d} \otimes \mathbb{R}^{L \times d}$, we denote $(B \otimes A)_{kk''} \equiv \sum_{k''} B_{kk''} A_{k''k} \in \mathbb{R}^{L \times d}$ and $(A \otimes A')_{kk''} \equiv \sum_{kk''} A_{kk''} A_{k''k'} \in \mathbb{R}^{L \times d}$. In this way, we define $\sqrt{A}$ to be the tensor such that $A = \sqrt{A} \otimes \sqrt{A}$. 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 $E_{\Xi}[\bullet]$. Therefore, the replicated replica symmetric free-energy is given by

\[
\lim_{s \to 0} \frac{\beta}{s} \Phi^{(s)}_{RS} = \sum_{k=1}^K \frac{1}{2} \text{tr} \left[ \tilde{m}_k^a Q_k \right] - \frac{1}{2} \sum_{k=1}^K \text{tr} \left[ \tilde{Q}_k^a V_k \right] - \frac{1}{2} \sum_{k=1}^K \text{tr} \left[ V_k^a \tilde{V}_k \right] \\
- \alpha \beta \Psi_{\text{out}}(m, Q, V) - \beta \Psi_w(m, Q, V)
\]

(155)
where we have defined two contributions

\[ \Psi_{\text{out}}(m, Q, V) = \beta^{-1} \sum_k \rho_k \mathbb{E}_k \ln Z(e_k, \omega_k, V_k) \]  

(156)

\[ \Psi_w(m, \hat{Q}, \hat{V}) = \frac{1}{\beta d} \mathbb{E}_k \ln \left( \int P_u(W) dW \prod_k e^{-\frac{1}{2} \left| \left| \frac{1}{\sqrt{2m_k}} W \mu_k + \Sigma_k \otimes \sqrt{Q_k} \otimes \xi_k \otimes W \right| \right|_2^2} \right) \]  

(157)

and introduced, for future convenience,

\[ \omega_k \equiv m_k + b + Q_k^{1/2} \xi_k. \]  

(158)

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

\[ \Psi_{\text{out}}(m, Q, V) = -\sum_k \rho_k \mathbb{E}_k \left[ M_{f(e_k, V_k^{1/2})} \left( V_k^{-1/2} \omega_k \right) \right]. \]  

(159)

Here and in the following the quantity

\[ M_f(s)(u) \equiv \min_{v \in \text{domain}(v)} \left[ \frac{1}{2} \| v - u \|_F^2 + f(v) \right] \]  

(160)

is the Moreau envelope of \( f : \text{domain}(v) \to \mathbb{R} \), whereas \( \| \bullet \|_F \) is the Frobenius norm. We can write the contribution \( \Psi_{\text{out}} \) in terms of a proximal

\[ h_k = V_k^{1/2} \text{Prox}_{f(e_k, V_k^{1/2})}(V_k^{-1/2} \omega_k) \equiv V_k^{1/2} \arg \min_u \left[ \frac{1}{2} \| u - V_k^{-1/2} \omega_k \|_F^2 + f(e_k, V_k^{1/2} u) \right]. \]  

(161)

as

\[ \Psi_{\text{out}}(m, Q, V) = -\sum_k \rho_k \mathbb{E}_k \left[ \frac{1}{2} \| V_k^{-1/2} h_k - V_k^{-1/2} \omega_k \|_F^2 + f(e_k, h_k) \right]. \]  

(162)

A similar expression can be obtained for \( \Psi_w \). Defining

\[ A = \left( \sum_k \hat{V}_k \otimes \Sigma_k \right)^{-1}, \quad B = \sqrt{d} \sum_k \mu_k \hat{m}_k^\top + \sum_k \Sigma_k \otimes \sqrt{Q_k} \otimes \Sigma_k. \]  

(163)

\( \Psi_w \) can be written as

\[ \Psi_w(m, \hat{Q}, \hat{V}) = \frac{1}{2d} \mathbb{E}_k [B \otimes A \otimes B] \]

\[ + \frac{1}{\beta d} \mathbb{E}_k \ln \left[ \int dW \exp \left( -\frac{\beta}{2} \| A^{-1/2} \otimes W - A^{-1/2} \otimes B \|_F^2 - \beta r(W) \right) \right]. \]  

(164)

It follows that, for \( \beta \to +\infty \),

\[ \Psi_w(m, \hat{Q}, \hat{V}) = \frac{1}{2d} \mathbb{E}_k [B \otimes A \otimes B] - \frac{1}{d} \mathbb{E}_k \left[ M_{r(A^{1/2} \otimes B)}(A^{1/2} \otimes B) \right]. \]  

(165)

As before, let us introduce the proximal

\[ G = A^{1/2} \otimes \text{Prox}_{r(A^{1/2} \otimes B)}(A^{1/2} \otimes B) \in \mathbb{R}^{L \times d} \]  

(166)

We can rewrite the prior contribution \( \Psi_w \) as

\[ \Psi_w(m, \hat{Q}, \hat{V}) = \frac{1}{2d} \mathbb{E}_k [B \otimes A \otimes B] - \frac{1}{d} \mathbb{E}_k \left[ \| A^{-1/2} \otimes G - A^{1/2} \otimes B \|_F^2 \right] + r(G). \]  

(167)
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,

$$f_{RS} = \text{Extr}_{m, Q, \nu, \beta} \left[ \sum_{k=1}^{K} m_k^T m_k + \frac{1}{2} \sum_{k=1}^{K} \text{tr} \left[ V_k^T Q_k \right] - \frac{1}{2} \sum_{k=1}^{K} \text{tr} \left[ Q_k^T V_k \right] - \frac{1}{2} \sum_{k=1}^{K} \text{tr} \left[ V_k^T V_k \right] - \alpha \Psi_{\text{out}}(m, Q, V) - \Psi_{\nu}(\bar{m}, \bar{Q}, \bar{V}) \right].$$ (168)

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 $\Psi_{\text{out}}$ as\(^1\)

\begin{align*}
\hat{Q}_k &= \alpha \rho_k \mathbb{E}_{\xi} \left[ f_k f_k^T \right], \quad (169a) \\
\hat{V}_k &= -\alpha \rho_k Q_k^{-1/2} \mathbb{E}_{\xi} \left[ f_k \xi \right], \quad (169b) \\
\hat{m}_k &= \alpha \rho_k \mathbb{E}_{\xi} \left[ f_k \right], \quad (169c) \\
b &= \sum_k \rho_k \mathbb{E}_{\xi} \left[ h_k - m_k \right] \iff \sum_k \rho_k \mathbb{E}_{\xi} \left[ V_k f_k \right] = 0. \quad (169d)
\end{align*}

where for brevity we have denoted

$$f_k \equiv V_k^{-1}(h_k - \omega_k).$$ (170)

Similarly, the saddle-point equations from $\Psi_{\text{out}}$ are

\begin{align*}
V_k &= \frac{1}{d} \mathbb{E}_{\xi} \left[ G \left( \hat{Q}_k \otimes \Sigma_k \right)^{-1/2} \otimes (I_k \otimes \Sigma_k) \right] \Xi_k^T, \quad (171a) \\
Q_k &= \frac{1}{d} \mathbb{E}_{\xi} \left[ G \Sigma_k G^T \right] \quad (171b) \\
m_k &= \frac{1}{\sqrt{d}} \mathbb{E}_{\xi} \left[ G \mu_k \right]. \quad (171c)
\end{align*}

To obtain the replica symmetric free energy, therefore, the given set of equation has to be solved, and the result then plugged in Eq. (168). 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

$$\epsilon_\ell = \frac{1}{n} \sum_{\ell=1}^{n} \ell \left( y+^\ell \frac{W^* \omega_{\ell}^\ast}{\sqrt{d}} + b^* \right),$$ (172)

in the $n \to +\infty$ limit. Denoting for brevity $\ell_k(x) \equiv \ell(e_k, x)$, the best way to proceed is to observe that $\mathbb{E}_{\nu(\omega_{\ell}^\ast)}[\mathcal{R}(W^*, b^*')] = -\lim_{\beta \to -\infty} \mathbb{E}_{\nu(\omega_{\ell}^\ast)}[\partial_\beta \ln Z_\beta] = \lambda \mathbb{E}_{\nu(\omega_{\ell}^\ast)}[\mathcal{R}(W^*, b^*')] + \epsilon_\ell,$

where

$$\epsilon_\ell = -\lim_{\beta \to +\infty} \partial_\beta (\beta \Psi_{\text{out}}) = \lim_{\beta \to +\infty} \sum_k \rho_k \int \ell_k(\eta) \frac{e^{-\frac{1}{2}(\eta - m_k)^T \hat{V}_k^{-1}(\eta - m_k) - \beta \ell_k(\eta)}}{\det(2\pi \beta \hat{V}_k) Z(e_k, \omega_k^\ast, \beta^{-1} \hat{V}_k^\ast)} d\eta.$$

\(^1\)To obtain the equation for $V$ it is convenient to use Stein’s lemma, so that $\mathbb{E}[\partial_\ell f_k] = \mathbb{E}[f_k \xi^T]$. 

\[19\]
In the $\beta \to +\infty$ limit, the integral concentrates on the minimizer of the exponent, that is, by definition, the proximal $h_k$. In conclusion, $\epsilon_\ell = \sum_k \rho_k \mathbb{E}[\ell(h_k)]$. By means of the same concentration result, the training error is

$$\epsilon_g = \frac{1}{n} \sum_{\nu=1}^n \mathbb{I} \left( \varphi \left( \frac{W^* x^\nu}{\sqrt{d}} + b^* \right) \neq y^\nu \right) \xrightarrow{n \to +\infty} \sum_{k=1}^K \rho_k \mathbb{E}_x \left[ \mathbb{I}(\varphi(h_k) \neq e_k) \right].$$  (174)

The expressions above hold in general, but, as anticipated, important simplifications can occur in the set of saddle-point equations (169) and (171) depending on the choice of the loss $\ell$ and of the regularization function $\gamma$.

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

$$\epsilon_g = \mathbb{E}_{y^{\text{new}}, x^{\text{new}}} \left[ \mathbb{I} \left( \varphi \left( \frac{W^* x^{\text{new}}}{\sqrt{d}} + b^* \right) \neq y^{\text{new}} \right) \right].$$  (175)

This expression can be rewritten as

$$\epsilon_g = \sum_k \rho_k \int \mathbb{I}(\varphi(\eta) = e_k) \mathbb{E}_{x^{\text{new}}} \left[ \delta \left( \eta - \frac{W^* x^{\text{new}}}{\sqrt{d}} - b^* \right) \right] d\eta$$  (176)

Once again, we write

$$\mathbb{E}_{x^{\text{new}}} \left[ \delta \left( \eta - \frac{W^* x^{\text{new}}}{\sqrt{d}} - b^* \right) \right] \xrightarrow{d \to +\infty} N(\eta|m_k^* + b^*, Q_k^*)$$  (177)

so that

$$\epsilon_g = \sum_{k=1}^K \rho_k \mathbb{E}_x \left[ \mathbb{I}(\varphi \left( m_k^* + Q_k^* \xi + b^* \right) \neq e_k) \right].$$  (178)

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 $\epsilon_g$, $\epsilon_\ell$ and $\epsilon_\ell$ we first need to find the fixed-point solutions of the saddle-point equations (169) and (171). 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 (that we adopted to be $10^{-5}$). In the simplest setting, i.e., the one discussed in Corollary 3, the update of $(m_k, Q_k, V_k)_{k \in [K]}$ is performed explicitly using eq. (11), where $\mathbb{E}_{\sigma, \mu} [\bullet]$ is a shorthand for the sum over the eigenvalues and eigenvectors of the assigned covariance matrices. The update of $(m_k, Q_k, V_k)_{k \in [K]}$ (right hand side of eq. (8)) is more involved, as it requires the computation of the proximal followed by a Gaussian average. Such average has been performed using a Monte Carlo strategy, i.e., by solving the equation for the proximal for a large number ($10^4 - 10^5$) of instances of $\xi$ and averaging the solution. We remark that in the case of the square loss, the proximal can be computed analytically and the integration can be performed explicitly, highly simplifying the fixed-point equations (see below eq. (191)). We have found that in practice fluctuations due to the adopted Monte Carlo pool were small enough to be negligible compared with the outcomes of direct numerical experiments.

The convergence to the the correct fixed point is guaranteed (in principle) by the convexity of the problem. However, a few delicate aspects have to be taken into account in the update process described above.

1. The update requires the computation of the proximals $G$ and $h_k$. 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 strong convexity of the problem defining the proximal. In our study of the cross-entropy loss function, for example, we computed the proximals $h_k$ numerically solving Eq. (194). 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. (198). We solved this issue performing an annealing in $\lambda$, i.e., solving for the proximal for decreasing values of the regularization strength.
2. 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 $\hat{V}_k$ in (169), which requires the inversion of $Q_k$. 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 $f_k$ satisfies the equation $f_k + \partial_x \ell_k(V_k f_k + \omega_k) = 0$ so that

$$\partial \omega_k f_k = -(I_K + \partial^2_x \ell_k(V_k f_k + \omega_k)V_k)^{-1}\partial^2_x \ell_k(V_k f_k + \omega_k)V_k.$$  

Using Stein’s lemma,

$$\hat{V}_k = -\alpha \rho_k \mathbb{E}_\xi [\partial \xi f_k] = \alpha \rho_k \mathbb{E}_\xi \left[(I_K + \partial^2_x \ell_k(V_k f_k + \omega_k)V_k)^{-1}\partial^2_x \ell_k(V_k f_k + \omega_k)\right].$$  

We found this equation numerically more stable than the one given in (169) when dealing with the cross-entropy loss.

Our implementation can be found at [58].


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 \( \ell \) and of the regularisation function \( r \). From the analysis developed in the previous Appendices, it is clear that the choices of \( \ell \) and \( r \) impact separately the set of equations (169) and (171) 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 \( \ell_2 \) regularization

In this Section we consider the relevant case of quadratic regularization, \( r(W) = \frac{1}{2}\|W\|_F^2 \). In this case the computation of \( \Psi_w \) can be performed explicitly via a Gaussian integration,

\[
\frac{1}{\beta} \Psi_w(m, \hat{Q}, \hat{V}) = \frac{1}{2d} \text{tr} \ln S - \frac{K \ln \beta}{2\beta} + \frac{1}{2} \text{tr} \left[ S \circ \left( \sum_{kk'} \hat{m}_k \hat{m}_{k'}^\top \otimes \mu_k \mu_{k'}^\top + \frac{1}{d} \sum_k \hat{Q}_k \otimes \Sigma_k \right) \right].
\]

(180)

Here we have introduced, for notation compactness,

\[
S \equiv \left( \lambda I_K \otimes I_d + \sum_k \hat{V}_k \otimes \Sigma_k \right)^{-1}.
\]

(181)

This form of \( \Psi_w \) allows us to write in a simpler way the set of eqs. (171), that can be re-written as

\[
Q_k = \text{tr}_d \left[ (I_K \otimes \Sigma_k) \circ S \otimes \left( \sum_{kk'} \hat{m}_k \hat{m}_{k'}^\top \otimes \mu_k \mu_{k'}^\top + \frac{1}{d} \sum_k \hat{Q}_k \otimes \Sigma_k \right) \circ S \right] \]

\[
m_k = \sum_{k'} \text{tr}_d \left[ S \circ (\hat{m}_{k'} \otimes \mu_{k'} \mu_{k'}^\top) \right] \]

\[
V_k = \frac{1}{d} \text{tr}_d \left[ (I_K \otimes \Sigma_k) \circ S \right].
\]

(182)

In the previous equations, by \( \text{tr}_d \) we denoted the trace with respect to the components living in the \( d \)-dimensional space of the dataset.

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

\[
Q_k = \frac{1}{d} \sum_{i=1}^d \sigma_i^k \left( \lambda I_K + \sum_k \sigma_i^k \hat{V}_k \right)^{-1} \left( \sum_{kk'} \mu_i^k \mu_i^{k'} \hat{m}_k \hat{m}_{k'}^\top + \sum_k \sigma_i^k \hat{Q}_k \right) \left( \lambda I_K + \sum_k \sigma_i^k \hat{V}_k \right)^{-1}
\]

\[
m_k = \frac{1}{d} \sum_{i=1}^d \sum_{k'} \mu_i^k \mu_i^{k'} \left( \lambda I_K + \sum_k \sigma_i^k \hat{V}_k \right)^{-1} \hat{m}_{k'}
\]

\[
V_k = \frac{1}{d} \sum_{i=1}^d \sigma_i^k \left( \lambda I_K + \sum_k \sigma_i^k \hat{V}_k \right)^{-1}.
\]

(183)

Introducing the joint density

\[
\frac{1}{d} \sum_{i=1}^d \prod_{k=1}^K \delta(\sigma_i^k - \sigma_i^k) \delta(\mu_i^k - \mu_i^k) \xrightarrow{d \to +\infty} \rho(\sigma, \mu),
\]

(184)
then we can write the saddle-point equations given in Corollary 3
\[
Q_k = E_{\sigma, \mu} \left[ \sigma^k \left( \lambda I_K + \sum_\kappa \sigma^\kappa V_k \right)^{-1} \left( \sum_{k'} \mu_k \cdot \mu_k' \cdot \hat{m}_k \cdot \hat{m}_k' \cdot \sum_\kappa \sigma^\kappa \hat{Q}_k \right) \left( \lambda I_K + \sum_\kappa \sigma^\kappa \hat{V}_k \right)^{-1} \right]
\]
\[
m_k = E_{\sigma, \mu} \left[ \mu^k \left( \lambda I_K + \sum_\kappa \sigma^\kappa V_k \right)^{-1} \sum_\kappa \sigma^\kappa \hat{m}_k \right]
\]
\[
V_k = E_{\sigma, \mu} \left[ \sigma^k \left( \lambda I_K + \sum_\kappa \sigma^\kappa \hat{V}_k \right)^{-1} \right].
\]

where the expectations $E_{\sigma, \mu}$ are taken with respect to the joint distribution $\rho$.

C.1.1 Uniform covariances

Let us consider the simpler case $\Sigma_k \equiv \Delta I_k$, 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
\[
Q_k \equiv Q, \quad V_k \equiv V, \quad \hat{Q}_k \equiv \frac{1}{\Delta} \hat{Q}_k, \quad \hat{V}_k \equiv \frac{1}{\Delta} \hat{V}, \quad \forall k.
\]

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

- $M \in \mathbb{R}^{K \times K}$ (resp. $\hat{M} \in \mathbb{R}^{K \times K}$) is the matrix obtained concatenating the vectors $m_k$ (resp. $\hat{m}_k$);
- $\Theta = (\mu_k \mu_{k'}^\top)_{kk'}$ is the Gram matrix of the means;
- $F \in \mathbb{R}^{K \times K}$ is the matrix obtained concatenating the vectors $f_k$;
- $H \in \mathbb{R}^{K \times K}$ is the matrix obtained concatenating the vectors $h_k$;
- $\Pi = \text{diag}(\rho_k) \in \mathbb{R}^{K \times K}$ is a diagonal matrix with elements $\Pi_{kk'} = \delta_{kk'} \rho_k$.

The saddle-point equations then can be rewritten as
\[
Q = \Delta \left( \lambda I_K + \hat{V} \right)^{-1} \left( \hat{Q} + \hat{M} \Theta \hat{M}^\top \right) \left( \lambda I_K + \hat{V} \right)^{-1}
\]
\[
M = \left( \lambda I_K + \hat{V} \right)^{-1} \hat{M} \Theta
\]
\[
V = \Delta \left( \lambda I_K + \hat{V} \right)^{-1},
\]

Here and in the following $1_K$ 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 $\ell(y, x) = \frac{1}{2} (y - x)^2$, then an explicit formula for the proximal can be found, namely
\[
f_k = (I_K + V)^{-1} (e_K - \omega_k)
\]
so that the second set of saddle-point equations (187) can be written as
\[
\hat{Q} = \alpha (I_K + V)^{-1} \left[ (I_K - M - b \otimes 1_K) \Pi (I_K - M - b \otimes 1_K)^\top + Q \right] (I_K + V)^{-1}
\]
\[
\hat{M} = \alpha (I_K + V)^{-1} (I_K - M - b \otimes 1_K) \Pi
\]
\[
\hat{V} = \alpha \Delta (I_K + V)^{-1}.
\]

Observe at this point that we can explicitly solve for $V$ using the equation for it in eqs. (187). In particular, $V$ satisfies the equation $\lambda V^2 + (\alpha + \lambda - \Delta) V = \Delta I_K$. Being $V$ positive definite, it follows that it is diagonal, $V = V I_K$ with diagonal element
\[
V = \frac{\Delta (1 - \alpha) - \lambda + \sqrt{\Delta - \alpha \Delta - \lambda}^2 + 4 \Delta \lambda}{2 \lambda}, \quad \hat{V} = \frac{\alpha \Delta}{1 + V},
\]
so that

\[ Q = \alpha \left[ (I_K - M - b \otimes 1_K)\Pi (I_K - M - b \otimes 1_K)^\top + Q \right] \]

\[ \hat{M} = -\alpha (I_K - M - b \otimes 1_K)\Pi \]

\[ \hat{\lambda} = \frac{\alpha (I_K - M - b \otimes 1_K)\Pi}{1 + V}. \]

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

\[ Q = \Delta (1-\alpha)^2 \left( \hat{Q} + \hat{M} \hat{\theta} \hat{M}^\top \right), \quad \hat{Q} = \frac{\alpha \left[ (I_K - M - b \otimes 1_K)\Pi (I_K - M - b \otimes 1_K)^\top + Q \right]}{\Delta^2 (1-\alpha)^2}, \]

\[ \hat{M} = -\alpha (I_K - M - b \otimes 1_K)\Pi \]

\[ \hat{\Delta} = \frac{\alpha (I_K - M - b \otimes 1_K)\Pi}{\Delta (1-\alpha)}. \]

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

\[ \ell(y, x) = -\sum_{k=1}^{K} y_k \ln \frac{e^{x_k}}{\sum_{k'=1}^{K} e^{x_{k'}}}. \]

If \( y \in \{e_k\}_{k \in [K]} \), the loss can be written in the form \( \ell(y, x) = -y^\top x + \ln \sum_{k} e^{x_k} \). If we introduce the softmax function \( \text{softmax}: \mathbb{R}^K \to \mathbb{R}^K \)

\[ \partial_x \ell(y, x) = -y + \text{softmax}(x), \quad \text{softmax}(x) \equiv \frac{\exp(x_k)}{\sum_{k} \exp(x_k)} \]

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

\[ V^{-1}(h_k - \omega_k) - e_k + \text{softmax}(h_k) = 0 \iff f_k = e_k - \text{softmax}(V f_k + \omega_k) \quad \forall k \in [K], \]

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 \( Q \to \lambda^{-2} Q, V \to \lambda^{-1} V, M \to \lambda^{-1} M, \)

\[ b \to \lambda^{-1} b, V \to \lambda V. \]

They become

\[ Q = \Delta \left( I_K + \hat{V} \right)^{-1} \left( \hat{Q} + \hat{M} \hat{\theta} \hat{M}^\top \right) \left( I_K + \hat{V} \right)^{-1}, \quad \hat{Q} = \alpha \Delta \mathbb{E}_{\mathbb{E}} \left[ FF^\top \right], \]

\[ \hat{M} = \alpha \mathbb{E}_{\mathbb{E}} \left[ F \Pi \right], \quad \hat{\lambda} = \frac{\alpha \mathbb{E}_{\mathbb{E}} \left[ (H - M)\Pi \right]}{\Delta (1-\alpha)}. \]

so that the dependence on \( \lambda \) disappears everywhere except in the equation for the proximal \( f_k \)

\[ f_k = \arg\min_x \left[ \frac{1}{2} x^\top V x + \lambda \ell \left( e_k, \frac{V x + \omega_k}{\lambda} \right) \right], \]

which, in the \( \lambda \to 0 \) limit, becomes

\[ f_k = \arg\min_x \left[ \frac{1}{2} x^\top V x + \min_\mu \left\{ (e_{\mu} - e_k)^\top (V x + \omega_k) \right\} \right]. \]

Note that in this limit, minimising the cross-entropy loss yields precisely the max-margin estimator [70].

**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

\[ \varphi(x) = \text{sign}(x) \]
the expression of the average errors is
\[
\begin{align*}
\epsilon_g &= \sum_{k \in [2]} \rho_k \mathbb{E}_q[\theta((-1)^k \omega_k^*)] = \sum_{k \in [2]} \rho_k \frac{1}{2} \text{erfc} \left(\frac{(-1)^{k-1} m_k^* + b^*}{\sqrt{2} \sigma_k} \right), \\
\epsilon_t &= \sum_{k \in [2]} \rho_k \mathbb{E}_q[\theta((-1)^k h_k^*)], \\
\epsilon_f &= \sum_{k \in [2]} \rho_k \mathbb{E}_q[f((-1)^k, h_k^*)].
\end{align*}
\]
(200)

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

C.2.1 Example: $\ell_1$ regularization

In this Section we derive the saddle-point equations for the case in which the two clusters have opposite means $\mu_1 = -\mu_2 \equiv \mu$, 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:
\[
\begin{align*}
V_1 &= V_2 \equiv V, \quad q_1 = q_2 \equiv q, \quad m_+ = -m_- \equiv m, \\
\hat{V}_+ &= \hat{V}_- \equiv \frac{1}{2} \hat{V}, \quad \hat{q}_+ = \hat{q}_- \equiv \frac{1}{2} \hat{q}, \quad \hat{m}_+ = -\hat{m}_- \equiv \frac{1}{2} \hat{m}.
\end{align*}
\]
(201)

We define
\[
\frac{1}{d} \sum_{i=1}^d \delta(\sigma - \sigma_i) \delta(\mu - \sqrt{d} \mu_i) \rightarrow \infty \quad p(\sigma, \mu)
\]
(203)

joint distribution of the covariance diagonal elements and of the mean elements. We will denote $\mathbb{E}_{\mu, \sigma}[\cdot]$ 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 $\ell_1$ regularization, i.e., $r(w) = \sum_i |w_i|$, and let us introduce the corresponding soft-thresholding operator:
\[
\text{Prox}_{\lambda V} |x| = \text{sign}(x) \max\{|x| - \lambda, 0\}.
\]
(204)

Observe that $\text{Prox}_{\lambda V} |(\alpha x)| = \alpha \text{Prox}_{\lambda V} |x|$ for $\alpha > 0$. Its derivative given by $\text{Prox}_{\lambda V} |x| = \theta(|x| > \lambda)$. The saddle point equations from the prior part simply read:
\[
\begin{align*}
V &= \frac{1}{V} \mathbb{E}_{\mu, \sigma, \xi} \left[ \text{Prox}_{\lambda V} \left( \frac{\hat{m} \mu + \sqrt{\hat{q} \sigma} \xi}{V \sigma} \right) \right], \\
q &= \mathbb{E}_{\mu, \sigma, \xi} \left[ \sigma \left( \text{Prox}_{\lambda V} \left( \frac{\hat{m} \mu + \sqrt{\hat{q} \sigma} \xi}{V \sigma} \right) \right)^2 \right], \\
m &= \mathbb{E}_{\mu, \sigma, \xi} \left[ \mu \text{Prox}_{\lambda V} \left( \frac{\hat{m} \mu + \sqrt{\hat{q} \sigma} \xi}{V \sigma} \right) \right].
\end{align*}
\]
(205)

(206)

(207)

The averages over $\xi$ can be performed explicitly using the simple expression of the proximal in this case. If we define the auxiliary functions
\[
\begin{align*}
\phi_0\pm(v, u, \lambda) &\equiv \frac{1}{2} \text{erfc} \left(\frac{\lambda \pm v}{\sqrt{2} u} \right), \\
\phi_1\pm(u, v, \lambda) &\equiv \sqrt{\frac{u}{2 \pi}} e^{-\frac{(u+\lambda v)^2}{2u}} - \frac{v \pm \lambda}{2} \text{erfc} \left(\frac{\lambda \pm v}{\sqrt{2} u} \right), \\
\phi_2\pm(u, v, \lambda) &\equiv -\sqrt{\frac{u}{2 \pi}} e^{-\frac{(u+\lambda v)^2}{2u}} \left(\lambda \pm v\right) + \frac{u + (\lambda \pm v)^2}{2} \text{erfc} \left(\frac{\lambda \pm v}{\sqrt{2} u} \right).
\end{align*}
\]
(208)

then
\[
\begin{align*}
V &= \frac{1}{V} \mathbb{E}_{\mu, \sigma} \left[ \phi_0\pm(\mu \hat{m}, \sigma \hat{q}, \lambda) + \phi_0\pm(\mu \hat{m}, \sigma \hat{q}, \lambda) \right], \\
q &= \mathbb{E}_{\mu, \sigma} \left[ \phi_1\pm(\mu \hat{m}, \sigma \hat{q}, \lambda) + \phi_1\pm(\mu \hat{m}, \sigma \hat{q}, \lambda) \right], \\
m &= \mathbb{E}_{\mu, \sigma} \left[ \mu \phi_2\pm(\mu \hat{m}, \sigma \hat{q}, \lambda) - \mu \phi_2\pm(\mu \hat{m}, \sigma \hat{q}, \lambda) \right].
\end{align*}
\]
(209)
Gaussian means, homogenous covariances — If \( p(\mu, \sigma) = \mathcal{N}(\mu, 1)\delta(\sigma - \Delta) \), i.e., the means have i.i.d. Gaussian entries and \( \Sigma = \Delta I_d \), then

\[
V = \frac{1}{V} \left[ \text{erfc} \left( \frac{\lambda + \bar{m}z}{\sqrt{2\Delta q}} \right) \right],
\]

\[
q = \frac{1}{\Delta V^2} \left\{ \frac{e^{-\frac{1}{2} \bar{m}^2 (\Delta q + \lambda)^2}}{2\pi(\Delta q + \lambda)^2} \frac{2(\Delta q)^2\lambda}{\bar{m}^2 + \Delta q} + \text{erfc} \left( \frac{\lambda + \bar{m}z}{\sqrt{2\Delta q}} \right) \right\},
\]

\[
m = \frac{1}{\Delta V} \left\{ \frac{e^{-\frac{1}{2} \bar{m}^2 (\Delta q + \lambda)^2}}{2\pi(\Delta q + \lambda)^2} \frac{2\Delta q\lambda}{\bar{m}^2 + \Delta q} + \text{erfc} \left( \frac{\lambda + \bar{m}z}{\sqrt{2\Delta q}} \right) \right\},
\]

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

\[ p(\sigma, \mu) = p\mathcal{N}(\mu|0, 1)\delta(\sigma - \Delta) + (1 - p)\delta(\mu|0) \delta(\sigma - \Delta_0). \]

The saddle-point equations are therefore

\[
V = \frac{1}{V} \left[ p\text{erfc} \left( \frac{\lambda + \bar{m}\mu}{\sqrt{2\Delta_0 q}} \right) \right] + (1 - p)\text{erfc} \left( \frac{\lambda}{\sqrt{2\Delta_0 q}} \right)
\]

\[
q = \frac{p}{\Delta_1 V^2} \left\{ \frac{e^{-\frac{1}{2} \bar{m}^2 (\Delta q + \lambda)^2}}{2\pi(\Delta q + \lambda)^2} \frac{2(\Delta q)^2\lambda}{\bar{m}^2 + \Delta q} + \text{erfc} \left( \frac{\lambda + \bar{m}z}{\sqrt{2\Delta q}} \right) \right\}
\]

\[
- \lambda(1 - p) \sqrt{\frac{\Delta_0 q}{2\pi}} e^{-\frac{\lambda^2}{2\Delta_0 q}} + \frac{1 - p}{2} (\Delta_0 q + \lambda^2) \text{erfc} \left( \frac{\lambda}{\sqrt{2\Delta_0 q}} \right)
\]

\[
m = \frac{p}{\Delta_1 V} \left\{ \frac{e^{-\frac{1}{2} \bar{m}^2 (\Delta q + \lambda)^2}}{2\pi(\Delta q + \lambda)^2} \frac{2\Delta q\lambda}{\bar{m}^2 + \Delta q} + \text{erfc} \left( \frac{\lambda + \bar{m}z}{\sqrt{2\Delta q}} \right) \right\}.
\]

In Section 3.1 we compare the performance obtained adopting an \( \ell_1 \) regularization with the corresponding one obtained using \( \ell_2, r(w) = \sum_i w_i^2 \). 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 \( \Psi_w \) can be written explicitly after a Gaussian integration as

\[
\Psi_w(\bar{m}, \bar{Q}, \bar{V}) = -\frac{1}{2d} \text{tr} \ln \left( \lambda I_d + \bar{V} \Sigma \right) + \frac{1}{2} \text{tr} \left( \left( \lambda I_d + \bar{V} \Sigma \right)^{-1} \left( \bar{m}^2 \mu \mu^\top + \frac{\bar{q}}{d} \Sigma \right) \right).
\]

In the setting given by eq. (211) the saddle point equations are then

\[
q = p \frac{\bar{m}^2 \Delta_1 + \bar{q} \Delta_2}{(\lambda + V \Delta_1)^2} + \frac{1 - p}{2} \frac{\bar{q} \Delta_2}{(\lambda + V \Delta_0)^2}
\]

\[
V = p \frac{\Delta_1}{\lambda + V \Delta_1} + \frac{1 - p}{2} \frac{\Delta_0}{\lambda + V \Delta_0}
\]

\[
m = \frac{\bar{m} p}{\lambda + V \Delta_1}.
\]
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 $\Sigma_k = \Delta I_d$ in the large $d$ limit, assuming that a dataset \{$(x^\nu, y^\nu)$\}$_{\nu \in [n]}$ of correctly labeled points is available. As usual, we will assume $n/d = \alpha$ finite. The distribution of a pair $(y, x)$ is given by

$$p(y, x | M) = \sum_k y_k \rho_k \exp \left( -\frac{1}{2\Delta} \|x - \mu_k\|^2 \right) \left( 2\pi \right)^{-d/2} \delta_k(x) .$$

(217)

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

$$p(M|y^\nu, x^\nu) \propto p([x^\nu]_\nu|M, [y^\nu]_\nu) P_\mu(M)$$

(218)

$$\propto P_\mu(M) \prod_{\nu=1}^n y_k^\nu \exp \left( -\frac{1}{2\Delta} \|x^\nu - \mu_k\|^2 \right) .$$

We will assume in the following the distribution

$$P_\mu(M) = \frac{\exp \left( -\frac{1}{2\Delta} tr[M \Theta^{-1} M^T] \right)}{(2\pi)^{dK/2} |\Theta|^{1/2}}$$

(219)

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

$$\mathbb{E} \left[ M^T M \right] = \Theta .$$

(220)

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

$$p(y^0 | x^0, \{y^\nu, x^\nu\}_\nu) \propto \mathbb{E}_{M|y^\nu, x^\nu} \left[ p(y, x | M) \right]$$

(221)

$$= \int dM P_\mu(M) \sum_k y_k^0 \rho_k \exp \left( -\frac{1}{2\Delta} \|x^0 - \mu_k\|^2 \right) \prod_{\nu=1}^n \sum_k y_k^\nu \exp \left( -\frac{1}{2\Delta} \|x^\nu - \mu_k\|^2 \right) .$$

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

$$p(y^0 | x^0, \{y^\nu, x^\nu\}_\nu) \propto \int dM P_\mu(M) \prod_{k=1}^K \rho_k^n \exp \left( -\frac{1}{2\Delta} \|x^0 - \mu_k\|^2 \right)$$

$$= \exp \left[ \sum_k y_k^0 \left( \ln \rho_k - \frac{1}{2\Delta} \|x^0\|^2 \right) \right] \times \exp \left[ \frac{1}{2\Delta} \text{tr} \left( \sum_{\nu=0}^n y^\nu \otimes x^\nu \right)^T (d\Delta \Theta^{-1} + \text{diag}(n + y^0)\Theta)^{-1} \left( \sum_{\nu=0}^n y^\nu \otimes x^\nu \right) \right] .$$

(222)

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

$$p(y^0 | x^0, \{y^\nu, x^\nu\}_\nu) \propto \exp \left[ \sum_k y_k \left( \ln \rho_k - \frac{1}{2\Delta} \|x^0\|^2 \right) \right] \times \exp \left[ \frac{1}{2\Delta} \text{tr} \left( \frac{1}{d} \sum_{\nu=0}^n y^\nu \otimes x^\nu \right)^T (\Delta \Theta^{-1} + \alpha \Pi)^{-1} \left( \sum_{\nu=0}^n y^\nu \otimes x^\nu \right) \right] .$$

(223)

Observe now that

$$\frac{1}{d} \sum_{\nu=0}^n y^\nu \otimes x^\nu \overset{n,d \to +\infty}{\longrightarrow} \alpha \rho_k \Theta \ast_k + \eta_k Z_k / \Delta , \quad \eta_k \equiv \sqrt{\Delta \left( 1 + \frac{\Delta}{\alpha \rho_k} \right) . \quad Z_k \sim \mathcal{N}(0, 1) ,}$$

(224)
so that, defining the vector \( \mathbf{a}^* = (a_k)_{k \in [K]} \) with elements

\[
a_k^* \equiv \alpha \rho_k \Theta_{\cdot,k} + \eta_k Z_k / \Delta,
\]

and neglecting the \( y^0 \)-independent contributions, the expression above can be rewritten as

\[
p(y^0|\mathbf{x}^0, \{\mathbf{y}^\nu, \mathbf{x}^\nu\}_\nu) \propto \exp \left[ \sum_k y^0_k \ln \rho_k + \left( \mathbf{a}^* + \frac{1}{2} y^0 \right)^\top (\Delta \Theta^{-1} + \alpha \Pi)^{-1} y^0 \right]
\]

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

\[
\epsilon_{BO}^g = \sum_k \rho_k \mathbb{P} \left[ \arg \max_{\kappa} \left( \ln \rho_{\kappa} + \left( \mathbf{a}^k + \frac{1}{2} e_{\kappa} \right)^\top (\Delta \Theta^{-1} + \alpha \Pi)^{-1} e_{\kappa} \right) \neq k \right].
\]

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

\[
\epsilon_{BO}^g = \mathbb{P} \left[ \frac{1}{\eta} < \max_{\kappa \in [K-1]} Z_{\kappa} + Z \right],
\]

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 \( \{(x^r, y^r)\}_{r=1}^{n_{\text{tot}}} \) with \( n_{\text{tot}} \) samples which we assume are independent. As a pre-processing step we center, normalise and flatten the inputs \( x^r \) into \( d \)-dimensional vectors. For both the MNIST [61] and Fashion-MNIST [62] data sets used in the experiments we have normalised the inputs by 255, such that components \( x^r_i \in [0,1] \). In what follows we focus on binary classification tasks and encode the labels as \( y^r \in \{-1,1\} \). For example, for the MNIST and Fashion-MNIST data sets we have \( d = 784 \) and \( n_{\text{tot}} = 7 \times 10^3 \), 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:

\[
\hat{P}(x, y) = \frac{1}{n_{\text{tot}}} \sum_{r=1}^{n_{\text{tot}}} \delta(x - x^r)\delta(y - y^r)
\]  

(229)

The question we want to answer is: how well can we approximate the learning curves \((\epsilon_g, \epsilon_1)\) 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 \( \hat{P} \):

\[
\hat{\mu}_k = \frac{1}{n_{\text{tot}}} \sum_{r=1}^{n_{\text{tot}}} x^r \mathbb{1}(x^r \in C_k), \quad \hat{\Sigma}_k = \frac{1}{n_{\text{tot}}} \sum_{r=1}^{n_{\text{tot}}} (x^r - \hat{\mu}_k)(x^r - \hat{\mu}_k)^T \mathbb{1}(x^r \in C_k)
\]

(230)

for \( k \in \{+, -\} \) labelling the two clusters. Similarly, the class probabilities \( \hat{\rho}_k \) are also estimated from the full data set:

\[
\hat{\rho}_k = \frac{1}{n_{\text{tot}}} \sum_{r=1}^{n_{\text{tot}}} \mathbb{1}(x^r \in C_k).
\]

(231)

The parameters \((\hat{\mu}_k, \hat{\Sigma}_k, \hat{\rho}_k)\) completely characterise the approximating Gaussian mixture distribution \( P_2 \), and together with Theorem 1 can be used to compute the theoretical learning curves \((\epsilon_g, \epsilon_1)\) 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 \( \varphi : \mathbb{R}^d \to \mathbb{R}^p \) is applied to the data prior to fitting. The only difference is that the empirical distribution \( \hat{P} \) is defined over the features \( \{(\varphi(x^r), y^r)\}_{r=1}^{n_{\text{tot}}} \) where \( \varphi(x^r) = \varphi(x^r) \), 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 \( \varphi = \text{erf}(Fx) \) with \( F \in \mathbb{R}^{p \times d} \) 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 \( n = nd \) 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 \( \varphi \) 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:

```python
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()
)
```

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 \(\varphi\) obtained by training 2-layer fully connected neural network, with \(\ell_2\) penalty and fixed \(\lambda = 0.05\). The different curves show the performance at different stages of training.

### Multiclass vs. binary approximation

In the cases previously discussed, we have considered a \(K = 2\) cluster approximation \(P_2\) to the empirical data distribution \(P\). 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 \(\lambda = \mathbf{W}x\) 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 \(\Sigma_k = \Delta I_d\) and with means:

\[
\mu_1 = e_1 + e_2, \quad \mu_2 = e_1 - e_2, \quad \mu_3 = -e_1 + e_2, \quad \mu_4 = -e_1 - e_2
\]  

(232)

where \(e_i \in \mathbb{R}^d\) 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 \((\mu_{\pm}, \Sigma_{\pm})\) 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 \(P_2\).
Figure 8: Two dimensional projection of the setting described in eq. (232). (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 $\ell_2$ 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 $\Delta = 0.5$.

**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 $q_{\pm} = W^\top \Sigma_{\pm} W$ – which is clearly not possible since $\Sigma_{\pm}$ 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 $\lambda$ below which the spurious fixed point develop seems to depend only on the conditioning number of the covariance matrices.

31
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 \(\ell_2\) 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.

References


