Asymptotic Bayes risk of semi-supervised multitask learning on Gaussian mixture

Minh-Toan Nguyen
GIPSA-lab, Université Grenoble Alpes

Romain Couillet
LIG-lab, Université Grenoble Alpes

Abstract

The article considers semi-supervised multitask learning on a Gaussian mixture model (GMM). Using methods from statistical physics, we compute the asymptotic Bayes risk of each task in the regime of large datasets in high dimension, from which we analyze the role of task similarity in learning and evaluate the performance gain when tasks are learned together rather than separately. In the supervised case, we derive a simple algorithm that attains the Bayes optimal performance.

1 INTRODUCTION

Multitask learning (MTL) is a machine learning method in which multiple tasks are learned simultaneously. It can facilitate knowledge transfer between tasks and can lead to more informative data representation (Ruder, 2017). Although learning from related tasks can help disseminate useful information learned from one task to other tasks, the presence of unrelated tasks can also be beneficial. With the prior knowledge that two given tasks are unrelated, the algorithm can learn to ignore irrelevant features of the data distribution, resulting in better data representation (Paredes et al., 2012).

In this work, we propose a simple model of MTL based on Gaussian mixtures that focuses on capturing the transfer of knowledge between tasks, leaving out the data representation aspect. Our paper extends the semi-supervised learning model studied in Lelarge and Miolane (2019a), which examines the added value of unlabeled data in a one-task classification. We consider here instead multiple classification tasks, for which the data in each task are partially labeled and come from two classes. Thanks to the simplicity of our model, we can define the correlation between two tasks as a number in $[-1, 1]$. We are interested in the performance gain when correlated tasks are learned together versus when they are learned separately, assuming the best algorithm is used. This leads to the concept of *Bayes risk*, defined as the minimal feasible probability of misclassifying a new data point not from the training dataset. Despite the randomness of data, in the limit where both the quantity and the dimensionality of the data are large with a fixed ratio, the Bayes risk converges towards a deterministic value.

Although the main objective of this study is to compute the minimum classification error, it is important to emphasize that the posterior distribution of a signal given the observed data is a more fundamental object, as it serves as a basis for deriving optimal estimators with respect to certain criteria. In the high-dimensional regime, the posterior law of a signal is a high-dimensional integral, and despite its complexity, it behaves like a simpler law. This property enables the exact calculations obtained in this work.

Contributions and related works.

As a first contribution, we derive an exact formula for the asymptotic Bayesian risk, based on a simple argument that is similar to the cavity method from statistical physics (Mezard and Montanari, 2009). Although not fully rigorous, the paper aims to provide a clear intuition of the asymptotic equivalence that occurs in high dimensions. This concept underlies most of the equations presented in the paper. The paper is designed to be accessible, and no prior knowledge of physics is required to understand its contents. Our work aligns with a body of research that studies the fundamental limit of various high-dimensional statistical models, including tensor models (Barbier et al., 2017; Lesieur et al., 2017; Lelarge and Miolane, 2019b), generalized linear model (Barbier et al., 2019) and Gaussian mixture model (Lesieur et al., 2016; Lelarge and Miolane, 2019a).

Secondly, we analyze the role of task correlations and how they interact with other elements of the model, such as the proportion of labeled data in each task. It is well known that unsupervised learning on a single task with Gaussian mixture data leads to a phase transition that separates the high and low noise regimes. We demonstrate that phase transition persists to the case of multitask and study how

Proceedings of the 26th International Conference on Artificial Intelligence and Statistics (AISTATS) 2023, Valencia, Spain. PMLR: Volume 206.

it is affected by task correlations. In the context of source task - target task, we identify the conditions in which the source task is most beneficial to the target task.

Finally, we derive a simple algorithm that achieves the optimal performance in the case of supervised learning. Although an optimal performance on a synthetic data set does not necessarily have a good performance on real data, this algorithm shows how the optimal algorithms on separate tasks should be modified when correlations are taken into account. This could offer useful insights for designing MTL methods in practice.

Although our focus is different, there is some connection between our work and theoretical studies that investigate optimization-based inference on simple data models. These studies compute the exact asymptotic performance of algorithms, and examine how this performance is influenced by factors such as choice of loss function, regularization, and number of model parameters. On the other hand, our work focuses on investigating the fundamental limit of statistical problems regardless of any specific algorithm. It is interesting, however, that in some cases, the optimization-based methods can nearly reach or achieve the optimal performance (Mai and Couillet, 2021; Thrampoulidis et al., 2020; Mignacco et al., 2020; Loureiro et al., 2021; Aubin et al., 2020). For multitask learning on Gaussian mixtures, Tiomoko et al. (2021b) obtain exact asymptotic results for least-square support vector machine using random matrix theory.

There are several reasons to study the GMM. Besides being amenable to theoretical analysis, it is the simplest model that captures the elements of MTL that we are interested in: task correlation and transferring of information. On the application aspect, it is remarked in Lesieur et al. (2016) that the Bayesian statistics under GMM as a prior can rediscover several key methods in machine learning such as the K-means or spectral clustering algorithms. There exists a close relationship between Bayesian statistics and algorithms. On one hand, Bayesian interpretations can be established for well-known algorithms such as PCA and SVM (Tipping and Bishop, 1999; Polson and Scott, 2011), which turn out to be standard estimators on fairly simple data models. On the other hand, by starting with a simple data model and devising an optimal algorithm based on specific criteria, one can enhance existing methods or create new ones. (Bishop, 1998; Krzakala et al., 2012).

Notation: We use the symbol $\langle \cdot, \cdot \rangle$ to denote the scalar (or inner) product of vectors. If $\boldsymbol{X} = (X_{ij})$, then $\boldsymbol{X}_{i\cdot} = (X_{ij})_j$ and $\boldsymbol{X}_{\cdot j} = (X_{ij})_i$. For $n \in \mathbb{N}$, we use $[n]$ to denote the set $\{1, 2, \ldots, n\}$. The notation $\boldsymbol{D}_{\boldsymbol{x}}$ represents the diagonal matrix with diagonal elements given by the vector $\boldsymbol{x}$. If indexed objects such as $\boldsymbol{X}_i$ are given, then $\boldsymbol{X}$ simply means $(\boldsymbol{X}_i)_i$.

The source code for the simulations in this paper is available at: https://github.com/Minh-Toan/Bayes-risk

2 MODEL

We consider T classification tasks, where task t consists of N_t data points in $\mathbb{R}^D$. The i-th data point in task t, denoted by $\boldsymbol{Y}_{ti}$, is given by

$$\boldsymbol{Y}_{ti} = V_{ti}\boldsymbol{U}_t + \sigma_t \boldsymbol{Z}_{ti} \tag{1}$$

where $\sigma_t > 0$. The random variables $\boldsymbol{V}, \boldsymbol{U}, \boldsymbol{Z}$ are independent, with

$$V_{ti} \overset{i.i.d}{\sim} \mathcal{U}(\{-1, 1\}),$$
$$Z_{ti} \overset{i.i.d}{\sim} \mathcal{N}(0, I_D),$$

and $\boldsymbol{U}_1, \ldots, \boldsymbol{U}_T$ are chosen uniformly randomly on the unit sphere $S^{D-1} = \{\boldsymbol{x} \in \mathbb{R}^D, \|\boldsymbol{x}\| = 1\}$, conditioned on the event

$$\langle \boldsymbol{U}_t, \boldsymbol{U}_{t'} \rangle = C_{tt'}, t \neq t'.$$

The matrix $\boldsymbol{C} = (\langle \boldsymbol{U}_t, \boldsymbol{U}_{t'} \rangle)_{t,t'=1}^T$ is called the *task-correlation matrix*. It follows from the definition that $\boldsymbol{C}$ is a positive definite matrix with diagonal entries all equal to 1. The tasks are said to be *connected* if for any two tasks t and t', there is a sequence of tasks $t_1, \ldots, t_k$ such that $C_{tt_1}, C_{t_1t_2} \ldots C_{t_kt'} \neq 0$.

In other words, the data in task t comes from two classes corresponding to two Gaussian distributions centered at $\pm\boldsymbol{U}_t$ with the same covariance $\sigma_t^2 I_D$. The positions of the centers are not known and can only be estimated from the data. The class of a data point $\boldsymbol{Y}_{ti}$ is indicated by V_{ti}, so each data point has probability $1/2$ of belonging to each class. A data point is said to be *labeled* if we know which class it belongs to, otherwise it is *unlabeled*. Independently of all other random variables, each data point in task t is labeled with probability η_t. The cases $\eta_t = 1$ and $\eta_t = 0$ correspond to supervised and unsupervised learning. $C_{tt'}$ measures the correlation between tasks t and t'. The parameters $\lambda_t = 1/\sigma_t^2$ are called the *signal to noise ratio* (SNR). As the SNR increases, the two classes separate and classification is easier. We study the model in the setting where the dimension and the amount of data in each task tends to infinity at a fixed rate $\alpha_t = \lim_{D\to\infty} N_t/D$, called the *sampling ratio*. Note that the model for semi-supervised learning studied in Lelarge and Miolane (2019a) corresponds to the case $T = 1$.

We have access to the dataset $\boldsymbol{Y} = (\boldsymbol{Y}_{ti})$, the labels as well as model parameters $(\sigma_t), (\eta_t), (\alpha_t)$ and $\boldsymbol{C}$. [1] Our job

[1] $\boldsymbol{\sigma}$ and $\boldsymbol{C}$ can indeed be estimated with vanishing errors as $D \to \infty$, given that a positive fraction of labeled data is available in each task, i.e. $\eta_t > 0$ for all t (Appendix D).

is to use that available information to classify a new data point $\boldsymbol{Y}_{\text{new}}$ in any given task t

$$\boldsymbol{Y}_{\text{new}} = V_{\text{new}}\boldsymbol{U}_t + \sigma_t \boldsymbol{Z}_{\text{new}} \tag{2}$$

We are interested in the minimal classification error, i.e. the Bayes risk

$$\inf_{\hat{V}} \mathbb{P}(\hat{V} \neq V_{\text{new}}) \tag{3}$$

where the infimum is taken over all estimators of V_{new}.

3 RESULTS

Before presenting the results, we need some definitions that will aid in formulating our findings in a clear and concise manner.

Definition 3.1. The inference of $\boldsymbol{X} \in \mathbb{R}^D$ from the data $\boldsymbol{Y}$ satisfies the *replica symmetric* (RS) property with *overlap* q if in the limit $D \to \infty$,

$$\langle \boldsymbol{X}, \boldsymbol{X}^1 \rangle,\ \langle \boldsymbol{X}^1, \boldsymbol{X}^2 \rangle,\ \langle \boldsymbol{X}, \hat{\boldsymbol{X}} \rangle,\ \|\hat{\boldsymbol{X}}\|^2 \tag{4}$$

all converge to the same limit q, where $\boldsymbol{X}^1, \boldsymbol{X}^2$ are sampled independently from the posterior of $\boldsymbol{X}$ given $\boldsymbol{Y}$, and $\hat{\boldsymbol{X}} = \mathbb{E}[\boldsymbol{X}|\boldsymbol{Y}]$, called the *MMSE estimator* of $\boldsymbol{X}$ given $\boldsymbol{Y}$. [2] In some contexts, we use $\hat{\boldsymbol{X}}$ to refer to a general estimator of $\boldsymbol{X}$, while the MMSE estimator of $\boldsymbol{X}$ is denoted as $\hat{\boldsymbol{X}}_{\text{MMSE}}$.

This property holds for a wide range of inference problems in the setting where the signal is generated from a known distribution. We assume that this property holds true for our model:

Assumption. *$\sigma_t^{-1}\boldsymbol{U}_t|\boldsymbol{Y}$ and $N_t^{-1/2}\boldsymbol{V}_t|\boldsymbol{Y}$ satisfies the RS property for all $t \in [T]$ with the overlaps denoted by q_{ut} and q_{vt} respectively.*

The inclusion of the normalizing factor σ_t in the definition of q_{ut} is for the purpose of convenience.

Later in the paper, we will require the following definition in order to prove the results:

Definition 3.2. Consider the following Gaussian channels

$$Y_i = \sqrt{\lambda_i} X_i + Z_i, \quad i = 1, \dots, n \tag{5}$$

with inputs X_i, outputs Y_i and SNRs λ_i. Let $\hat{\boldsymbol{X}} = \mathbb{E}[\boldsymbol{X}|\boldsymbol{Y}]$. The *overlap functions* $F_{\boldsymbol{X},i} : \mathbb{R}^n \to \mathbb{R}$ are defined as

$$F_{\boldsymbol{X},i}(\boldsymbol{\lambda}) = \mathbb{E}[\hat{X}_i X_i] = \mathbb{E}[\hat{X}_i^2] \tag{6}$$

$F_{\boldsymbol{X},i}$ is also referred to as the *overlap* of the signal X_i.

[2] MMSE stands for minimum mean-squared error.

The main result of the article unfolds as follows.

Result. *i) Under the setting of the model, as $D \to \infty$, the Bayes risk converges to*

$$1 - \Phi(\sqrt{q_{ut}}),$$

where $\Phi(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{t} e^{-x^2} dx$

ii) The overlaps q_{ut}, q_{vt} satisfies the following equations

$$q_{ut} = [\boldsymbol{M} - \boldsymbol{M}(\boldsymbol{I} + \boldsymbol{D}\boldsymbol{M})^{-1}]_{tt} \tag{7a}$$
$$q_{vt} = \eta_t + (1 - \eta_t) F(q_{ut}) \tag{7b}$$

with

$$\boldsymbol{M} = \{C_{tt'}/\sigma_t\sigma_{t'}\}_{t,t'=1}^{T}$$
$$\boldsymbol{D} = \text{diag}\{\alpha_t q_{vt}\}_{t=1}^{T}$$
$$F(q) = \mathbb{E}[\tanh(\sqrt{q}Z + q)], \quad Z \sim \mathcal{N}(0,1).$$

Remark 3.1. When $q_{ut} = 0$, the Bayes risk of task t is equal to 0.5, which corresponds to the level of classification error of a random guess. In this case, we say that the classification of task t is *impossible*. On the other hand, if q_{ut} is positive, the classification of task t is said to be *feasible*.

Remark 3.2. The fixed point equations (7a) and (7b) may not uniquely determine the overlaps. Specifically, for unsupervised learning with high SNR, two solutions exist: the zero solution is unstable while the non-zero solution is stable, and the stable solution is naturally chosen as overlaps. In other cases, there is only one solution.

We can perform a sanity check of the result by considering the following special cases: if the similarity between any two tasks is zero, the result implies that MTL has the same asymptotic Bayes risks as learning task separately, which is obvious since the data from different tasks are independent, while if $\sigma_t = \sigma$ and $C_{tt'} = 1$ for all t, t', i.e. the data distributions are identical for all tasks, the asymptotic Bayes risks of all tasks are equal to that of a single task with parameters $\alpha = \sum_t \alpha_t$ and $\alpha\eta = \sum_t \alpha_t \eta_t$ (Appendix B).

4 CONSEQUENCES

We present in this section some implications of the main result.

4.1 Supervised learning.

For supervised learning with only one task, the minimal classification error of a new data point $\boldsymbol{Y}_{\text{new}}$ is achieved by the estimator $\hat{V}_{\text{new}} = \text{sgn}(\langle \boldsymbol{Y}_{\text{new}}, \bar{\boldsymbol{Y}} \rangle)$, where $\bar{\boldsymbol{Y}} = N^{-1} \sum_i V_i \boldsymbol{Y}_i$ (Lelarge and Miolane, 2019a). In the multi-task case, if $\boldsymbol{Y}_{\text{new}}$ is a new data point in task t, the following algorithm achieves the optimal performance:

1. Compute

$$\bar{\boldsymbol{Y}}_t = \frac{1}{N_t}\sum_{i=1}^{N_t} V_{ti}\boldsymbol{Y}_{ti}$$

2. Compute

$$\tilde{\boldsymbol{Y}}_t = \sum_{s=1}^{T} a_{ts}\bar{\boldsymbol{Y}}_s$$

where $\boldsymbol{A} = (a_{ts})_{t,s=1}^{T} = \boldsymbol{M}\boldsymbol{D}_{\boldsymbol{\alpha}}(\boldsymbol{I} + \boldsymbol{M}\boldsymbol{D}_{\boldsymbol{\alpha}})^{-1}$.

3. The asymptotic Bayes risk is achieved by

$$\hat{V}_{\text{new}} = \text{sgn}(\langle \boldsymbol{Y}, \tilde{\boldsymbol{Y}}_t \rangle). \tag{8}$$

We can see that the optimal estimator for multiple tasks modifies the optimal estimators for separated tasks $\bar{\boldsymbol{Y}}_t$ by taking into account the correlations between tasks as well as their levels of difficulty and the relative sizes, measured by $\boldsymbol{C}$, (σ_t) and (α_t) respectively. Interestingly, this optimal algorithm coincides with the method proposed in Tiomoko et al. (2021a) using a different approach.

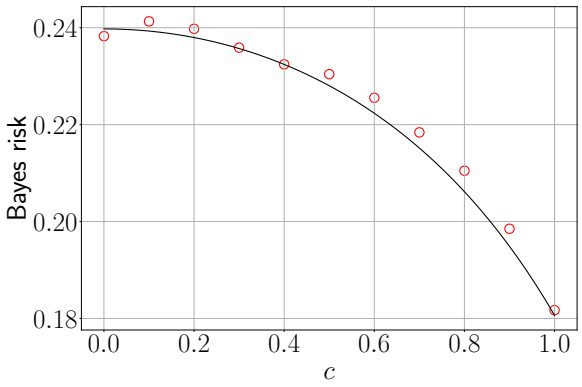

FIGURE 1: Bayes risk vs performance of the asymptotic optimal algorithm. $\alpha_1 = \alpha_2 = 1$, $\sigma_1 = 1$, $\sigma_2 = 0.5$, $D = 1000$.

4.2 Unsupervised learning and phase transition.

A particularly interesting behavior that only occurs in the case of unsupervised learning is phase transition. One of the most well-known example of this phenomenon is *BBP phase transition* (Baik et al., 2005) which concerns a single learning task with $\lim_{D\to\infty} N/D = 1$. When $\lambda = 1/\sigma^2 \leq 1$, no estimator can achieve a smaller classification error than 0.5. In other words, the classification is objectively impossible since the two classes are statistically identical. On the other hand, we say that a task is *feasible* if one can obtain a classification error smaller than 0.5. It turns out that phase transition persists to the case of multitask. Fig. 2 shows the performance of task 1 in terms of SNRs in the case of two tasks with $N_1 = N_2 = D$ and correlation $c = 0.7$. The classification is impossible in the region delimited by the black curve.

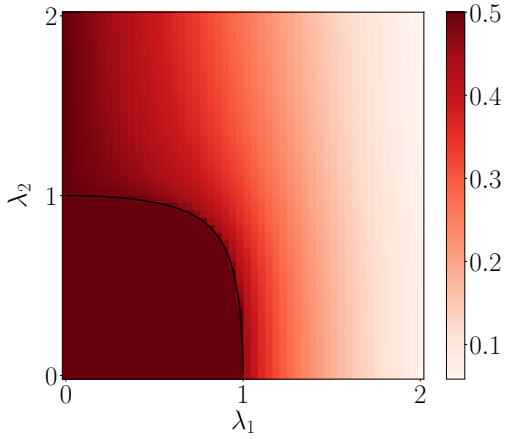

FIGURE 2: Bayes risk of Task 1 in terms of SNR of each task. Two tasks are unsupervised, with $N_1 = N_2 = D$ and correlation $c = 0.7$. The classification is impossible in the region delimited by the black curve. The impossible region is identical for two tasks.

The simulation also shows that the impossible regions are identical for both tasks. In other words, two correlated tasks are either feasible or impossible. In the general case with any number of tasks, tasks are feasible or impossible together, given that they are connected.

Note that phase transition disappears as soon as a positive proportion of labeled data is available, since supervised learning restricted on labeled data already produces a non-trivial performance.

In the case of two tasks with $N_1 = N_2 = D$, the region of impossible classification is given by

$$\left\{(\lambda_1, \lambda_2) \in [0,1]^2 : (1-\lambda_1^2)(1-\lambda_2^2) \geq c^4\lambda_1^2\lambda_2^2\right\} \tag{9}$$

as shown in Figure 3. As the task correlation c increases from 0 to 1, this region shrinks from the unit square $[0,1]^2$ to a quarter of a disk.

Another special case where an explicit formula for the impossible region can be obtained is when there are T tasks with $N_1 = \cdots = N_T = D$, with correlation $c > 0$ between any two of them, and $\lambda_t = \lambda$ for all t. It can be shown that the classification is impossible whenever

$$\lambda \leq \frac{1}{\sqrt{1 + (T-1)c^2}}. \tag{10}$$

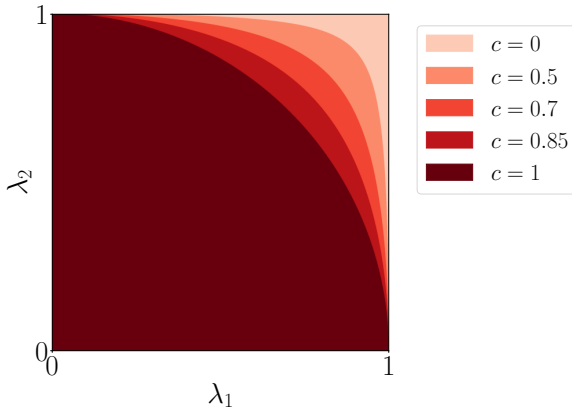

FIGURE 3: The region of impossible classification shrinks as the task correlation increases. When two tasks are uncorrelated ($c = 0$), the region of impossible classification is the whole square $[0, 1]^2$. As c increases from 0 to 1, the impossible region shrinks from the unit square $[0, 1]^2$ to a quarter of a disk.

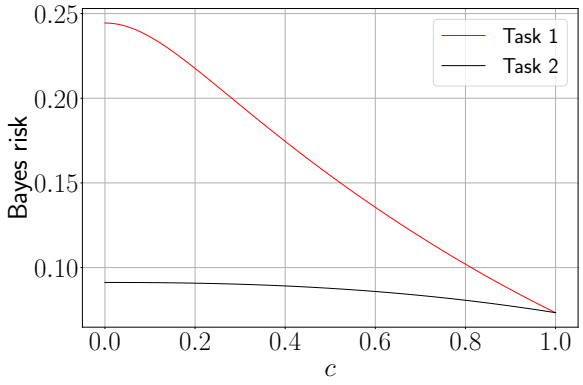

FIGURE 4: Two-task setting: Bayes risks as a function of the task correlation c, with proportions of labeled data $\eta_1 = 0$, $\eta_2 = 1$, oversampling ratios $\alpha_1 = 0.1$, $\alpha_2 = 0.2$ and SNRs $\lambda_1 = \lambda_2 = 4$. When two tasks are highly correlated ($c \gtrsim 0.5$), the performance of task 1 is significantly improved.

4.3 Semi-supervised learning.

To reduce the number of model parameters in the simulation, we here focus on a specific setting consisting of one *source task* and one *target task*. The source task is comparatively easy: it can be fully labeled, have a high SNR, or have a larger dataset. We want to see how the target task benefits from the source task.

Figure 4 illustrates the effect of task correlation. The task correlation c ranges from 0 to 1. Note that the correlations c and $-c$ are essentially the same, since one can be transformed to another by switching labels in one task. The first task (target task) is composed of a small dataset ($\alpha_1 = 0.1$) without label ($\eta_1 = 0$), while the second task (source task) consists of a fully labeled dataset ($\eta_2 = 1$) with twice as much data ($\alpha_2 = 0.2$). If two tasks are highly correlated ($c \gtrsim 0.5$), the performance of the target task can be significantly improved. When c is near zero, the decrease in Bayes risk is slow, in order of $O(c^2)$. Note that two tasks have the same SNR ($\lambda_1 = \lambda_2 = 4$), so when $c = 1$ they have the same data distribution and can be combined into a single task, yielding a identical performance.

In Figure 5, we compute the rate of error reduction in the target task as a result of transferring information from the source task. We found that MTL is most effective when the SNR of the target task is near the phase transition and is smaller than that of the source task, while the proportion of labeled data is low.

Intuitively, there are three reasons for this. Firstly, the labeled data from the target task is more valuable than that of source task, even in this case where two tasks are highly correlated ($c = 0.8$). This leads to lower gain when the proportion of labeled data in the target task is high. Secondly,

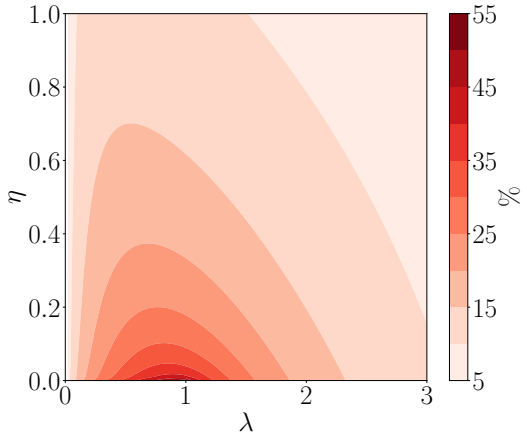

FIGURE 5: Percentage of reduction of Bayes risk in term of SNR and proportion of labeled data of the target task, with parameters $c = 0.8$, $N_1 = N_2 = D$, $\lambda_1 = 2$, $0 \leq \lambda_2 \leq 3$, $\eta_1 = 1$, $0 \leq \eta_2 \leq 1$.

if the source task is more difficult than the target task, i.e. the SNR is higher in the target task, then the source task is not very useful. Finally, near the phase transition where the target task struggles, labeled data from the source task can offer valuable help.

5 CAVITY ARGUMENT

The various equations obtained in the paper are underpinned by the phenomenon of asymptotic equivalence that occurs in the high-dimensional limit. In this limit, a fairly complicated statistical model decouples into independent components, and the inference can be performed sepa-

rately in each component. This decoupling phenomenon is proven using the so-called *cavity method*. The following lemma plays a crucial role in the cavity argument presented in this paper:

Lemma 5.1. *Suppose we want to estimate the signal $X \in \mathbb{R}$ with prior P_X from the data $\boldsymbol{Y}$ that can be split into two parts as follows. The first part, denoted by $\boldsymbol{Y}^x$, consists of the following observation on X,*

$$\boldsymbol{Y}^x = X\boldsymbol{U} + \boldsymbol{Z}, \tag{11}$$

where

- *$\boldsymbol{U} \in R^D$ is unknown with prior P_U,*
- *$\boldsymbol{Z} \sim \mathcal{N}(0, I_D)$,*
- *X, $\boldsymbol{U}$ and $\boldsymbol{Z}$ are independent.*

The second dataset, denoted by $\boldsymbol{Y}^u$, is independent of $\boldsymbol{X}$. Suppose that the law $\boldsymbol{U}|\boldsymbol{Y}^u$ has the RS property with overlap q. Then in the limit $D \to \infty$,

i) The posterior of $\boldsymbol{X}$ given $\boldsymbol{Y}$ is asymptotically equivalent to the law $\bar{P}$ defined as

$$\frac{d\bar{P}(x|\boldsymbol{Y})}{dP_X(x)} \propto \exp\Big(x\langle \boldsymbol{Y}^x, \hat{\boldsymbol{U}}\rangle - \frac{1}{2}qx^2\Big) \tag{12}$$

where $\hat{\boldsymbol{U}} = \mathbb{E}[\boldsymbol{U}|\boldsymbol{Y}]$. As a consequence, the statistics $S = \langle \boldsymbol{Y}^x, \hat{\boldsymbol{U}}\rangle$ is asymptotically sufficient for estimating $\boldsymbol{X}$ from $\boldsymbol{Y}$.

ii) $S/\sqrt{q}$ converges in law to $\sqrt{q}X + \xi$, where ξ follows standard normal distribution and is independent of X. As a result, estimating X from $\boldsymbol{Y}$ is asymptotically equivalent to estimating X from the output of a Gaussian channel with SNR q.

Proof. Since X is independent of $\boldsymbol{U}$ and $\boldsymbol{Y}^u$, we have

$$\begin{aligned}\frac{dP(x|\boldsymbol{Y})}{dP_X(x)} &= \int dP(\boldsymbol{u}|\boldsymbol{Y}^u)P(x|\boldsymbol{u}, \boldsymbol{Y}^x)\\ &\propto \int dP(\boldsymbol{u}|\boldsymbol{Y}^u)\exp\Big(x\langle \boldsymbol{Y}^x, \boldsymbol{u}\rangle - \frac{1}{2}x^2\|\boldsymbol{u}\|^2\Big)\\ &:= \mathcal{A}\end{aligned}$$

Define

$$\mathcal{B} = \exp\Big(x\langle \boldsymbol{Y}^x, \hat{\boldsymbol{U}}\rangle - \frac{1}{2}qx^2\Big) \tag{13}$$

To prove (i), we will show that $\mathbb{E}[(\mathcal{A} - \mathcal{B})^2] \to 0$ in the high-dimensional limit $D \to \infty$ for any value of x. To do this, it is sufficient to show that $\mathbb{E}[\mathcal{A}^2]$, $\mathbb{E}[\mathcal{B}^2]$ and $\mathbb{E}[\mathcal{A}\mathcal{B}]$ converge to the same limit, using the RS property of $\boldsymbol{U}|\boldsymbol{Y}^u$. Indeed, $\mathbb{E}[\mathcal{A}^2]$ can be written as

$$\mathbb{E}\exp\Big(\sum_{a=1}^{2} x\langle \boldsymbol{Y}^x, \boldsymbol{U}^a\rangle - \frac{1}{2}x^2\|\boldsymbol{U}^a\|^2\Big)$$

where $\boldsymbol{U}^1, \boldsymbol{U}^2$ are sampled independently from $\boldsymbol{U}|\boldsymbol{Y}^u$. Substituting $\boldsymbol{Y}^x = X\boldsymbol{U} + \boldsymbol{Z}$ into the previous expression, we obtain

$$\mathbb{E}\exp\Big(\sum_{a=1}^{2} xX\langle \boldsymbol{U}, \boldsymbol{U}^a\rangle + x\langle \boldsymbol{Z}, \boldsymbol{U}^a\rangle - \frac{1}{2}x^2\|\boldsymbol{U}^a\|^2\Big)$$

Taking the expectation over $\boldsymbol{Z}$ and using the fact that $\mathbb{E}[e^{\langle \boldsymbol{a}, \boldsymbol{Z}\rangle}] = e^{\frac{1}{2}\|\boldsymbol{a}\|^2}$, we have

$$\mathbb{E}[\mathcal{A}^2] = \mathbb{E}\exp\Big(\sum_{a=1}^{2} xX\langle \boldsymbol{U}, \boldsymbol{U}^a\rangle + x^2\langle \boldsymbol{U}^1, \boldsymbol{U}^2\rangle\Big)$$

It follows from RS property of $\boldsymbol{U}|\boldsymbol{Y}^u$ that

$$\lim_{D\to\infty} \mathbb{E}[\mathcal{A}^2] = \mathbb{E}\exp\big(2qXx + qx^2\big) \tag{14}$$

To calculate the limits of $\mathbb{E}[\mathcal{A}\mathcal{B}]$ and $\mathbb{E}[\mathcal{B}^2]$, we follow exactly the same procedure, which involves substituting the definition of $\boldsymbol{Y}^x$, taking the expectation over $\boldsymbol{Z}$, and using the RS property. This leads us to the same limit as (14), thereby proving (i).

It follows immediately from the asymptotic equivalence between $P(x|\boldsymbol{Y})$ and $\bar{P}(x|\boldsymbol{Y})$ that the statistics $\langle \boldsymbol{Y}^x, \hat{\boldsymbol{U}}\rangle$ is asymptotically sufficient for estimating X from $\boldsymbol{Y}$. This means that all of the relevant information about X can be extracted from $\langle \boldsymbol{Y}^x, \hat{\boldsymbol{U}}\rangle$ instead of from $\boldsymbol{Y}$, without any loss of information in high dimensional limit.

Now we have

$$\langle \boldsymbol{Y}^x, \hat{\boldsymbol{U}}\rangle = \langle X\boldsymbol{U} + \boldsymbol{Z}, \hat{\boldsymbol{U}}\rangle = X\langle \boldsymbol{U}, \hat{\boldsymbol{U}}\rangle + \langle \boldsymbol{Z}, \hat{\boldsymbol{U}}\rangle.$$

Given that $\langle \boldsymbol{Z}, \hat{\boldsymbol{U}}\rangle \sim \mathcal{N}(0, \|\hat{\boldsymbol{U}}\|^2)$ and $\boldsymbol{Z}$ is independent of X, in the limit $D \to \infty$, this inner product converges in distribution to $\sqrt{q}\xi$, where ξ is a standard normal random variable independent of $\boldsymbol{X}$. Therefore

$$\frac{\langle \boldsymbol{Y}^x, \hat{\boldsymbol{U}}\rangle}{\sqrt{q}} \xrightarrow{d} \sqrt{q}X + \xi, \quad D \to \infty,$$

which proves (ii) since the left hand side of the last expression is also a sufficient statistics of X given $\boldsymbol{Y}$. □

To give an application of Lemma 5.1 and to familiarize readers with the cavity argument before delving into the proof of the main results in the paper, we will analyze the following tensor model studied in Miolane (2017). Our goal is to estimate the signals $\boldsymbol{U}$ and $\boldsymbol{V}$ from the following observations:

$$Y_{ij} = \sqrt{\frac{\lambda}{N}}U_iV_j + Z_{ij}, i \in [N_u], j \in [N_v] \tag{15}$$

Here, we assume that $U_i \overset{i.i.d}{\sim} P_U, V_j \overset{i.i.d}{\sim} P_V$ and the noises Z_{ij} follow independent standard Gaussian distributions for

all i, j. We study the model in the limit as N, N_u, N_v tend to infinity with fixed ratios $N_u/N \to \alpha_u$ and $N_v/N \to \alpha_v$. Furthermore, we assume that $\boldsymbol{U}, \boldsymbol{V}$ and $\boldsymbol{Z} = (Z_{ij})$ are independent. It can be shown that both $N_u^{-1/2}\boldsymbol{U}|\boldsymbol{Y}$ and $N_v^{-1/2}\boldsymbol{V}|\boldsymbol{Y}$ satisfies the replica symmetry property, with overlaps q_u and q_v respectively. We will use Lemma 5.1 to derive the fixed point equations that satisfied by q_u, q_v.

Let $i \in [N_u]$ be fixed. The cavity method involves dividing the data $\boldsymbol{Y}$ into two parts. The first part, denoted as $\boldsymbol{Y}^1$, includes the observations related to U_i, given by

$$\boldsymbol{Y}_{i\cdot} = \sqrt{\frac{\lambda}{N}} U_i \boldsymbol{V} + \boldsymbol{Z}_{i\cdot} \tag{16}$$

while the remaining data is denoted as $\boldsymbol{Y}^2$. Since the dataset $\boldsymbol{Y}^1$ only contains an insignificant amount of information relevant to $\boldsymbol{V}$ (one can see that by comparing the sizes of $\boldsymbol{Y}^1$ and $\boldsymbol{Y}^2$), estimating $\boldsymbol{V}$ from $\boldsymbol{Y}$ is essentially the same as estimating $\boldsymbol{V}$ from $\boldsymbol{Y}^2$. Therefore, $N_v^{-1/2}\boldsymbol{V}|Y^2$ also satisfies the RS property with overlap q_v. It is easy to check that the Lemma 5.1 is applicable for this model, with U_i and $\sqrt{\lambda/N}\boldsymbol{V}$ respectively playing the role of X and $\boldsymbol{U}$ in the lemma. As a result, estimating U_i from $\boldsymbol{Y}$ is asymptotically equivalent to estimating the signal U_i from the output of a Gaussian channel with SNR $\lambda\alpha_v q_v$.

For distinct $i, k \in N_u$, since $\boldsymbol{Z}_{i\cdot}$ and $\boldsymbol{Z}_{k\cdot}$ are independent, it can be seen from the proof of Lemma 5.1-ii that the noises ξ_i and ξ_k of the equivalent Gaussian channels associated with U_i, U_k are independent. Therefore $\hat{\boldsymbol{U}}_i$, which depends on ξ_i and U_i, are asymptotically independent for all i. By the law of large number

$$q_u = \lim_{N_u \to \infty} \frac{1}{N_u} \sum_{i=1}^{N_u} \hat{U}_i^2 = F_U(\lambda q_v) \tag{17}$$

where F_U is the overlap function of the Gaussian channel with signal U. Repeating the same argument for V_j with $j \in N_v$, we obtain the fixed point equations for q_u, q_v:

$$q_u = F_U(\lambda\alpha_v q_v)$$
$$q_v = F_V(\lambda\alpha_u q_u)$$

Note that fixed point equations may not uniquely determine overlaps, as they can have multiple solutions. However, rigorous methods (Barbier and Macris (2019)) demonstrate that overlaps can be uniquely determined as the minimax point of a certain function.

6 PROOFS

6.1 Fixed point equations

Reformulation as a tensor model. Let $\tilde{\boldsymbol{U}}_t = \sqrt{D}\boldsymbol{U}_t$, it is shown in Appendix F that in the limit $D \to \infty$, $\tilde{U}_{tj}$ are asymptotically Gaussian with covariance

$$\mathbb{E}[\tilde{U}_{tj}\tilde{U}_{t'j'}] = C_{tt'}\delta_{jj'} \tag{18}$$

Let $\boldsymbol{W}_t = \sqrt{D}\boldsymbol{U}_t/\sigma_t$, the original model can be written as a collection of one-dimensional Gaussian channels

$$Y_{ijt} = \frac{1}{\sqrt{D}} V_{ti} W_{tj} + Z_{tij} \tag{19}$$

for $1 \le t \le T, 1 \le i \le N_t, 1 \le j \le D$. As $D \to \infty$, the random variables W_{tj} are asymptotically Gaussian with covariance

$$\mathbb{E}[W_{tj}W_{t'j'}] = M_{tt'}\delta_{jj'} \tag{20}$$

where $M_{tt'} = C_{tt'}/(\sigma_t\sigma_{t'})$.

Next, the information conveyed by the labels can be absorbed into the prior distribution of $\boldsymbol{V}$. Specifically, if the value of V_{ti} is unknown, then its prior remains uniform over $\{-1, 1\}$. Otherwise, if it is known that $V_{ti} = 1$, then the prior of V_{ti} is $\delta(v-1)$. Note that in this case, the posterior coincides with the prior.

The RS property of $\sigma_t^{-2}\boldsymbol{U}_t|\boldsymbol{Y}$ implies that $D^{-1/2}\boldsymbol{W}_t|\boldsymbol{Y}$ also has the RS property with overlap $\boldsymbol{q}_{ut}$.

In summary, the problem can be cast as a tensor model, whereby the objective is to estimate the signals $\boldsymbol{V}_t$ and $\boldsymbol{W}_t$ based on prior information regarding these vectors and noisy observations of the tensor products $\boldsymbol{V}_t \otimes \boldsymbol{W}_t$.

Cavity argument. A crucial step in the analysis is to show that in the high-dimensional limit, estimating $\boldsymbol{V}_t$ and $\boldsymbol{W}_t$ given $\boldsymbol{Y}$ is asymptotically equivalent to estimating the coordinates of these vectors from Gaussian channels with independent noises. The original model is thus equivalent to a much more decoupled model and the inference can be done separately on each channel.

To obtain the fixed point equations, we follow the same approach as the example presented in Section 5. We assume that the proportion of unlabeled data is positive in any task. By taking the limit of these proportions to zero, we can derive the result for the supervised case. Fix $t \in [T]$ and $i \in [N_t]$ such that V_{ti} is unknown. We divide the data $\boldsymbol{Y}$ into two parts: $\boldsymbol{Y}^1$ consisting of the observations concerning V_{ti}, namely

$$\boldsymbol{Y}_{ti} = \frac{1}{\sqrt{D}} V_{ti}\boldsymbol{W}_t + \boldsymbol{Z}_{ti}$$

and the remaining data $\boldsymbol{Y}^2$. Since the dataset $\boldsymbol{Y}^1$ only contains an insignificant amount of information relevant to $\boldsymbol{W}_t$, estimating $\boldsymbol{W}_t$ from $\boldsymbol{Y}$ is essentially the same as estimating $\boldsymbol{W}_t$ from $\boldsymbol{Y}^2$. Therefore, $D^{-1/2}\boldsymbol{W}_t|\boldsymbol{Y}^2$ also satisfies the RS property with overlap q_u. It is easy to check that the Lemma 5.1 is applicable, with V_{ti} and $D^{-1/2}\boldsymbol{W}_t$ respectively playing the role of X and $\boldsymbol{U}$ in the lemma. As

a result, estimating V_{ti} from $\boldsymbol{Y}$ is asymptotically equivalent to estimating the signal V_{ti} from the output of the Gaussian channel with SNR q_{ut}. For distinct $i, k \in [N_t]$, since $\boldsymbol{Z}_{ti}$ and $\boldsymbol{Z}_{tk}$ are independent, it can be seen from the proof of Lemma 5.1-ii that the noises ξ_i and ξ_k of the equivalent Gaussian channels associated with V_{ti}, V_{tk} are also independent. Therefore V_{ti}, which depends on ξ_i and V_{ti}, are asymptotically independent for all i such that V_{ti} is unlabeled. By the law of large number,

$$\begin{aligned} r_{vt} &:= \lim_{N_t \to \infty} \frac{1}{(1-\eta_t)N_t} \sum_i \hat{V}_{ti}^2 \\ &= F_v(q_{ut}) \end{aligned} \tag{21}$$

where the sum is over all $i \in [N_t]$ such that V_{ti} is unlabeled and F_v is the overlap function of the Gaussian channel with Rademacher signal. From Appendix E.1,

$$F_v(q) = \mathbb{E}[\tanh(\sqrt{q}Z + q)], \quad Z \sim \mathcal{N}(0,1). \tag{22}$$

On the other hand, from the definition of r_{vt}, we have

$$q_{vt} = \eta_t + (1-\eta_t) r_{vt} \tag{23}$$

The fixed point equation (7b) follows from (21), (22) and (23).

Following exactly the same cavity argument, the estimation of W_{tj} given $\boldsymbol{Y}$ is asymptotically equivalent the the estimation of the signal W_{tj} from the output of the Gaussian channel with SNR $\alpha_t q_{vt}$. Moreover, the noises corresponding to the signals W_{tj} and $W_{t'j'}$ are asymptotically independent for $(t,j) \neq (t',j')$. When $j \neq j'$, the signals W_{tj} and $W_{t'j'}$ are independent. As a result, the inference on the equivalent Gaussian channels can be performed independently on groups of T scalar Gaussian channels $(W_{tj})_{t=1}^T$. By the law of large number,

$$q_{ut} = \lim_{D \to \infty} \frac{1}{D} \sum_{j=1}^{D} \hat{W}_{tj}^2 = F_{w,t}(\{\alpha_t q_{vt}\}_{t=1}^T) \tag{24}$$

where $F_{w,t}$ are overlap functions of the Gaussian channel with signal $\mathcal{N}(0, \boldsymbol{M})$. The explicit formula for $F_{w,t}$ are computed in Appendix E.2, which gives the fixed point equation (7a).

6.2 Bayes risk and optimal algorithm

Suppose we want to classify a new data point $\boldsymbol{Y}_{\text{new}}$ in task t

$$\boldsymbol{Y}_{\text{new}} = V_{\text{new}} \boldsymbol{U}_t + \sigma_t \boldsymbol{Z}_{\text{new}} \tag{25}$$

It is easy to check that Lemma 5.1 can be applied to this problem, with $V_{\text{new}}, \boldsymbol{U}_t$ playing the role of $X, \boldsymbol{U}$ in the lemma, as the posterior $\sigma_t^{-1}\boldsymbol{U}_t|\boldsymbol{Y}$ satisfies the RS property with overlap q_{ut}. As a result, in high dimensional limit, estimating V_{new} given $\boldsymbol{Y}, \boldsymbol{Y}_{\text{new}}$ is essentially the same as estimating the signal V_{new} from the output of the Gaussian channel with SNR q_{ut}. This implies that the minimal classification error of V_{new} is given by that of the Gaussian channel with Rademacher signal and SNR q_{ut}, which is (Appendix E.1)

$$1 - \Phi(\sqrt{q_{ut}}),$$

According to Lemma 5.1, $S = \langle \boldsymbol{Y}_{\text{new}}, \hat{\boldsymbol{U}}_t \rangle / \sqrt{q_{ut}}$ is sufficient for estimating V_{new}. Moreover, S converges in law to the output of the Gaussian channel with signal V_{new} and SNR q_{ut}. The estimator that minimizes the Bayes risk for this channel is simply $\text{sgn}(S)$, which leads to the optimal estimator of V_{new} as $\text{sgn}(\langle \boldsymbol{Y}_{\text{new}}, \hat{\boldsymbol{U}}_t \rangle)$. The next step is to determine the value of $\hat{\boldsymbol{U}}_t$. We will take advantage of the fact that the vectors $\boldsymbol{U}_t$ are asymptotically Gaussian, so our subsequent argument will rely on the reformulation (18) of the model. We will need the following result

Lemma 6.1. *The following collection of Gaussian channels*

$$Y_i = c_i X_i + Z_i, \quad i = 1, \dots, n \tag{26}$$

with inputs X_i, outputs Y_i, SNR c_i^2 and independent standard Gaussian noises Z_i, is equivalent to a single Gaussian channel with signal X, output $\langle \boldsymbol{c}, \boldsymbol{Y} \rangle / \|\boldsymbol{c}\|$ and SNR $\sum_{i=1}^n c_i^2$. Moreover,

Proof. It is straightforward to verify that the statistics $S := \langle \boldsymbol{c}, \boldsymbol{Y} \rangle / \|\boldsymbol{c}\|$ is sufficient for estimating X from $\boldsymbol{Y}$. Moreover, $S = \|\boldsymbol{c}\| X + \xi$ where $\xi = \|\boldsymbol{c}\|^{-1} \langle \boldsymbol{c}, \boldsymbol{Z} \rangle$ is standard Gaussian and independent of X. This proves the claim of the lemma. □

Remark 6.1. From the proof of Lemma 6.1 we can also see that the noise ξ of the simplified channel comes from the noises of the original channels.

The Lemma 6.1 implies that, for each (t,j) fixed, the following Gaussian channels

$$Y_{tij} = \frac{1}{\sqrt{D}} V_{ti} W_{tj} + Z_{tij}, \quad i = 1, \dots, N_t$$

which share the same signal W_{tj}, can be simplified into a single Gaussian channel with output $\sqrt{N_t}\bar{Y}_{tj}$ and SNR $N_t/D \simeq \alpha_t$, where $\bar{Y}_{tj}$ is the j-th coordinate of the vector $\bar{\boldsymbol{Y}}_t$ in the algorithm.

For $(t,j) \neq (t',j')$, the noises of the simplified Gaussian channels associated with W_{tj} and $W_{t'j'}$ are independent, as a consequence of Remark 6.1. Additionally, the signals W_{tj} and $W_{t'j'}$ are independent if $j \neq j'$. Therefore, the inference on the simplified Gaussian channels can be carried out independently on each group of T channels with

signals $(W_{tj})_{t=1}^T$. The MMSE estimator on each of these groups can be computed explicitly as

$$(\hat{W}_{tj})_{t=1}^T = \boldsymbol{B}(\sqrt{N_t}\bar{Y}_{tj})_{t=1}^T$$

where

$$\boldsymbol{B} = \boldsymbol{M}\boldsymbol{D}_{\boldsymbol{\alpha}}^{1/2}(\boldsymbol{I} + \boldsymbol{D}_{\boldsymbol{\alpha}}^{1/2}\boldsymbol{M}\boldsymbol{D}_{\boldsymbol{\alpha}}^{1/2})^{-1}$$

(Appendix E.2). Equivalently,

$$\hat{\boldsymbol{W}}_t = \sum_s B_{ts}\sqrt{N_s}\bar{\boldsymbol{Y}}_s$$

Dividing both sides by $\sqrt{D}$ and using $N_t/D \simeq \alpha_t$, we have

$$\tilde{\boldsymbol{Y}}_t := \sigma_t^{-1}\hat{\boldsymbol{U}}_t \simeq \sum_s A_{ts}\bar{\boldsymbol{Y}}_s \tag{27}$$

where $A_{ts} = B_{ts}\sqrt{\alpha_s}$. Therefore,

$$\boldsymbol{A} = \boldsymbol{M}\boldsymbol{D}_{\boldsymbol{\alpha}}(\boldsymbol{I} + \boldsymbol{M}\boldsymbol{D}_{\boldsymbol{\alpha}})^{-1}$$

as given in the optimal algorithm. The optimal estimator for V_{new} is $\text{sgn}(\langle \boldsymbol{Y}_{\text{new}}, \hat{\boldsymbol{U}}_t\rangle) = \text{sgn}(\langle \boldsymbol{Y}_{\text{new}}, \tilde{\boldsymbol{Y}}_t\rangle)$.

7 CONCLUSION

This paper proposed a Gaussian mixture model of multi-tasking learning, in which each task is a semi-supervised classification problem. We derived an explicit formula for the Bayes risk, from which the behaviors of the model is studied through various numerical simulations.

The model in this paper concerns with Gaussian and Rademacher random variables. However, our method also works for more general tensor models with random variables of finite second moments.

Acknowledgement. We would like to thank the reviewers for their valuable feedback and insightful comments, which have significantly contributed to the improvement of this paper. We would also like to express our appreciation to Malik Tiomoko for insightful discussions on the algorithmic aspects of the model and to Hugues Souchard de Lavoreille for his internship report, which the first author consulted numerous times while working on this paper. Our research is supported by MIAI.

Bibliography

B. Aubin, F. Krzakala, Y. Lu, and L. Zdeborová. Generalization error in high-dimensional perceptrons: Approaching bayes error with convex optimization. *Advances in Neural Information Processing Systems*, 33: 12199–12210, 2020.

J. Baik, G. B. Arous, and S. Péché. Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices. *The Annals of Probability*, 33(5):1643–1697, 2005.

J. Barbier and N. Macris. The adaptive interpolation method: a simple scheme to prove replica formulas in bayesian inference. *Probability theory and related fields*, 174:1133–1185, 2019.

J. Barbier, N. Macris, and L. Miolane. The layered structure of tensor estimation and its mutual information. In *2017 55th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 1056–1063. IEEE, 2017.

J. Barbier, F. Krzakala, N. Macris, L. Miolane, and L. Zdeborová. Optimal errors and phase transitions in high-dimensional generalized linear models. *Proceedings of the National Academy of Sciences*, 116(12):5451–5460, 2019.

C. Bishop. Bayesian pca. *Advances in neural information processing systems*, 11, 1998.

F. Krzakala, M. Mézard, F. Sausset, Y. Sun, and L. Zdeborová. Statistical-physics-based reconstruction in compressed sensing. *Physical Review X*, 2(2):021005, 2012.

M. Lelarge and L. Miolane. Asymptotic bayes risk for gaussian mixture in a semi-supervised setting. In *2019 IEEE 8th International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*, pages 639–643. IEEE, 2019a.

M. Lelarge and L. Miolane. Fundamental limits of symmetric low-rank matrix estimation. *Probability Theory and Related Fields*, 173(3):859–929, 2019b.

T. Lesieur, C. De Bacco, J. Banks, F. Krzakala, C. Moore, and L. Zdeborová. Phase transitions and optimal algorithms in high-dimensional gaussian mixture clustering. In *2016 54th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 601–608. IEEE, 2016.

T. Lesieur, L. Miolane, M. Lelarge, F. Krzakala, and L. Zdeborová. Statistical and computational phase transitions in spiked tensor estimation. In *2017 IEEE International Symposium on Information Theory (ISIT)*, pages 511–515. IEEE, 2017.

B. Loureiro, G. Sicuro, C. Gerbelot, A. Pacco, F. Krzakala, and L. Zdeborová. Learning gaussian mixtures with generalized linear models: Precise asymptotics in high-dimensions. *Advances in Neural Information Processing Systems*, 34:10144–10157, 2021.

X. Mai and R. Couillet. Consistent semi-supervised graph regularization for high dimensional data. *J. Mach. Learn. Res.*, 22:94–1, 2021.

M. Mezard and A. Montanari. *Information, physics, and computation*. Oxford University Press, 2009.

F. Mignacco, F. Krzakala, Y. Lu, P. Urbani, and L. Zdeborova. The role of regularization in classification of

high-dimensional noisy gaussian mixture. In *International Conference on Machine Learning*, pages 6874–6883. PMLR, 2020.

L. Miolane. Fundamental limits of low-rank matrix estimation: the non-symmetric case. *arXiv preprint arXiv:1702.00473*, 2017.

B. R. Paredes, A. Argyriou, N. Berthouze, and M. Pontil. Exploiting unrelated tasks in multi-task learning. In *Artificial intelligence and statistics*, pages 951–959. PMLR, 2012.

N. G. Polson and S. L. Scott. Data augmentation for support vector machines. *Bayesian Analysis*, 6(1):1–23, 2011.

S. Ruder. An overview of multi-task learning in deep neural networks. *arXiv preprint arXiv:1706.05098*, 2017.

C. Thrampoulidis, S. Oymak, and M. Soltanolkotabi. Theoretical insights into multiclass classification: A high-dimensional asymptotic view. *Advances in Neural Information Processing Systems*, 33:8907–8920, 2020.

M. Tiomoko, R. Couillet, and F. Pascal. Pca-based multi task learning: a random matrix approach. *arXiv preprint arXiv:2111.00924*, 2021a.

M. Tiomoko, H. Tiomoko, and R. Couillet. Deciphering and optimizing multi-task learning: a random matrix approach. In *ICLR 2021-9th International Conference on Learning Representations*, 2021b.

M. E. Tipping and C. M. Bishop. Probabilistic principal component analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(3):611–622, 1999.

A Setting and main result

We summarize here the general setting and main results of the paper. We consider T tasks, where task t consists in classifying N_t data points in $\mathbb{R}^D$ that belong to two different Gaussian clusters with the same covariance $\sigma_t^2 I_D$. The dataset of each task is partially labeled. The model is studied in the high dimensional setting $D \to \infty$ with the following parameters supposed to be known:

- $\boldsymbol{C} = (C_{tt'})_{t,t'=1}^T$: task correlations, with $C_{tt} = 1$ for all t.
- $\alpha_t = \lim_{D\to\infty} N_t/D$: oversampling ratios
- $\lambda_t = 1/\sigma_t^2$: signal-to-noise ratios (SNRs)
- η_t: proportion of labeled data in task t

We are interested in the minimal probability of misclassifying a new data point in task t, i.e. the Bayes risk of task t.

Result. *Under the setting of the model, as $D \to \infty$, the Bayes risk of task t converges to*

$$1 - \Phi(\sqrt{q_{ut}}),$$

where $\Phi(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^t e^{-x^2} dx$ and $(q_{ut}, q_{vt})_{t=1}^T$ is the stable solution of the system of equations

$$q_{ut} = [\boldsymbol{M} - \boldsymbol{M}(\boldsymbol{I} + \boldsymbol{D}\boldsymbol{M})^{-1}]_{tt} \tag{28a}$$
$$q_{vt} = \eta_t + (1 - \eta_t)F(q_{ut}) \tag{28b}$$

with

$$\boldsymbol{M} = \{C_{tt'}/\sigma_t\sigma_{t'}\}_{t,t'=1}^T$$
$$\boldsymbol{D} = \mathrm{diag}\{\alpha_t q_{vt}\}_{t=1}^T$$
$$F(q) = \mathbb{E}[\tanh(\sqrt{q}Z + q)], \quad Z \sim \mathcal{N}(0,1).$$

B Special cases

We check the main result with the following special cases.

B.1 Uncorrelated tasks

We consider here the case in which $C_{tt'} = 0$ for all $t \neq t'$, the matrix $\boldsymbol{M}$ is diagonal and we obtain the following equations for each t

$$q_{ut} = \frac{1}{\sigma_1^2} \frac{\alpha_t q_{vt}}{\sigma_t^2 + \alpha q_{vt}}$$
$$q_{vt} = \eta_t + (1 - \eta_t)F(q_{ut})$$

which is the same as the fixed point equations when the tasks are learned separately.

B.2 The data for each task follows the same distribution

We consider here the case in which $C_{tt'} = 1$ and $\sigma_t = \sigma$ for all $t, t' = 1, \ldots, T$. We have

$$\boldsymbol{M} = \frac{1}{\sigma^2} \mathbb{1}\mathbb{1}^T, \quad \boldsymbol{D}\boldsymbol{M} = \frac{\boldsymbol{u}\mathbb{1}^T}{\sigma^2}$$

where $\boldsymbol{u} = (\alpha_t q_{vt})_{t=1}^\top$ and $\mathbb{1} = \underbrace{(1, \ldots, 1)^\top}_{T\ 1s}$. Applying the formula

$$(\boldsymbol{I} + \boldsymbol{u}\boldsymbol{v}^T)^{-1} = \boldsymbol{I} - \frac{\boldsymbol{u}\boldsymbol{v}^T}{1 + \boldsymbol{u}^T\boldsymbol{v}}$$

for $\boldsymbol{v} = \mathbb{1}/\sigma^2$, we obtain

$$(\boldsymbol{I} + \boldsymbol{DM})^{-1} = \boldsymbol{I} - \frac{\boldsymbol{u}\mathbb{1}^T}{\sigma^2 + \boldsymbol{u}^T\mathbb{1}}$$

so

$$\boldsymbol{M} - \boldsymbol{M}(\boldsymbol{I} + \boldsymbol{DM})^{-1} = \frac{1}{\sigma^2}\frac{\boldsymbol{u}^T\mathbb{1}}{\sigma^2 + \boldsymbol{u}^T\mathbb{1}}\mathbb{1}\mathbb{1}^T \tag{29}$$

It follows from the equation (28a) that for all t,

$$q_{ut} = \frac{1}{\sigma^2}\frac{\boldsymbol{u}^T\mathbb{1}}{\sigma^2 + \boldsymbol{u}^T\mathbb{1}} := q_u \tag{30}$$

Define α, η as

$$\alpha = \sum_t \alpha_t, \quad \alpha\eta = \sum_t \alpha_t\eta_t \tag{31}$$

We have

$$\begin{aligned}
\boldsymbol{u}^T\mathbb{1} &= \sum_t \alpha_t q_{ut} \\
&= \sum_t \alpha_t(\eta_t + (1-\eta_t)F(q_u)) \\
&= \alpha\eta + \alpha(1-\eta)F(q_u) \\
&= \alpha q_v
\end{aligned} \tag{32}$$

where q_v is defined as

$$q_v = \eta + (1-\eta)F(q_u) \tag{33}$$

then from (30) and (32), we have

$$q_u = \frac{1}{\sigma^2}\frac{\alpha q_v}{\sigma^2 + \alpha q_v} \tag{34}$$

Since (33) and (34) are exactly the equations for the case of single task learning with parameters α and η, the multitask learning problem is reduced to one single task with parameters α, η given by (31).

C Unsupervised learning and phase transition

C.1 Region of impossible recovery

In the unsupervised case, the fixed point equations are

$$q_{ut} = [\boldsymbol{M} - \boldsymbol{M}(\boldsymbol{I} + \boldsymbol{DM})^{-1}]_{tt} \tag{35a}$$
$$q_{vt} = F(q_{ut}) \tag{35b}$$

which always admits $(\boldsymbol{q}_u, \boldsymbol{q}_v) = (\boldsymbol{0}, \boldsymbol{0})$ as solution. The classification is impossible if and only if this solution is stable. To analyze the stability of (35) around zero, let $q_{ut}, q_{vt} = O(h)$ where $h \to 0$. For vectors A and B of the same dimension, we denote $A \simeq B$ if $|A - B| \simeq O(h^2)$, where $|\,.\,|$ denotes the Euclidean norm. From

$$F(q) = \mathbb{E}[\tanh(\sqrt{q}Z + q)], \tag{36}$$

(Appendix E.1), using the Taylor expansion $\tanh(x) = x - x^3/3 + o(x^3)$, we get

$$q_{vt} = F(q_{ut}) \simeq q_{ut}$$

On the other hand,

$$\begin{aligned} q_{ut} &= [\boldsymbol{M} - \boldsymbol{M}(\boldsymbol{I} + \boldsymbol{D}\boldsymbol{M})^{-1}]_{tt} \\ &\simeq [\boldsymbol{M} - \boldsymbol{M}(\boldsymbol{I} - \boldsymbol{D}\boldsymbol{M})]_{tt} \\ &= [\boldsymbol{M}\boldsymbol{D}\boldsymbol{M}]_{tt} \\ &= \sum_{s=1}^{T} M_{ts}^2 \alpha_s q_{vs} \end{aligned}$$

Let

$$\boldsymbol{P} = (M_{ts}^2 \alpha_s)_{s,t=1}^T = \Big(\frac{C_{ts}^2}{\sigma_t^2 \sigma_s^2}\alpha_s\Big)_{s,t=1}^T = (\lambda_s \lambda_t C_{st}^2 \alpha_s)_{s,t=1}^T \tag{37}$$

In a small neighborhood of $(\mathbf{0}, \mathbf{0})$, the system of equations can be approximated up to an error of $O(h^2)$ by

$$\boldsymbol{q}_v = \boldsymbol{q}_u \tag{38}$$

$$\boldsymbol{q}_u = \boldsymbol{P}\boldsymbol{q}_v \tag{39}$$

Therefore the fixed point $(\mathbf{0}, \mathbf{0})$ is stable if and only if the module of each eigenvalue of $\boldsymbol{P}$ is not larger than 1. Using the property that AB and BA has the same eigenvalues for general square matrices A, B, the matrix $\boldsymbol{P}$ has the same eigenvalues as the following symmetric matrix

$$\boldsymbol{R} = (\sqrt{\alpha_s \alpha_t}\lambda_s \lambda_t C_{st}^2)_{s,t=1}^T \tag{40}$$

Note that $\boldsymbol{R}$ is a positive semidefinite (p.s.d) matrix, since it can be written as Hadamard product of p.s.d. matrices. Therefore, the classification is impossible if and only if all eigenvalues of $\boldsymbol{R}$ are not greater than 1.

When $C_{tt'} = c$ for all $t \neq t'$ and $\lambda_t = \lambda, \alpha_t = 1$ for all t, we have

$$\boldsymbol{R} = \lambda^2(c^2 \mathbb{1}\mathbb{1}^T + (1 - c^2)I) \tag{41}$$

Note that the matrix $\mathbb{1}\mathbb{1}^T$ has eigenvalues $0, \ldots, 0, T$, so the largest eigenvalue of $\boldsymbol{R}$ is $\lambda^2(1 + (T-1)c^2)$, from with we obtain the condition for impossible classification

$$\lambda^2(1 + (T-1)c^2) \leq 1 \tag{42}$$

which becomes $\lambda \leq 1$ for the special case $T = 1$.

When $T = 2$ with task correlation c and $\alpha_1 = \alpha_2 = 1$, we have

$$\boldsymbol{R} = \begin{pmatrix} \lambda_1^2 & c^2 \lambda_1 \lambda_2 \\ c^2 \lambda_1 \lambda_2 & \lambda_2^2 \end{pmatrix} \tag{43}$$

It is clear that the (λ_1, λ_2)-domain of impossible classification is a subset of $[0, 1]^2$, otherwise at least one task is achievable. All eigenvalues of $\boldsymbol{R}$ are less than 1 if and only if $\mathrm{Tr}(\boldsymbol{I} - \boldsymbol{R}) \geq 0$ and $\det(\boldsymbol{I} - \boldsymbol{R}) \geq 0$. The first condition is already satisfied for $(\lambda_1, \lambda_2) \in [0, 1]^2$ while the second condition is equivalent to

$$(1 - \lambda_1^2)(1 - \lambda_2^2) \leq c^4 \lambda_1^2 \lambda_2^2 \tag{44}$$

C.2 Connected tasks are either all feasible or impossible

In the unsupervised case, tasks are considered connected if any two tasks are directly or indirectly correlated through other tasks. We will prove that if tasks are connected, then either all tasks are feasible or all tasks are impossible. As a reminder, for any task t, the value of q_{ut} is always non-negative. If $q_{ut} = 0$, then the task t is impossible; otherwise, it is feasible.

Consider T Gaussian channels with outputs $(Y_t)_{t=1}^T$, signals $(X_t)_{t=1}^T$ having joint distribution $\mathcal{N}(0, \boldsymbol{M})$ and independent standard Gaussian noises. The SNRs for each channel are $(\alpha_t q_{vt})_{t=1}^T$. Then the right-hand side of (35a) corresponds to the overlap between the signal X_t and its MMSE estimator (Appendix E.2).

Suppose by contradiction that the tasks can be split into non-empty sets such S and S' such that $q_{ut} = 0$ for all $t \in S$ while $q_{ut} > 0$ for all $t \in S'$. Since the tasks are connected, there exists correlated tasks t, t' such that $t \in S, t' \in S'$. Therefore, there exists t, t' such that $q_{ut} = 0, q_{ut'} > 0$ and $C_{tt'} \neq 0$.

Since $\mathbb{E}[X_t X_{t'}] = M_{tt'} = C_{tt'}/(\sigma_t \sigma_{t'}) \neq 0$, X_t is correlated with $X_{t'}$. Moreover, as $q_{vt'} = F(q_{ut'})$ and $q_{ut'} > 0$, we have $q_{vt'} > 0$. Therefore X_t is not independent of $\boldsymbol{Y} = \left\{\sqrt{\alpha_s q_{vs}} X_s + Z_s\right\}_{s=1}^T$, leading to $q_{ut} = \mathbb{E}[X_t \mathbb{E}[X_t|\boldsymbol{Y}]] > 0$, a contradiction.

D Estimating model parameters from data

Although it is assumed that the model parameters $\boldsymbol{C}$ and (σ_t) are available for the analysis, we show here that they can indeed be estimated with vanishing errors as $D \to \infty$, given that a positive fraction of labeled data is available in each task, i.e. $\eta_t > 0$ for all t. First consider the supervised learning case. Let

$$\bar{\boldsymbol{Y}}_t = \frac{1}{N_t} \sum_{i=1}^{N_t} V_{ti} \boldsymbol{Y}_{ti} \tag{45}$$

Then we have

$$\bar{\boldsymbol{Y}}_t = \boldsymbol{U}_t + \sqrt{\frac{\sigma_t^2}{N_t}} \bar{\boldsymbol{Z}}_t \tag{46}$$

where

$$\bar{\boldsymbol{Z}}_t = \frac{1}{\sqrt{N_t}} \sum_{i=1}^{N_t} V_{ti} \boldsymbol{Z}_{ti} \tag{47}$$

It is clear that $\bar{\boldsymbol{Z}}_t \overset{i.i.d}{\sim} \mathcal{N}(0, I_D)$ for $t = 1, \ldots, T$. We consider the following estimator of $C_{tt'}$ for $t \neq t'$:

$$\hat{C}_{tt'} = \langle \bar{\boldsymbol{Y}}_t, \bar{\boldsymbol{Y}}_{t'} \rangle \tag{48}$$

Insert (46) into the definition of $\hat{C}_{tt'}$ and use the fact that $\langle \bar{\boldsymbol{Z}}_t, \bar{\boldsymbol{Z}}_{t'} \rangle = O(\sqrt{D})$, $\langle \bar{\boldsymbol{U}}_t, \bar{\boldsymbol{Z}}_{t'} \rangle = O(1)$, which are direct consequences of Central Limit Theorem, we obtain $\hat{C}_{tt'} = C_{tt'} + O(D^{-1/2})$. Moreover

$$\|\bar{\boldsymbol{Y}}_t\|^2 = 1 + \frac{\sigma_t^2}{\alpha_t} + O(D^{-1/2}), \tag{49}$$

from which σ_t can also be estimated.

In the case where the proportion of labeled data is positive for all tasks, we can restrict the above estimators on the labeled data and obtain the approximate values of $\boldsymbol{C}$ and (σ_t) with errors converging to zero when $D \to \infty$.

E Simple Gaussian channels

E.1 Rademacher signal.

Consider the Gaussian channel given by

$$Y = \sqrt{\lambda} X + Z, \tag{50}$$

where the Rademacher signal X takes values of 1 and -1 with equal probabilities and the standard Gaussian noise Z is independent of X. We have

$$\begin{aligned} P(x|Y) &= \frac{P(x)P(Y|x)}{P(Y)} \\ &\propto e^{-(Y - \sqrt{\lambda}x)^2/2} \\ &\propto e^{\sqrt{\lambda} Y x}, \end{aligned} \tag{51}$$

from which we obtain the posterior distribution as

$$P(x|Y) = \frac{e^{\sqrt{\lambda}Yx}}{2\cosh(\sqrt{\lambda}Y)} \tag{52}$$

and the MMSE estimator $\hat{X}_{\text{MMSE}} = \mathbb{E}[X|Y]$ as

$$\hat{X} = \sum_{x=\pm 1} xP(x|Y) = \tanh(\sqrt{\lambda}Y). \tag{53}$$

The overlap between the MMSE estimator and the signal is therefore

$$\begin{aligned}
\mathbb{E}[X\hat{X}_{\text{MMSE}}] &= \mathbb{E}[X\tanh(\sqrt{\lambda}(\sqrt{\lambda}X + Z)] \\
&= \frac{1}{2}\mathbb{E}[\tanh(\lambda + \sqrt{\lambda}Z)] - \frac{1}{2}\mathbb{E}[\tanh(-\lambda + \sqrt{\lambda}Z)] \\
&= \frac{1}{2}\mathbb{E}[\tanh(\lambda + \sqrt{\lambda}Z)] - \frac{1}{2}\mathbb{E}[\tanh(-\lambda - \sqrt{\lambda}Z)] \\
&= \mathbb{E}[\tanh(\sqrt{\lambda}Z + \lambda)]
\end{aligned} \tag{54}$$

Next, the error $\mathbb{P}(\hat{X} \neq X)$ for any estimator $\hat{X}$ of X is minimized by the maximum-likelihood estimator:

$$\begin{aligned}
\hat{X}_{\text{ML}} &= \operatorname*{argmax}_{x=\pm 1} P(x, Y) \\
&= \operatorname*{argmax}_{x=\pm 1} e^{\sqrt{\lambda}Yx}
\end{aligned} \tag{55}$$

This gives us the maximum-likelihood estimator as:

$$\hat{X}_{\text{ML}} = \text{sgn}(Y). \tag{56}$$

The Bayes risk is therefore

$$\begin{aligned}
\mathbb{P}(X \neq \hat{X}_{ML}) &= \frac{1}{2}\mathbb{P}(X = 1, \hat{X}_{ML} = -1) + \frac{1}{2}\mathbb{P}(X = -1, \hat{X}_{ML} = 1) \\
&= \frac{1}{2}\mathbb{P}(X = 1, Y < 0) + \frac{1}{2}\mathbb{P}(X = -1, Y > 0) \\
&= \mathbb{P}(X = -1, Y > 0) \\
&= \mathbb{P}(Z > \sqrt{\lambda})
\end{aligned}$$

E.2 Correlated Gaussian signals

Consider T Gaussian channels, where the signals $X_1, \ldots, X_T$ have a joint distribution of $\mathcal{N}(0, \boldsymbol{M})$ and are independent of Gaussian noises $Z_1, \ldots, Z_T$ that are independently distributed as $\mathcal{N}(0, 1)$. Specifically, we have:

$$Y_t = \sqrt{\lambda_t}X_t + Z_t, \quad t = 1, \ldots, T.$$

Let $\hat{X}_t = \mathbb{E}[X|\boldsymbol{Y}]$ be the MMSE estimator for X_t. Since $(X_t, Y_1, \ldots, Y_T)$ is a Gaussian vector, $\hat{X}_t$ is a linear combination of $Y_1, \ldots, Y_T$. Therefore

$$\begin{aligned}
\text{MMSE}_t &:= \mathbb{E}[(X_t - \hat{X}_t)^2] \\
&= \min_{\boldsymbol{\beta}_t \in \mathbb{R}^T} \mathbb{E}\big[(X_t - \langle \boldsymbol{\beta}_t, \boldsymbol{Y}\rangle)^2\big].
\end{aligned}$$

This can be written as a quadratic optimization problem

$$\text{MMSE}_t = \min_{\boldsymbol{\beta}_t \in \mathbb{R}^T} \left\{ M_{tt} - 2\boldsymbol{a}_t^T \boldsymbol{\beta}_t + \boldsymbol{\beta}_t^T \boldsymbol{A} \boldsymbol{\beta}_t \right\}$$

with

$$\boldsymbol{a}_t = (\mathbb{E}[X_t Y_s])_{s=1}^T = \left(\sqrt{\lambda_t} M_{ts} \right)_{s=1}^T = \boldsymbol{D}_{\boldsymbol{\lambda}}^{1/2} \boldsymbol{M} \boldsymbol{e}_t$$
$$\boldsymbol{A} = (\mathbb{E}[Y_s Y_{s'}])_{s,s'=1}^T = \left(\sqrt{\lambda_s \lambda_{s'}} M_{ss'} + \delta_{ss'} \right)_{s,s'=1}^T = \boldsymbol{I} + \boldsymbol{D}_{\boldsymbol{\lambda}}^{1/2} \boldsymbol{M} \boldsymbol{D}_{\boldsymbol{\lambda}}^{1/2}.$$

This optimization problem admits a unique minimizer $\boldsymbol{\beta}_t = \boldsymbol{A}^{-1} \boldsymbol{a}_t$, from which we obtain

$$\hat{\boldsymbol{X}} = \boldsymbol{M} \boldsymbol{D}_{\boldsymbol{\lambda}}^{1/2} (\boldsymbol{I} + \boldsymbol{D}_{\boldsymbol{\lambda}}^{1/2} \boldsymbol{M} \boldsymbol{D}_{\boldsymbol{\lambda}}^{1/2})^{-1} \boldsymbol{Y} \tag{57}$$
$$\text{MMSE}_t = [\boldsymbol{M} (\boldsymbol{I} + \boldsymbol{D}_{\boldsymbol{\lambda}} \boldsymbol{M})^{-1}]_{tt} \tag{58}$$
$$\mathbb{E}[X_t \hat{X}_t] = [\boldsymbol{M} - \boldsymbol{M} (\boldsymbol{I} + \boldsymbol{D}_{\boldsymbol{\lambda}} \boldsymbol{M})^{-1}]_{tt}. \tag{59}$$

F The uniform prior is asymptotically Gaussian

To generate $(\boldsymbol{U}_1, \ldots, \boldsymbol{U}_T)$ according to the prior distribution specified in the model, we follow these steps:

1. Generate $\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_T \overset{i.i.d}{\sim} \mathcal{N}(\boldsymbol{0}, I_D)$.
2. Orthonormalize $\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_T$ using Gram-Schmidt process, we obtain orthonormal vectors $\boldsymbol{S}_1, \ldots, \boldsymbol{S}_T$
3. $(\boldsymbol{U}_1, \ldots, \boldsymbol{U}_T) = (\boldsymbol{S}_1, \ldots, \boldsymbol{S}_T) \boldsymbol{C}^{1/2}$, where $(\boldsymbol{U}_1, \ldots, \boldsymbol{U}_T)$ denotes the $D \times T$ matrix with columns $\boldsymbol{U}_1, \ldots, \boldsymbol{U}_T$.

In the high dimensional limit, the vector $\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_T$ are asymptotically orthogonal, so the orthonormalizing step produces approximately $n^{-1/2}(\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_T)$, which implies that if $\boldsymbol{W}_t = \sqrt{D} \boldsymbol{U}_t$, then $\boldsymbol{W}_t$'s are asymptotically Gaussian with covariance

$$\mathbb{E}[W_{ti} W_{t'j}] = \delta_{ij} C_{tt'} \tag{60}$$

It is worth noting that this is a direct consequence of the equivalence between the canonical and microcanonical ensembles in statistical physics.