Master of Science Program
in Physics of Complex Systems

Learning Generalized Linear Models with Superstatistical Covariates

Author:

Leonardo Defilippis

Supervisor:

Bruno Loureiro
DIENS, École Normale Superieure de Paris,
CNRS

Co-Supervisor:

Alfredo Braunstein
DISAT, Politecnico di Torino

Abstract

While machine learning is making significant advances in recent years, the problem of its theoretical understanding still remains an open challenge. One key aspect is the ability to predict the generalization of learning algorithms' predictions, which is crucial for assessing their reliability in various domains such as medicine, biology, finance, and signal processing. Previous studies in supervised learning have considered, in the vast majority of cases, Gaussian distributed covariates. However, in practical applications of machine learning, the data distribution may diverge from Gaussianity in many ways, such as fluctuations, heavy-tails or structured patterns. This work aim to investigate, employing the heuristic replica method from statistical physics, the supervised learning of generalized linear models when the covariates are distributed according to a superstatistical model, meaning that each covariate is drawn from a Gaussian distribution with random covariance following a generic probability distribution ρ. The regime of our interest is the one of finite sample complexity, which is the ratio of sample size with respect to the covariates' size, with both of them taken infinitely large. The choice of ρ can affect drastically the resulting covariates' distribution, which may present heavy-tails or even infinite variance. In particular we derive equations to predict the minimal estimation error that is achievable by any algorithm given the data, studying the *Bayes optimal* setting for this problem. We compare these results to the ones of empirical risk minimization. We then compute the leading order of the estimation error curves with respect to the sample complexity, showing that it does not depend on the choice of ρ and it is compatible with the Gaussian covariates' case. Our findings align with the Gaussian universality principle, which has been proven rigorously for several problems, stating that non-Gaussian distributed data can be effectively described by Gaussian distributions with matching first two moments.

Contents

1 Introduction

In recent years, machine learning and artificial intelligence have made significant advances, leading to remarkable achievements in a wide range of applications. However, the theoretical understanding of the performance of such algorithms remains an open challenge, especially in high-dimensional contexts involving large datasets and a high number of parameters. Establishing the reliability and improving the implementation of these methods requires a deep understanding of their theoretical limits. Over the last few decades, many aspects of inference and learning problems have been investigated, including regression, classification and matrix factorization. While exact results exist for some inference and learning problems [43, 12], part of the research in this field rely on heuristic tools from statistical physics, such as the replica method [32, 38]. Despite its non-rigorous nature, the replica method has demonstrated its reliability whenever analogous exact results are available [6, 7, 18].
This work follows this research direction, with a particular focus on generalized linear models (GLMs) [36, 30]. Introduced as extensions of linear models, GLMs offer a simple formulation and high versatility, making them applicable in various fields such as statistics, communication, signal processing and more [47, 13, 8, 44]. In GLMs, given data composed of predictive features, referred to as covariates, and response variables, known as labels, each label is obtained as a scalar activation function (which can be nonlinear) of a linear combination of the covariates. Learning a GLM involves estimating the weights of this linear combination used to generate the labels. GLM can be also seen as a single node of a neural network. An illustrative representation of GLM can be found in fig. 1.

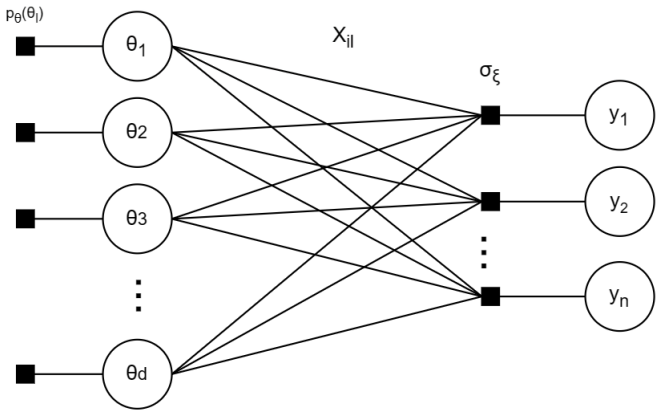

Figure 1: Illustration of a GLM. The covariates are $\mathbf{x}_i = (X_{i1}, \dots, X_{id})^\top$ and the labels y_i $i \in [n]$, the weights are θ_l, $l \in [d]$ (drawn from p_θ) and the activation function is σ_ξ.

A crucial assumption in many previous studies has been that the covariates are distributed according to a Gaussian, or at least a mixture of Gaussians [33, 27, 4, 29], with few exceptions [1, 40]. However, in real-world data, we often observe structural patterns, outliers, non-Gaussian fluctuations, heavy-tailed distributions and in general deviations from Gaussianity [2, 42]. Therefore, it is important to study GLMs under non-Gaussian assumptions.
In this work, we employ the replica method to provide asymptotic results for the estimation error in learning GLMs with linear and probit activation functions, relaxing the assumption of Gaussian covariates. The framework we develop may be useful to extend our results to

different choices of activation functions. Specifically, we assume that the covariates follow a *superstatistical* model [10, 9], inspired by the literature in statistical physics. In this model, the features are drawn from a Gaussian distribution with a covariance matrix equal to $\Delta \mathbb{I}_d/d$, where Δ itself is a random variable. By appropriately choosing the distribution of Δ (e.g., an inverse gamma distribution in our work), the covariates' distribution can exhibit heavy-tailed behavior or even infinite variance. We are interested in the regime of proportionality between the sample size and the covariates dimension, when both of them are infinitely large.

1.1 Motivations and related works

Our work was inspired by a recent paper [1] studying the task of classificating superstatistical features and, similarly to it, finds its motivation in the context of *Gaussian universality*. This principle implies that in numerous problem, the asymptotic performance of learning non-Gaussian data can be effectively described by Gaussian distributed data as long as the first two moments match. A recent work by Montanari and Saed [35], extending a previous study [20], proved such a principle in the context of GLMs, assuming pointwise normality of the distribution of the features (see also [16, 14]). Moreover, this universality principle has been proven rigorously in the context of compressed sensing [34] and lasso regression [39]. Other related works can be found in [31, 19]. However these proofs require specific assumption and if they are not satisfied the Gaussian universality principle may break. Therefore we aim to explore the validity of this principle for GLMs under the assumption of a superstatistical model for the covariates, which may include distributions with heavy-tails or even infinite covariance.

1.2 Overview and our contributions

We provide an overview of the present research work. We study for the first time the problem of supervised learning (*i.e.* learning data with labels) given covariates that follow a superstatistical distribution and labels generated by a GLM. This problem is tackled by employing the replica method from statistical physics.We first compute the best achievable estimation error for a given dataset through *Bayes optimal* estimation. At a later stage we compare this optimal performance with usual estimation approaches like *empirical risk minimization* (ERM) methods, in particular when the latters are optimized. At last, to validate our results, we show the agreement between our predictions for the error and numerical exeriments on synthetic data. The main contibutions of our work are the following ones:

- we study the task of learning a generalized linear model with features distributed according the aforementioned superstatistical model, aiming to evaluate its asymptotic performance, in a general framework;
- we focus on computing the theoretical limits of this task, *i.e.* the best performance any algorithm can achieve given the data, by studying the problem in the *Bayes optimal* setting (see [46], [24] as examples); in particular we find explicit results for the choices of linear and probit activation function;
- we evaluate the performance of the estimation by empirical risk minimization (ERM) in particular for ridge and lasso regression;
- we evaluate the decay rates[1] of all the learning curves for large sample complexity (number

[1] By decay rate we mean the leading order of the estimation error curves with respect to the sample complexity.

of samples divided by number of features) and show they match the case with Gaussian covariates;

- we optimize the regression parameter λ in ridge regression and show that this achieve the same performance as Bayes optimal evaluation even for non-Gaussian covariates.

This manuscript is organized as follows: in section 2 we define the setting of the problem, the task to achieve and outline the main points of the methodology applied to obtain the results; in section 3 we present the main results of this work, both the equations that allow to compute the theoretical performance of the learning algorithms and the numerical solution of these equation, presenting the plot of the learning curves; we then compare the results of Bayes optimal estimation to the ones of optimized ERM; in section 4 we compare our results to numerical experiments, using in particular the tool of Generalized Approximate Message Passing (GAMP) algorithms as a way to perform Bayes optimal estimation. Finally in the appendices we present all the detailed computations that are not included in the main sections, along with some technicalities on replica symmetric matrix algebra, Bayes optimal setting and its implications and the choice of the inverse gamma distribution for the superstatistical model.

2 Problem setting

The data

Consider the supervised learning problem with training data $\mathcal{D} = \left\{(\mathbf{x}_i, y_i) \in \mathbb{R}^d \times \mathcal{Y} : i \in [n]\right\}$ and with covariates $\mathbf{x}_i$ indipendently and identically drawn from a superstatistics model:

$$P(\mathbf{x}) = \int_0^\infty \mathcal{N}\left(\mathbf{x} \mid \boldsymbol{\mu},\ \frac{\Delta}{d}\mathbb{I}_d\right) \rho(\Delta)\, \mathrm{d}\Delta = \mathbb{E}_\Delta\left[\mathcal{N}\left(\mathbf{x} \mid \boldsymbol{\mu},\ \frac{\Delta}{d}\mathbb{I}_d\right)\right] \tag{1}$$

where $\mathbb{I}_d$ is the identity matrix of size d, $\mathcal{N}(\mathbf{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma})$ is the notation for the Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ and $\rho : \mathbb{R}^+ \to \mathbb{R}$ is a probability density function for Δ. This type of model has been intensively studied by the statistical physics community [10, 9, 25, 45] and, depending on the choice for $\rho(\Delta)$, the distribution $P(\mathbf{x})$ can have significantly different properties from those of a Gaussian distribution. While the derivation of our results will keep $\rho(\Delta)$ generic, for the numerical simulations presented in this report, we will consider Δ following the *inverse Gamma* distribution $\rho(x \mid a, b) = b^a(1/x)^{a+1}\exp(-b/x)/\Gamma(a)$, where $\Gamma(\cdot)$ denotes the Gamma function.[2] We include further details about the choice of the inverse Gamma distribution for $\rho(\Delta)$ in appendix D, where we also show in (40) that it implies power-law tails for $P(x)$

$$P(\mathbf{x}) \propto \left(2b + ||\sqrt{d}(\mathbf{x} - \boldsymbol{\mu})||^2\right)^{-a-\frac{d}{2}}.$$

By varying the shape parameter a and the scale parameter b, we can consider different non-Gaussian distribution $P(\mathbf{x})$, with heavy-tails or even infinite covariance. In particular, as we also explain in D, we will consider the case $\rho_>(\Delta \mid a) := \rho(\Delta \mid a > 1, b = a - 1)$ with finite $\overline{\Delta} := \mathbb{E}_\Delta[\Delta] = 1$, $\forall a > 1$, and $\rho_<(\Delta \mid a) := \rho(\Delta \mid a \in (0, 1], b = 1)$ with the first moment $\mathbb{E}_\Delta[\Delta] = +\infty$ not defined for $a \in (0, 1]$. In fig. 2 we show some examples of $P(x)$ for $d = 1$.

The labels y_i are generated from a generalized linear model (GLM)

$$y_i = \sigma_\xi\left(\boldsymbol{\theta}_*^\top \mathbf{x}_i\right), \quad i \in [n],$$

[2] $\Gamma(z) = \int_0^\infty t^{z-1}e^{-t}dt$, $\mathrm{Re}[z] > 0$

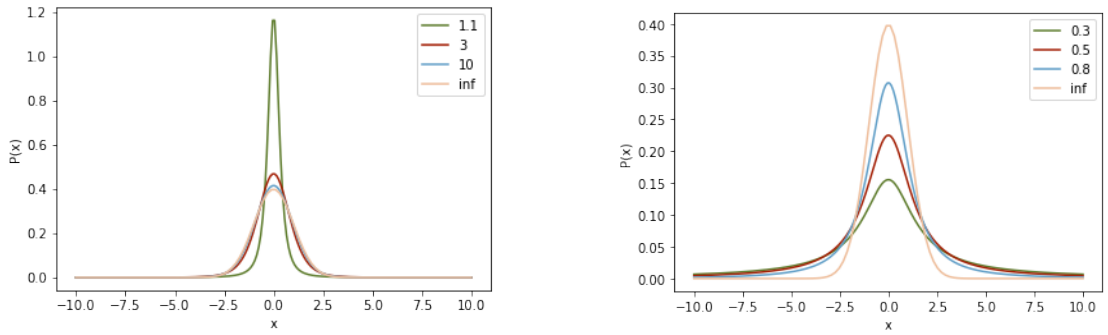

Figure 2: $P(x)$ for $d = 1$. The left plot is obtained considering $\rho_{>}(\Delta \mid a)$, while the right plot corresponds to the choice $\rho_{<}(\Delta \mid a)$. Labels in the legend indicate the value of a, in particular the label $a = \inf$ (meaning $a \to \infty$) correspond to a Gaussian $P(x) = \mathcal{N}(x \mid 0, 1)$.

where the activation function $\sigma_\xi : \mathbb{R} \to \mathbb{R}$ may contain randomness, here parametrized by the noise $\xi \sim p_\xi(\xi)$, $p_\xi : D_\xi \subseteq \mathbb{R} \to \mathbb{R}$, and $\boldsymbol{\theta}_* \in \mathbb{R}^d$ is a vector of indipendent and identically distributed weights $\theta_{*,\,l} \sim p^*_\theta(\theta)$ ("prior" distribution), $l \in [d]$. The labels vector $\mathbf{y} = (y_1, \dots, y_n)^\top$ is hence drawn from the conditioned probability distribution (called output channel or likelihood)

$$P^*_{Y|X}(\mathbf{y} \mid \mathbf{X}\boldsymbol{\theta}_*) = \prod_{i\in[n]} p^*_{out}(y_i \mid \boldsymbol{\theta}_*^\top \mathbf{x}_i) = \prod_{i\in[n]} \int_{D_\xi} p_\xi(\xi)\, \delta(y_i - \sigma_\xi(\boldsymbol{\theta}_*^\top \mathbf{x}_i))\, \mathrm{d}\xi$$

with $\mathbf{X} \in \mathbb{R}^{n\times d}$ the matrix with the covariate $\mathbf{x}_i$ as i^{th} row ($i \in [n]$) and $\delta(\cdot)$ the Dirac's delta distribution.

The task

The aim is to learn the model by reconstructing the weights vector through an estimator $\hat{\boldsymbol{\theta}}$ in the limit $d, n \to \infty$, with fixed *sample complexity* $n/d = \alpha \in \mathbb{R}$, assuming that the weights θ_l ($l \in [d]$) and the labels y_i ($i \in [n]$) are respectively generated from some distributions $p_\theta(\theta_l)$ and $p_{out}(y_i \mid \boldsymbol{\theta}^\top \mathbf{x}_i)$. This scenario is often referred to as *teacher-student* setting, where a *teacher* generates the labeled data $\mathcal{D}$, drawing the weights from a target distribution p^*_θ, called (teacher) *prior* in this manuscript, and producing the label through an output channel p^*_{out} (these two are referred to as teacher distributions), while a *student* tries to learn the teacher model using the data and assuming the prior to be p_θ and the output channel p_{out} (which are also called strudent distributions).
Our choice for the estimation error is the mean square error with respect to the true weights $\varepsilon_{\text{est}}(\hat{\boldsymbol{\theta}}) := d^{-1}||\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_*||^2$, which is the way we are going to evaluate the performance of differents methods for estimating the weights $\boldsymbol{\theta}$. The main results shown in this work specifically concern the *Bayes optimal* setting, in which the (student) prior distribution p_θ and the output channel p_{out} used in the reconstruction coincide with the (teacher) ones that actually generated the true weights and the labels, respectively p^*_θ and p^*_{out}. In this setting, the optimal estimator for the weights in the MSE sense, *i.e.* the one that minimize the square error, is the Minimal Mean Square Error (MMSE) estimator

$$\hat{\boldsymbol{\theta}}(\mathcal{D}) := \arg\min_{\boldsymbol{\theta}} \mathbb{E}_{\boldsymbol{\theta}_*|\mathcal{D}} ||\boldsymbol{\theta} - \boldsymbol{\theta}_*||^2 = \mathbb{E}_{\boldsymbol{\theta}_*|\mathcal{D}} [\boldsymbol{\theta}_*]\,, \tag{2}$$

where the latter is the expected value with respect to the posterior distribution

$$p(\boldsymbol{\theta} \mid \mathcal{D}) = \frac{1}{Z(\mathcal{D})} \prod_{l\in[d]} p_\theta(\theta_l) \prod_{i\in[n]} p_{out}\left(y_i \mid \boldsymbol{\theta}^\top \mathbf{x}_i\right). \tag{3}$$

In general cases with incomplete information (which means absence of correspondence between teacher and student distrubutions), the estimator defined in the last equality of (2) is called *mean posterior* estimator (MP), but it is not optimal unless teacher and student distributions coincide (see appendix C.1). The study of the Bayes optimal setting allows us to compute the theoretical limits of the learning task for a fixed teacher. In comparison, any other estimator would achieve either worse or, at best, the same performance.
The posterior distribution (3) mantains its relevance also when an estimator different from the MPE is used, since other estimators are linked to it. Some examples are the *maximum a posteriori* estimator $\hat{\boldsymbol{\theta}}_{MAP} = \arg\max_{\boldsymbol{\theta}} p(\boldsymbol{\theta} \mid \mathcal{D})$ and the *minimal mean absolute* estimator $\hat{\boldsymbol{\theta}}_{MMA} = \text{Median}_{\theta|\mathcal{D}}\left(\boldsymbol{\theta} \mid \mathcal{D}\right)$.
Another common procedure to reconstruct the weights vector is the *empirical risk minimization* (ERM), whose estimator is

$$\hat{\boldsymbol{\theta}}(\mathcal{D}) = \arg\min_{\boldsymbol{\theta}} \sum_{i\in[n]} \ell(y_i, \boldsymbol{\theta}^\top \mathbf{x}_i) + \sum_{l\in[d]} \lambda r(\theta_l). \tag{4}$$

The functions $\ell : \mathbb{R}^2 \to \mathbb{R}$ and $r : \mathbb{R} \to \mathbb{R}$ are respectively called *loss* and *regularization* and their expression can be chosen depending on the specific problem considered. The argument of the minimization in (4) is also called *empirical risk*, hence the name of this procedure.
We introduce the following *Gibbs measure* over the weights $\boldsymbol{\theta}$:

$$\mu_\beta(\boldsymbol{\theta} \mid \mathcal{D})\mathrm{d}\boldsymbol{\theta} = \frac{1}{Z_\beta(\mathcal{D})} \prod_{l\in[d]} \underbrace{\exp(-\beta\lambda r(\theta_l))}_{p_\theta(\theta_l)} \prod_{i\in[n]} \underbrace{\exp\left(-\beta\ell\left(y_i, \boldsymbol{\theta}^\top \mathbf{x}_i\right)\right)}_{p_{out}(y_i|\boldsymbol{\theta}^\top \mathbf{x}_i)} \mathrm{d}\boldsymbol{\theta}, \tag{5}$$

where we have also defined the (unnormalised) functions p_θ and p_{out} that respectively play an analogous role to the prior and the likelihood in (3). It is easy to see that in the limit $\beta \to \infty$ the Gibbs measure concentrates around the configurations defined in (4). Hence, the ERM estimator can be seen as a MP estimator [3] in the following sense:

$$\hat{\boldsymbol{\theta}} = \lim_{\beta\to\infty} \langle \boldsymbol{\theta} \rangle_{\mu_\beta} = \lim_{\beta\to\infty} \frac{1}{Z_\beta(\mathcal{D})} \int \mathrm{d}\boldsymbol{\theta} \mu_\beta(\boldsymbol{\theta} \mid \mathcal{D})\boldsymbol{\theta} \tag{6}$$

The method

It is straightforward to estabilish a connection between this learning problem and the statistical physics of a disordered system. Consider a system with energy for the *configuration* $\boldsymbol{\theta}$, at fixed *disorder* $\mathcal{D}$, given by the following (rescaled[4]) *Hamiltonian*:

$$-\beta\mathcal{H}_\mathcal{D}(\boldsymbol{\theta}) := \sum_{l\in[d]} \log p_\theta(\theta_l) \sum_{i\in[n]} \log p_{out}\left(y_i \mid \boldsymbol{\theta}^\top \mathbf{x}_i\right),$$

[3] It is clear that, following this approach, the Gibbs measure μ_β defined in (5) plays for ERM the same role of the (student) posterior distribution (3) in MP estimation. Hence we will often refer to both of them as "posterior" for simplicity. Moreover we will also use the common notation $Z(\mathcal{D})$, droppping the subscript β, when the estimation method is not specified.

[4] The factor β, proportional to the inverse of the temperature in statistical physics, has no particular meaning in general (in particular for the evaluation of the MP estimator). Nevertheless, we choose to write it in order to estabilish a direct connection to physics' notation. On the contrary, for ERM we already have introduced a parameter β and we require $\beta \to \infty$ for the Gibbs measure to concentrate around the estimator. In physical language this correspond to say that the ERM estimator is the ground state of the system we have defined.

the customary way for a physicist to examine the properties of such systems is to evaluate the *free energy density* $f_{\mathcal{D}} := -\beta^{-1}d^{-1}\log Z(\mathcal{D})$, where the *partition function* $Z(\mathcal{D}) := \int \mathrm{d}\boldsymbol{\theta} e^{-\beta \mathcal{H}_{\mathcal{D}}(\boldsymbol{\theta})}$ is precisely the normalization of the posterior defined in (3) and in (5). In fact, the derivatives of the free energy can give access to the average of the observables of the system, for instance the energy, the heat capacity or the entropy.
Different specific realizations of the disorder (in our case the data $\mathcal{D}$) result in statistical fluctuations of the free energy density, thus it is assumed that, in the *thermodynamical limit* $d \to \infty$, $n/d = \alpha \in \mathbb{R}$, the *self averaging* property holds, *i.e.* the value of the free energy density for any $\mathcal{D}$ concentrate around its *typical value* $\mathbb{E}_{\mathcal{D}} f_{\mathcal{D}}$, which would simplify the study of the asymptotic performance of the estimators.
Hence, the quantity we are interested in is [5]

$$f = -\frac{1}{\beta}\lim_{d\to\infty}\frac{1}{d}\mathbb{E}_{\mathcal{D}}\log Z_\beta(\mathcal{D}), \tag{7}$$

often called *quenched* free energy density, where $\mathbb{E}_{\mathcal{D}}$ represent the average with respect to the teacher distributions. Note that we should take the limit $\beta \to \infty$ for ERM (see (6)), while we don't need it for Bayes optimal estimation, hence in that case we will consider $\beta = 1$.
The logarithm in the expressions in (7) can make the computation of the average hard or unfeasible. A heuristic tool used in statistical physics to avoid this issue is the *replica trick*, which shift the focus of the study from $\mathbb{E}_{\mathcal{D}}\log Z(\mathcal{D})$ to the more tractable $\mathbb{E}_{\mathcal{D}} Z(\mathcal{D})^r$, $r \in \mathbb{N}$. This corresponds to study r *replicas* of the system, with configurations $\boldsymbol{\theta}^1, \dots, \boldsymbol{\theta}^r$, each drawn from the posterior distribution (3) or (5).

The details of the replica method and the explicit computation of the free energy can be found in appendix A.
The usual procedure during the replica computations is to write (7) as a saddle point problem (see (14)) in the dimension d with respect to the following *overlap* parameters:

$$q^{ab} = d^{-1}\boldsymbol{\theta}^{a\top}\boldsymbol{\theta}^b, \quad a, b = 0, \dots, r,$$

where we use the notation $\boldsymbol{\theta}^0$ for the true weights $\boldsymbol{\theta}_*$.
The standard procedure at this point is to assume that the self averaging property holds also for the overlaps, which concentrate around their typical as $d \to \infty$, assuming therefore the *replica symmetric* ansatz for solution of the saddle point problem:

$$q^{00} = r^0 = d^{-1}\mathbb{E}_{\mathcal{D}}\boldsymbol{\theta}_*^\top\boldsymbol{\theta}_*, \qquad q^{0a} = m = d^{-1}\mathbb{E}_{\mathcal{D}}\left\langle \boldsymbol{\theta}_*^\top\boldsymbol{\theta}^{(1)}\right\rangle_1, \ 1 \le a \le r,$$
$$q^{aa} = r = d^{-1}\mathbb{E}_{\mathcal{D}}\left\langle \boldsymbol{\theta}^{(1)\top}\boldsymbol{\theta}^{(1)}\right\rangle_1, \ 1 \le a \le r, \quad q^{ab} = q = d^{-1}\mathbb{E}_{\mathcal{D}}\left\langle \boldsymbol{\theta}^{(1)\top}\boldsymbol{\theta}^{(2)}\right\rangle_2, \ 1 \le a < b \le r,$$

where we use the notation $\langle\cdot\rangle_k, k \in \mathbb{N}$, for the expectation operator with respect to the conditional posterior distribution $p(\boldsymbol{\theta}^{(1)} \mid \mathcal{D}) \dots p(\boldsymbol{\theta}^{(k)} \mid \mathcal{D})$ for MP estimation or $\mu_\beta(\boldsymbol{\theta}^{(1)} \mid \mathcal{D}) \dots \mu_\beta(\boldsymbol{\theta}^{(k)} \mid \mathcal{D})$ for ERM.
The self averaging assumption allows us to evaluate (in the thermodynamical limit) the estimation error ε_{est} as a function of the concentrated overlaps independently from the data $\mathcal{D}$:

$$\begin{aligned}\varepsilon_{\text{est}}(\hat{\boldsymbol{\theta}}) &= \mathbb{E}_{\mathcal{D}}\frac{||\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_*||^2}{d} = d^{-1}\mathbb{E}_{\mathcal{D}}\left[\left\langle \boldsymbol{\theta}^{(1)\top}\boldsymbol{\theta}^{(2)}\right\rangle_2 - 2\left\langle \boldsymbol{\theta}_*^\top\boldsymbol{\theta}^{(1)}\right\rangle_1 + \boldsymbol{\theta}_*^\top\boldsymbol{\theta}_*\right] \\ &= q - 2m + r^0 \overset{B.O.}{=} r^0 - q,\end{aligned} \tag{8}$$

[5] We write the limit $\lim_{d\to\infty}$ as a shorthand for $\lim_{d,n\to\infty,\, n/d=\alpha}$

The last equality is true only for Bayes optimal setting, since $q = m$ due to the Nishimori identity. A brief proof can be found in appendix C.2.
In conclusion, the main advantage of the replica approach is that it bypasses the unfeasible high-dimensional problem of sampling from the posterior distribution, giving direct access to the overlap parameters, estimated through numerically solvable self-consistent equations.

3 Main results

In this section are reported the self consistent equations obtained through the replica approach (detailed derivation in appendix A.1) for Bayes optimal estimation and empirical risk minimization, for different choices of output channels and risk functions.
For simplicity we have chosen $\boldsymbol{\mu} = \mathbf{0}$ in (1) to derive our results.
The result for the parameter r^0 found in (16) is generic and it applies to all cases studied, as $p^*_\theta(\theta_*) = \mathcal{N}(\theta_* \mid 0, 1)$:

$$r^0 = \mathbb{E}_{\theta_*}[\theta_*] = 1.$$

For each case we also show some plots for the mean square error ε_{est} (computed solving numerically the self consistent equations) at varying sample complexity, for different value of a in $\rho_>(\Delta \mid a)$ or $\rho_<(\Delta \mid a)$, as defined in section 2. All results are compared to the case of Gaussian covariates, which in the plots is indicated by the label $a = \text{inf}$. In fact, as discussed in appendix D, the Gaussian case correspond to $\rho_>(\Delta \mid a)$ in the limit $a \to \infty$.
In particular we are interested in the dependance of ε_{est} from α at large sample complexity $\alpha \gg 1$: considering the leading term of $\varepsilon_{\text{est}} = \alpha^{-c} + O(\alpha^{-c-1})$, $c \in \mathbb{R}^+$, we find that in all the studied cases the *decay rate* $c = 1$, as in the Gaussian covariates case, indipendently from the choice of $\rho(\Delta)$. We also perform a linear regression of the curves in a *log-log* plot and measure the decay rate c as the slope of the fitted lines [6], in order to compare it with our prediction. The results are reported in Table 1. [7]
At a later stage we compare the theoretical performance of the MMSE estimator (*i.e.* Bayes optimal eastimation) and ridge regression, after optimizing the rgression parameter λ.

3.1 Bayes optimal setting

3.1.1 Linear channel

The linear output channel corresponds to $y_i = \boldsymbol{\theta}_*^\top \mathbf{x}_i + \sigma\xi$, where the noise $\xi \sim \mathcal{N}(\xi \mid 0, 1)$. In this case $p_{out}(y \mid z) = \mathcal{N}(y \mid z, \sigma^2)$. The saddle point equations coming from the replica computations are the following:

$$q = \frac{\hat{q}}{1+\hat{q}}, \qquad\qquad \hat{q} = \alpha \mathbb{E}_\Delta \left[\frac{\Delta}{\Delta(1-q)+\sigma^2} \right]$$

Their detailed derivation can be found in A.2.
By solving them numerically it is possible to compute the means square error as $\varepsilon_{\text{est}} = 1 - q$ (see (8)). In fig.3, 4, we show some results obtained for noise variance $\sigma^2 = 0.1$. In particular the plots on the right of both figures (in *log-log* scale) shows the behaviour of ε_{est} at large sample complexity.

[6] In fact $\varepsilon_{\text{est}} \sim \alpha^c \implies \log \varepsilon_{\text{est}} \sim -c \log \alpha + \text{const}$.
[7] Note that the linear fit results strongly depends on the range of values of α and most importantly on the precision threshold chosen for solving the self consistent equations iteratively. Hence, the value of the slopes are presented here just as a reference and to compare them with our predictions, but they serve no other purpose.

a	BO linear	BO probit	ERM ridge	ERM lasso
0.3	-1.0031(6)	-	-1.0016(3)	-
0.5	-1.012(2)	-	-1.0041(7)	-
0.8	-1.049(6)	-	-1.013(2)	-
1.005	-1.088(2)	-	-1.087(2)	-1.089(2)
1.1	-1.050(2)	-1.002(3)	-1.050(2)	-1.050(2)
3	-1.00019(4)	-0.9995(2)	-1.00016(3)	-1.00023(4)
10	-1.00010(2)	-	-1.00007(1)	-1.00015(3)
∞	-1.00009(2)	-0.9994(2)	-1.00006(1)	-1.00014(2)

Table 1: Decay rates (with error) of ε_{est} at large α, *i.e.* the slopes from the linear regression of the learning curves in the log-log scale (right plots of figs. 3, 4, 5, 6, 7, 8). The distribution $\rho(\Delta)$ is $\rho_<(\Delta \mid a)$ when $a < 1$ and to $\rho_>(\Delta \mid a)$ when $a > 1$.

Considering the leading order in $\varepsilon_{\text{est}} = 1 - q = q_0 \alpha^{-c} + O(\alpha^{-(c+1)}$, $c \in \mathbb{R}^+$ and $q_0 \in \mathbb{R}$ a constant, we find that

$$\hat{q} = \frac{\alpha}{\sigma^2} \mathbb{E}_\Delta \left[\Delta \left(1 - \frac{\Delta q_0 \alpha^{-c}}{\sigma^2} + O(\alpha^{-c-1}) \right) \right] \implies 1 - q = \frac{1}{1+\hat{q}} = \frac{\sigma^2}{\alpha \overline{\Delta}} + O(\alpha^{-2}) \implies c = 1$$

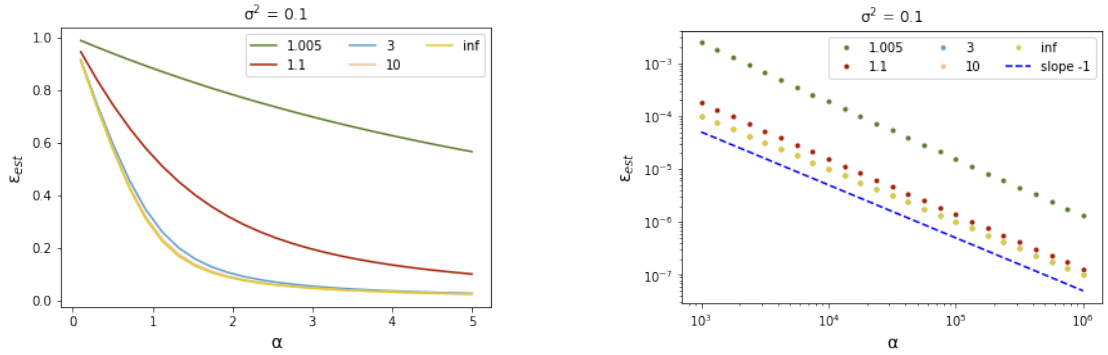

Figure 3: Estimation error in Bayes optimal setting with linear output channel at varying sample complexity. The noise variance is $\sigma^2 = 0.1$. Different colors correspond to different values for the parameter a in $\rho_>(\Delta \mid a)$, as stated in the legend.

3.1.2 Probit channel

The *probit* output channel corresponds to $y_i = \text{sign}\left(\boldsymbol{\theta}_*^\top \mathbf{x}_i + \sigma\xi\right)$, where the noise $\xi \sim \mathcal{N}(\xi \mid 0, 1)$. This is a common way to generate dicotomized labels in classification problems. In this case $p^*_{out}(y|z) = \text{erfc}\left(\frac{-yz}{\sqrt{2}\sigma}\right)/2$, $y \in \{-1, 1\}$. [8] The saddle point equations coming from the replica

[8] $\text{erfc}(x) = \frac{2}{\pi} \int_x^\infty \text{d}t e^{-t^2}$ is the *complementary error function*.

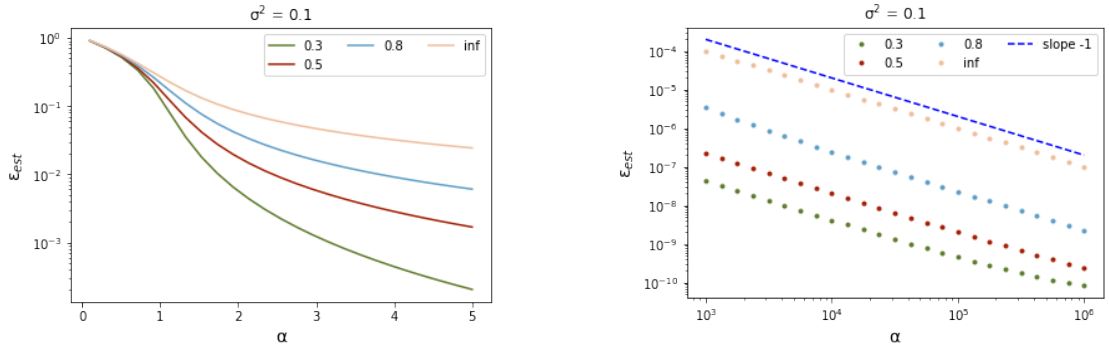

Figure 4: Estimation error in Bayes optimal setting with linear output channel at varying sample complexity. The noise variance is $\sigma^2 = 0.1$. Different colors correspond to different values for the parameter a in $\rho_<(\Delta \mid a)$, as stated in the legend.

computations are the following:

$$q = \frac{\hat{q}}{1+\hat{q}}, \qquad \hat{q} = \frac{\alpha}{q\sqrt{2\pi}} \mathbb{E}_{\Delta,\eta}\left[k_q(\Delta)\left(1+k_q(\Delta)^2\right) e^{-\frac{k_q(\Delta)^2}{2}\eta^2} \eta \log \operatorname{erfc}\left(\eta \frac{k_q(\Delta)}{\sqrt{2}}\right)\right],$$

where $\eta \sim \mathcal{N}(\eta \mid 0, 1)$ and $k_q(\Delta) := \sqrt{\Delta q}/\sqrt{\sigma^2 + \Delta(1-q)}$. Their detailed derivation can be found in A.2.
Again we can solve these equation numerically and in fig. 5 we show some results for noise variance $\sigma^2 = 0.1$. In particular the plot on the left (in *log-log* scale) shows the behaviour of ε_{est} at large sample complexity. Considering the leading order in $\varepsilon_{\text{est}} = 1 - q = q_0 \alpha^{-c} + O(\alpha^{-(c+1)})$, $c \in \mathbb{R}^+$ and $q_0 \in \mathbb{R}$ a constant, we find that

$$k_q(\Delta) = \frac{\sqrt{\Delta}\left(1 - \frac{1}{2} q_0 \alpha^{-c}\right)}{\sqrt{\sigma^2 + \Delta q_0 \alpha^{-c}}} + O(\alpha^{-c-1}) = \sqrt{\frac{\Delta}{\sigma^2}}\left(1 - \frac{1}{2} q_0 \alpha^{-c}\right)\left(1 - \frac{\Delta}{2\sigma^2} q_0 \alpha^{-c}\right) + O(\alpha^{-c-1}) \implies$$

$$\hat{q} = \frac{\alpha}{\sqrt{2\pi}} \mathbb{E}_{\Delta,\eta}\left[\sqrt{\frac{\Delta}{\sigma^2}}\left(1 + \frac{\Delta}{\sigma^2}\right) e^{-\frac{\Delta}{2\sigma^2}\eta^2} \eta \log \operatorname{erfc}\left(\eta \sqrt{\frac{\Delta}{2\sigma^2}}\right)\right] + O(1) \implies$$

$$1 - q = \frac{\sqrt{2\pi}}{\alpha}\left(\mathbb{E}_{\Delta,\eta}\left[\sqrt{\frac{\Delta}{\sigma^2}}\left(1 + \frac{\Delta}{\sigma^2}\right) e^{-\frac{\Delta}{2\sigma^2}\eta^2} \eta \log \operatorname{erfc}\left(\eta \sqrt{\frac{\Delta}{2\sigma^2}}\right)\right]\right)^{-1} + O(\alpha^{-2}) \implies c = 1,$$

3.2 Empirical risk minimization

In this section we show the results for the performance evaluation of the estimator (4) in solving our task, when the teacher distributions are p^*_θ and p^*_{out}.

3.2.1 Ridge regression

Ridge regression is a method that considers the risk function composed of quadratic loss $\ell(y, z) = (y - z)^2/2$ and L_2 regularization $r(\theta) = \theta^2/2$.[9] In this section we consider a teacher model with the usual Gaussuan prior for the weights and labels generated by a linear output channel (as

[9]The regularization term in (4) can be seen as the L_2 norm acting on $\boldsymbol{\theta}$: $\sum_l \theta_l^2/2 = ||\boldsymbol{\theta}||^2/2$.

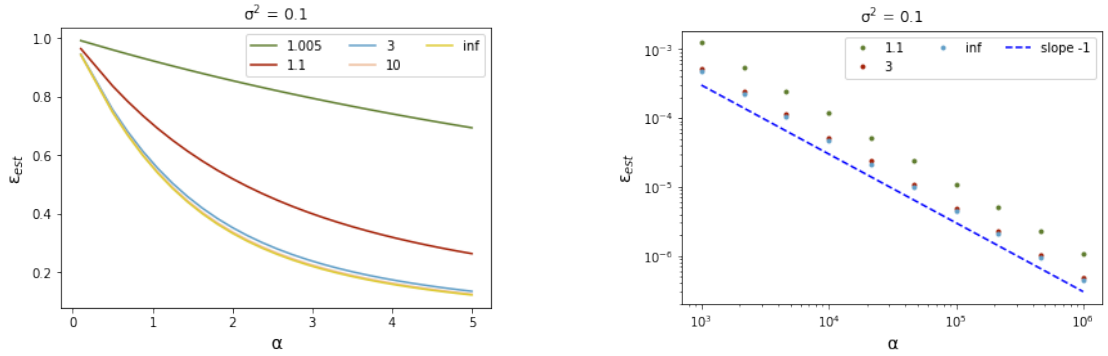

Figure 5: Estimation error in Bayes optimal setting with probit output channel at varying sample complexity. The noise variance is $\sigma^2 = 0.1$. Different colors correspond to different values for the parameter a in $\rho_>(\Delta \mid a)$, as stated in the legend.

defined in section 3.1.1). The saddle points equations coming from the replica computations are the following:

$$m = \frac{\hat{M}}{\hat{R}+\lambda}, \qquad r = \frac{\hat{M}^2+\hat{\chi}}{(\hat{R}+\lambda)^2}, \qquad \chi = \frac{1}{\hat{R}+\lambda},$$
$$\hat{M} = \alpha \mathbb{E}_\Delta \frac{\Delta}{1+\Delta\chi}, \qquad \hat{R} = \alpha \mathbb{E}_\Delta \frac{\Delta}{1+\Delta\chi}, \qquad \hat{\chi} = \alpha \mathbb{E}_\Delta \frac{\Delta(\sigma^2+\Delta(1-2m+r))}{(1+\Delta\chi)^2},$$

where we have defined $\chi = \beta(r-q)$. A detailed derivation can be found in appendix A.3. These equation can be solved numerically, allowing to compute the mean square error of the estimator as a function of the overlap parameters $\varepsilon_{\text{est}} = 1-2m+r$, as seen in (29). We show in fig. 6, 7 some results for ε_{est} at varying sample complexity for noise variance $\sigma^2 = 0.1$ and $\lambda = 0.3$. In particular the plot on the left of both figures (in *log-log* scale) shows the behaviour of ε_{est} at large sample complexity. Considering the leading order in $\varepsilon_{\text{est}} = 1-2m+r = \varepsilon_0\alpha^{-c}+O(\alpha^{-(c+1)})$ and $\chi = \chi_0\alpha^{-c}+O(\alpha^{-(c+1)})$, $c \in \mathbb{R}^+$ and $\varepsilon_0 \in \mathbb{R}$ a constant, we find that

$$\begin{cases} \hat{M} = \hat{R} = \alpha(\overline{\Delta} + O(\alpha^{-c})) \\ \hat{\chi} = \alpha(\sigma^2\overline{\Delta} + O(\alpha^{-c})) \end{cases} \implies \begin{cases} m = \frac{\hat{M}}{\hat{R}}\left(1-\frac{\lambda}{\alpha\overline{\Delta}}\right) + O(\alpha^{-2}) \\ r = \frac{\hat{M}^2}{\hat{R}^2}\left(1+\frac{\sigma^2}{\alpha\overline{\Delta}}\right)\left(1-2\frac{\lambda}{\alpha\overline{\Delta}}\right) + O(\alpha^{-2}) \\ \chi = \alpha^{-1}\overline{\Delta}^{-1} + O(\alpha^{-2}) \end{cases}$$

$$\implies \begin{cases} \varepsilon_{\text{est}} = 1-2m+r = \alpha^{-1}\sigma^2\overline{\Delta}^{-1} + O(\alpha^{-2}) \\ \chi = \alpha^{-1}\overline{\Delta}^{-1} + O(\alpha^{-2}) \end{cases} \implies c = 1$$

3.2.2 Lasso regression

Lasso regression is a method that considers the risk function composed of the quadratic loss function $\ell(y,z) = (y-z)^2/2$ and the L_1 regularization $r(\theta) = |\theta|$.[10] As we did for the ridge regression, we consider a teacher model with the usual Gaussian prior for the weights and labels

[10] The regularization term in (4) can be seen as the L_1 norm acting on $\boldsymbol{\theta}$: $\sum_l |\theta_l| = ||\boldsymbol{\theta}||_1$.

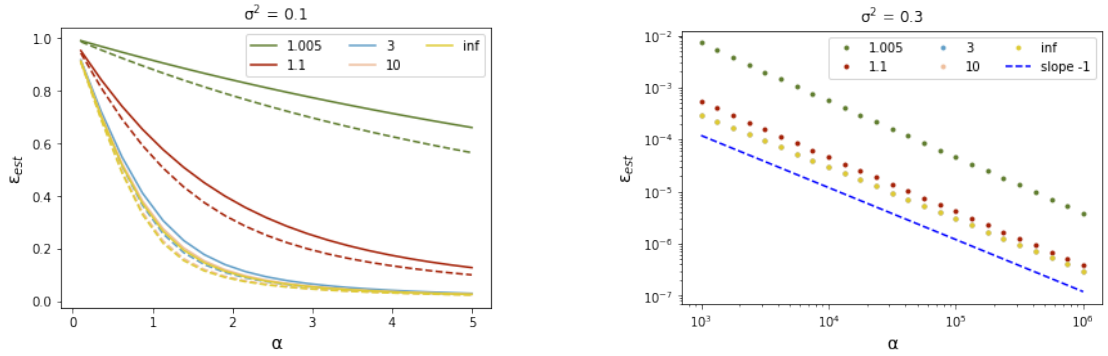

Figure 6: Estimation error for ridge regression at varying sample complexity. The noise variance is $\sigma^2 = 0.1$. Different colors correspond to different values for the parameter a in $\rho_>(\Delta \mid a)$, as stated in the legend.

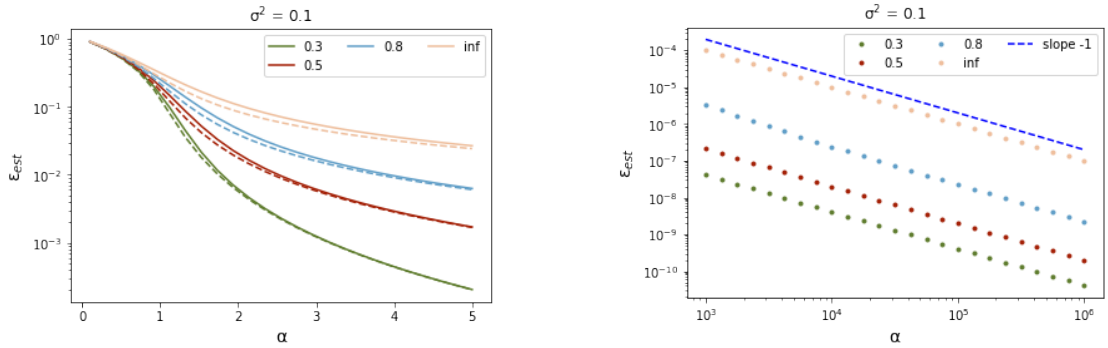

Figure 7: Estimation error for ridge regression at varying sample complexity. The noise variance is $\sigma^2 = 0.1$. Different colors correspond to different values for the parameter a in $\rho_<(\Delta \mid a)$, as stated in the legend.

generated by a linear output channel (as defined in section 3.1.1). The saddle points equations coming from the replica computations are the following:

$$
\begin{aligned}
m &= \frac{\hat{M}}{\hat{R}}\phi_{\text{erfc}}, & r &= \frac{1}{\hat{R}^2}\phi_\theta, & \chi &= \frac{1}{\hat{R}}\phi_{\text{erfc}}, \\
\hat{M} &= \alpha \mathbb{E}_\Delta \frac{\Delta}{1+\Delta\chi}, & \hat{R} &= \alpha \mathbb{E}_\Delta \frac{\Delta}{1+\Delta\chi}, & \hat{\chi} &= \alpha \mathbb{E}_\Delta \frac{\Delta(\sigma^2 + \Delta(1-2m+r))}{(1+\Delta\chi)^2},
\end{aligned}
$$

where we have defined

$$
\phi_{\text{erfc}} := \text{erfc}\left(\frac{\lambda}{\sqrt{2(\hat{M}^2+\hat{\chi})}}\right), \quad \phi_\theta := \left(\hat{M}^2 + \hat{\chi} + \lambda^2\right)\phi_{\text{erfc}} - \lambda\sqrt{\frac{2\left(\hat{M}^2+\hat{\chi}\right)}{\pi}}e^{-\frac{\lambda^2}{2(\hat{M}^2+\hat{\chi})}}.
$$

A detailed derivation can be found in appendix A.3.
As in all the other cases, these equations can be solved numerically, and we show in fig. 8

some results for ε_{est} at varying sample complexity for noise variance $\sigma^2 = 0.1$ and $\lambda = 0.3$. In particular the plot on the left (in *log-log* scale) shows the behaviour of ε_{est} at large sample complexity. Considering the leading order in $\varepsilon_{\text{est}} = 1 - 2m + r = \varepsilon_0 \alpha^{-c} + O(\alpha^{-(c+1)})$ and $\chi = \chi_0 \alpha^{-c} + O(\alpha^{-(c+1)})$, $c \in \mathbb{R}^+$ and $\varepsilon_0 \in \mathbb{R}$ a constant, we find that [11]

$$
\begin{cases} \hat{M} = \hat{R} = \alpha(\overline{\Delta} + O(\alpha^{-c})) \\ \hat{\chi} = \alpha(\sigma^2 \overline{\Delta} + O(\alpha^{-c})) \end{cases}
$$

$$
\implies \begin{cases} \phi_{\text{erfc}} = 1 - \dfrac{2\lambda}{\sqrt{2\pi(\alpha^2 \overline{\Delta}^2 + \alpha\sigma^2\overline{\Delta})}} + O(\alpha^{-2}) = 1 - \sqrt{\dfrac{2}{\pi}} \lambda \alpha^{-1} \overline{\Delta}^{-1} + O(\alpha^{-2}) \\ \phi_\theta = \left(\alpha^2 \overline{\Delta}^2 + \alpha\sigma^2\overline{\Delta}\right)\left(1 - \sqrt{\dfrac{2}{\pi}} \lambda \alpha^{-1} \overline{\Delta}^{-1}\right) - \lambda\alpha\overline{\Delta}\sqrt{\dfrac{2}{\pi}}\left(1 + \dfrac{\sigma^2}{2\alpha\overline{\Delta}}\right)\left(1 - \dfrac{\lambda^2}{2\alpha^2\overline{\Delta}^2}\right) + O(\alpha^{-2}) \\ \quad = \alpha^2\overline{\Delta}^2 + \alpha\sigma^2\overline{\Delta} - 2\sqrt{\dfrac{2}{\pi}}\lambda\alpha\overline{\Delta} + O(1) \end{cases}
$$

$$
\implies \begin{cases} m = 1 - \sqrt{\dfrac{2}{\pi}} \lambda \alpha^{-1} \overline{\Delta}^{-1} + O(\alpha^{-2}) \\ r = \left(\alpha^2 \overline{\Delta}^2 + \alpha\sigma^2\overline{\Delta} - 2\sqrt{\dfrac{2}{\pi}}\lambda\alpha\overline{\Delta} + O(1)\right)\left(\alpha(\overline{\Delta} + O(\alpha^{-c}))\right)^{-2} \\ \quad = 1 + \sigma^2 \alpha^{-1} \overline{\Delta}^{-1} - 2\sqrt{\dfrac{2}{\pi}} \lambda \alpha^{-1} \overline{\Delta}^{-1} + O(\alpha^{-2}) \\ \chi = \left(1 - \sqrt{\dfrac{2}{\pi}} \lambda \alpha^{-1} \overline{\Delta}^{-1} + O(\alpha^{-2})\right)\left(\alpha(\overline{\Delta} + O(\alpha^{-c}))\right)^{-1} = \alpha^{-1}\overline{\Delta}^{-1} + O(\alpha^{-2}) \end{cases}
$$

$$
\implies \begin{cases} \varepsilon_{\text{est}} = 1 - 2m + r = \alpha^{-1}\sigma^2\overline{\Delta}^{-1} + O(\alpha^{-2}) \\ \chi = \alpha^{-1}\overline{\Delta}^{-1} + O(\alpha^{-2}) \end{cases} \implies c = 1
$$

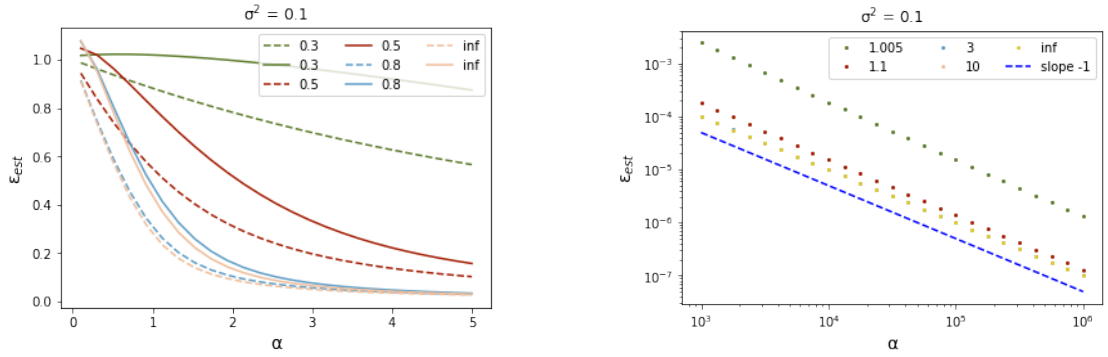

Figure 8: Estimation error for lasso regression at varying sample complexity. The noise variance is $\sigma^2 = 0.1$. Different colors correspond to different values for the parameter a in $\rho_>(\Delta \mid a)$, as stated in the legend.

[11] We use $\text{erfc}(z) = 1 - \frac{2}{\sqrt{\pi}} z + O(z^3)$

3.3 Comparison and optimization of λ

In this section we consider the same teacher model defined in 3.1.1, 3.2.1 and 3.2.2, with Gaussian prior $p_\theta(\theta) = \mathcal{N}(\theta \mid 0, 1)$ and linear output channel / likelihood $p_{out}(y \mid z) = \mathcal{N}(y \mid z, \sigma^2)$. We distinguish the possible estimators of the true weights that have been studied in the previous sections by tintroducing the notation $\hat{\boldsymbol{\theta}}^{BO}$ for the mean posterior estimator in Bayes optimal setting (*i.e.* the minimal mean square error estimator), $\hat{\boldsymbol{\theta}}_\lambda^{L_2}$ the estimator obtained through the ridge regression and $\hat{\boldsymbol{\theta}}_\lambda^{L_1}$ the one obtained through the lasso regression. By definition of the MMSE estimator, we expect $\varepsilon_{\text{est}}\left(\hat{\boldsymbol{\theta}}^{BO}\right) \leq \varepsilon_{\text{est}}\left(\hat{\boldsymbol{\theta}}_\lambda^{L_2}\right), \varepsilon_{\text{est}}\left(\hat{\boldsymbol{\theta}}_\lambda^{L_1}\right), \forall\lambda$. One advantage of having equations able to predict the performance of ERM is that it is possible to optimize the algorithm by selecting the parameter λ that minimizes the mean square error:[12]

$$\lambda_\gamma^* = \arg\min_\lambda \varepsilon_{\text{est}}\left(\hat{\boldsymbol{\theta}}_\lambda^{L_\gamma}\right), \quad \gamma = 1, 2.$$

Note that the optimized λ_γ^* is a function of the parameters used to generate the data (in our case the noise variance σ^2 in the output channel and the parameter a in $\rho(\Delta \mid a)$) and the sample complexity α. In fig. 9 we show the results of this optimization procedure for ridge regression, with a comparison to the MMSE

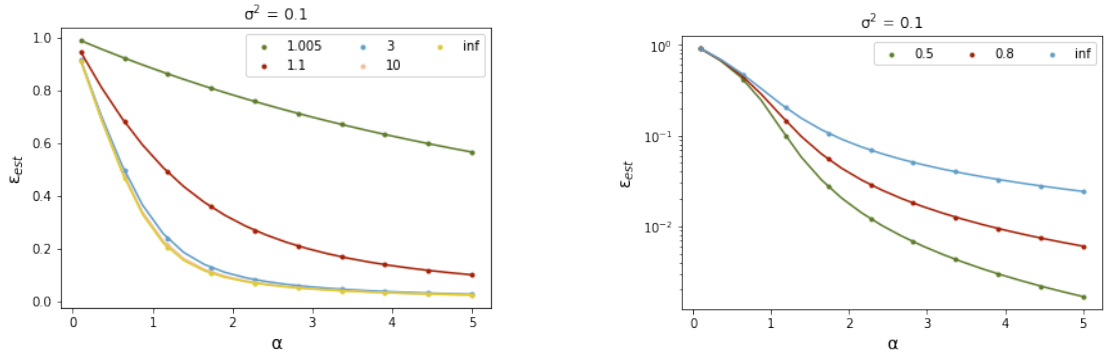

Figure 9: Estimation error for optimized ridge regression (points) compared to Bayes optimal estimation (lines) at varying sample complexity. The noise variance is $\sigma^2 = 0.1$. Different colors correspond to different values for the parameter a in $\rho_>(\Delta \mid a)$ (left) or $\rho_<(\Delta \mid a)$ (right), as stated in the legend.

4 Application to synthetic data

In this section we compare numerical experiments to the performance predictions obtained through the replica approach for all the cases shown in 3. All experiments are done considering $d = 1000$ and generating 20 instances of the problem.

[12] In practical applications of ERM, when precise asymptotics are not known, the choice of λ can be made through the procedure of *cross-validation*: the data are divided in k subsets of the same size, training the model on $k-1$ of them and testing the predictions on the remaining one. The procedure can be repeated using each of the subsets for testing. Eventually one chooses te value of λ minimizing the test error.

4.1 Bayes optimal setting

Following previous literature, Bayes optimal estimation for GLM can be efficiently performed using a Generalized Approximate Message Passing (GAMP) algorithm. This polynomial (with respect to d) procedure allows to avoid the sampling from the posterior (3), which is computationally costly in high-dimensions, performing optimal estimation in this specific setting [6]. The algorithm and further details are shown in C.3, while its derivation and state evolution can be found in [23, 41].
Fig. 10, 11 show the comparison of these numerical experiments obtained from GAMP and the estimation error curves shown in 3.1.1.

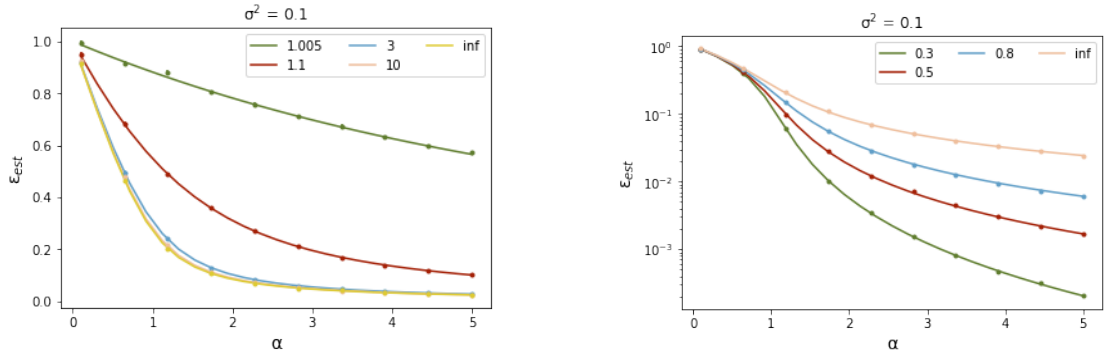

Figure 10: Estimation error in Bayes optimal setting with linear output channel at varying sample complexity: theoretical prediction (lines) and average from numerical GAMP experiments (points) with error. The noise variance is $\sigma^2 = 0.1$. Different colors correspond to different values for the parameter a in $\rho_>(\Delta \mid a)$ (left) or $\rho_<(\Delta \mid a)$ (right), as stated in the legend.

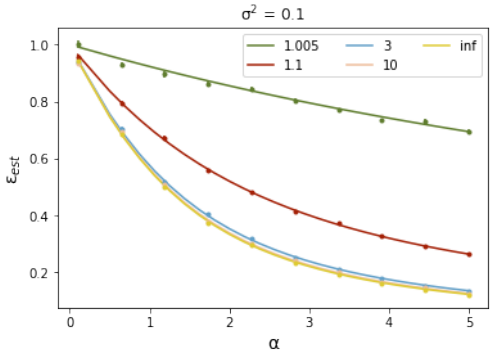

Figure 11: Estimation error in Bayes optimal setting with probit output channel at varying sample complexity: theoretical prediction (lines) and average from numerical GAMP experiments (points) with error. The noise variance is $\sigma^2 = 0.1$. Different colors correspond to different values for the parameter a in $\rho_>(\Delta \mid a)$, as stated in the legend.

4.2 Empirical risk minimization

In fig. 12, 13 we show the comparison between numerical experiments of ERM and the estimation error curves obtained from replica computations.

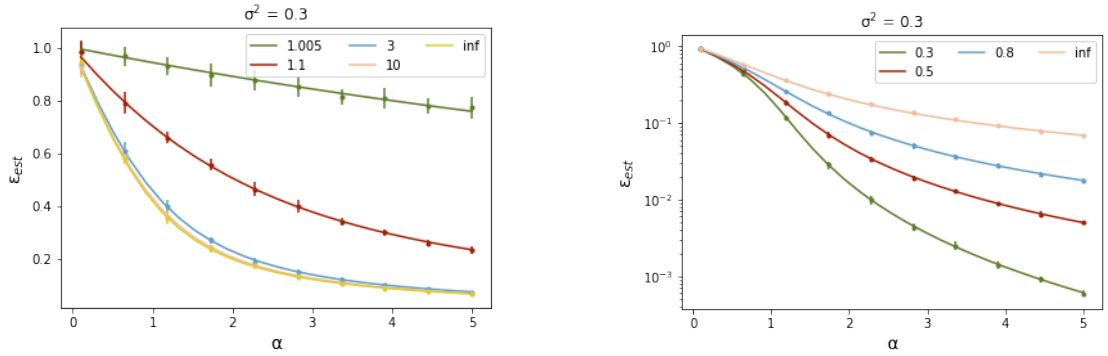

Figure 12: Estimation error from ridge regression at varying sample complexity: theoretical prediction (lines) and average from numerical experiments (points) with error. The noise variance is $\sigma^2 = 0.3$. Different colors correspond to different values for the parameter a in $\rho_>(\Delta \mid a)$ (left) or $\rho_<(\Delta \mid a)$ (right), as stated in the legend.

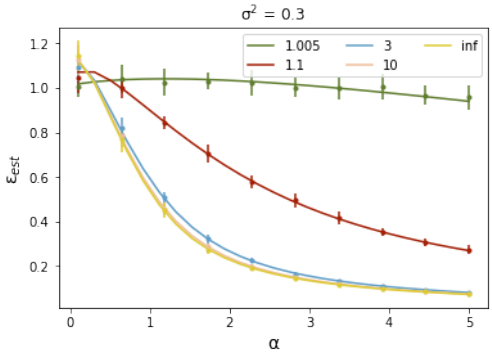

Figure 13: Estimation error from lasso regression at varying sample complexity: theoretical prediction (lines) and average from numerical experiments (points) with error. The noise variance is $\sigma^2 = 0.3$. Different colors correspond to different values for the parameter a in $\rho_>(\Delta \mid a)$, as stated in the legend.

5 Conclusions

In this research work we have studied the problem of learning generalized linear models in high-dimensions when the covariates are drawn from the superstatistical distribution $P(\mathbf{x})$ in (1). This choice allows to consider very different non-Gaussian covariates' distributions (heavy-tails, infinite variance) with the advantage of having a Gaussian conditioned probability density function $P(\mathbf{x} \mid \Delta) = \mathcal{N}(\mathbf{x} \mid \boldsymbol{\mu}, d^{-1}\Delta\mathbb{I}_d)$, which simplifies the computations. We reached our goal of evaluating the best performance (given the data) any algorithm can achieve in case of linear and probit labels, considering the Bayes optimal setting. We verified, in the linear case, that optimized ridge regression and lasso regression performance are comparable to the Bayes optimal ones. We were also able to show that our predictions obtained through the heuristic replica method agree with numerical experiments performed on synthetic data. One of the most important results we have obtained is the computation of the decay rates of the learning curves, which are the same indipendently of the distribution for Δ and equal to the known Gaussian covariates case (where $P(\mathbf{x}) = \mathcal{N}(\mathbf{x} \mid \boldsymbol{\mu}, d^{-1}\overline{\Delta}\mathbb{I}_d)$) with matching $\mathbb{E}_\Delta[\Delta] = \overline{\Delta}$. This implies that even though the covariates' superstatistical distribution may have heavy-tails or infinite variance, ultimately this does not affect the performance of the estimation, in particular when the sample size n is much larger (but still proportional to) the covariates' dimension d, *i.e.* the regime of large sample complexity. This results finds its place in the context of the Gassian universality principle and represent a further steps towards understanding how learning algorithms perform on realistic datasets and how much the latters can be described by the first two moments of they distribution.

The natural progression of this work would be to perform numerical experiments on realistic data whose features are supposed to follow a superstatistical model and validate the results here presented. Moreover one could use the results of A.1 and consider different choice for the output channel or the risk function (e.g., studying ERM for classification). The results presented in this manuscript can be straightforwardly applied to the case of covariates distributed according to Huber's contamination model [21, 22], considering the case where each covariate can be drawn from a Gaussian or a superstatistical model with probabilities respectively equal to $1 - \epsilon$ and ϵ, for some $\epsilon \in [0, 1]$. Further steps can be the introduction of structure in the noise of the output channel, for instance drawn from a superstatistical distribution (or a contamination model) likewise. Another development could be a rigorous proof of our results, exploiting previous exact results on GLMs.

References

[1] U. Adomaityte, G. Sicuro, and P. Vivo, *Classification of Superstatistical Features in High Dimensions*, May 2023.

[2] R. J. Alder, R. E. Feldman, and M. S. Taqqu, *A Practical Guide to Heavy Tails: Statistical Techniques and Applications*, Journal of the American Statistical Association, 94 (1999), p. 653.

[3] M. Andrea, *Estimating random variables from random sparse observations*, European Transactions on Telecommunications, 19 (2008).

[4] C. Baldassi, E. M. Malatesta, M. Negri, and R. Zecchina, *Wide flat minima and optimal generalization in classifying high-dimensional Gaussian mixtures*, Journal of Statistical Mechanics: Theory and Experiment, 2020 (2020), p. 124012.

[5] J. Barbier, *Overlap matrix concentration in optimal Bayesian inference*, Jan. 2020.

[6] 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 (2019), pp. 5451–5460.

[7] 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 (2019), pp. 1133–1185.

[8] A. R. Barron and A. Joseph, *Toward fast reliable communication at rates near capacity with Gaussian noise*, IEEE, June 2010, pp. 315–319.

[9] C. Beck, *Superstatistics: Theory and Applications*, Continuum Mechanics and Thermodynamics, 16 (2004).

[10] C. Beck and E. G. D. Cohen, *Superstatistics*, Physica A: Statistical Mechanics and its Applications, 322 (2003).

[11] M. Chen, C. Gao, and Z. Ren, *A General Decision Theory for Huber's ϵ-Contamination Model*, Jan. 2017.

[12] L. Clarté, B. Loureiro, F. Krzakala, and L. Zdeborová, *Theoretical characterization of uncertainty in high-dimensional linear classification*, Machine Learning: Science and Technology, 4 (2023).

[13] A. C. C. Coolen, M. Sheikh, A. Mozeika, F. Aguirre-Lopez, and F. Antenucci, *Replica analysis of overfitting in generalized linear regression models*, Journal of Physics A: Mathematical and Theoretical, 53 (2020), p. 365001.

[14] Y. Dandi, L. Stephan, F. Krzakala, B. Loureiro, and L. Zdeborová, *Universality laws for Gaussian mixtures in generalized linear models*, Feb. 2023.

[15] D. Delpini and G. Bormetti, *Minimal model of financial stylized facts*, Physical Review E, 83 (2011), p. 041111.

[16] F. Gerace, F. Krzakala, B. Loureiro, L. Stephan, and L. Zdeborová, *Gaussian Universality of Perceptrons with Random Labels*, May 2022.

[17] F. Gerace, B. Loureiro, F. Krzakala, M. Mezard, and L. Zdeborova, *Generalisation error in learning with random features and the hidden manifold model*, in Proceedings of the 37th International Conference on Machine Learning, PMLR, Nov. 2020, pp. 3452–3462.

[18] C. Gerbelot, A. Abbara, and F. Krzakala, *Asymptotic Errors for Teacher-Student Convex Generalized Linear Models (Or: How to Prove Kabashima's Replica Formula)*, IEEE Transactions on Information Theory, 69 (2023).

[19] S. Goldt, B. Loureiro, G. Reeves, F. Krzakala, M. Mezard, and L. Zdeborova, *The Gaussian equivalence of generative models for learning with shallow neural networks*, in Proceedings of the 2nd Mathematical and Scientific Machine Learning Conference, PMLR, Apr. 2022, pp. 426–471.

[20] H. Hu and Y. M. Lu, *Universality Laws for High-Dimensional Learning with Random Features*, Oct. 2022.

[21] P. J. Huber, *Robust Estimation of a Location Parameter*, The Annals of Mathematical Statistics, 35 (1964), pp. 73–101.

[22] ——, *A Robust Version of the Probability Ratio Test*, The Annals of Mathematical Statistics, 36 (1965), pp. 1753–1758.

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

[24] Y. Kabashima, F. Krzakala, M. Mézard, A. Sakata, and L. Zdeborová, *Phase transitions and sample complexity in Bayes-optimal matrix factorization*, IEEE Transactions on Information Theory, 62 (2016).

[25] K. Kiyono and H. Konno, *Log-amplitude statistics for Beck-Cohen superstatistics*, Physical Review E, 87 (2013), p. 052104.

[26] N. Langrené, G. Lee, and Z. Zhu, *Switching to non-affine stochastic volatility: A closed-form expansion for the Inverse Gamma model*, International Journal of Theoretical and Applied Finance, 19 (2016), p. 1650031.

[27] M. Lelarge and L. Miolane, *Asymptotic Bayes risk for Gaussian mixture in a semi-supervised setting*, Sept. 2019.

[28] J. Li, R. Zhao, J.-T. Huang, and Y. Gong, *Learning small-size DNN with output-distribution-based criteria*, in Interspeech 2014, ISCA, Sept. 2014, pp. 1910–1914.

[29] 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*, in Advances in Neural Information Processing Systems, vol. 34, Curran Associates, Inc., 2021, pp. 10144–10157.

[30] P. McCullagh, *Generalized linear models*, European Journal of Operational Research, 16 (1984), pp. 285–292.

[31] S. Mei and A. Montanari, *The Generalization Error of Random Features Regression: Precise Asymptotics and the Double Descent Curve*, Communications on Pure and Applied Mathematics, 75 (2022), pp. 667–766.

[32] M. Mezard, G. Parisi, and M. A. Virasoro, *Spin Glass Theory And Beyond: An Introduction To The Replica Method And Its Applications*, World Scientific Publishing Company, Nov. 1987.

[33] F. Mignacco, F. Krzakala, Y. M. Lu, and L. Zdeborová, *The role of regularization in classification of high-dimensional noisy Gaussian mixture*, Feb. 2020.

[34] A. Montanari and P. M. Nguyen, *Universality of the elastic net error | 2017 IEEE International Symposium on Information Theory (ISIT)*.

[35] A. Montanari and B. Saeed, *Universality of empirical risk minimization*, Oct. 2022.

[36] J. A. Nelder and R. W. M. Wedderburn, *Generalized Linear Models*, Journal of the Royal Statistical Society. Series A (General), 135 (1972).

[37] D. B. Nelson, *ARCH models as diffusion approximations*, Journal of Econometrics, 45 (1990), pp. 7–38.

[38] H. Nishimori, *Mean-Field Theory of Phase Transitions*, in Statistical Physics of Spin Glasses and Information Processing: An Introduction, H. Nishimori, ed., Oxford University Press, July 2001.

[39] A. Panahi and B. Hassibi, *A Universal Analysis of Large-Scale Regularized Least Squares Solutions*, in Advances in Neural Information Processing Systems, vol. 30, Curran Associates, Inc., 2017.

[40] A. Pensia, V. Jog, and P.-L. Loh, *Robust regression with covariate filtering: Heavy tails and adversarial contamination*, May 2021.

[41] S. Rangan, *Generalized Approximate Message Passing for Estimation with Random Linear Mixing*, Aug. 2012.

[42] W. J. Reed and B. D. Hughes, *From gene families and genera to incomes and internet file sizes: Why power laws are so common in nature*, Physical Review E, 66 (2002), p. 067103.

[43] P. Sur and E. J. Candès, *A modern maximum-likelihood theory for high-dimensional logistic regression*, Proceedings of the National Academy of Sciences of the United States of America, 116 (2019), pp. 14516–14525.

[44] T. Tanaka, *A statistical-mechanics approach to large-system analysis of CDMA multiuser detectors*, IEEE Transactions on Information Theory, 48 (2002), pp. 2888–2910.

[45] E. Van Der Straeten and C. Beck, *Superstatistical fluctuations in time series: Applications to share-price dynamics and turbulence*, Physical Review E, 80 (2009), p. 036108.

[46] Y. Xu, Y. Kabashima, and L. Zdeborova, *Bayesian signal reconstruction for 1-bit compressed sensing*, Journal of Statistical Mechanics: Theory and Experiment, 2014 (2014), p. P11015.

[47] Q. Zou and H. Yang, *Replica Analysis for Generalized Linear Regression with IID Row Prior*, Sept. 2021.

A Replica trick and free energy computation

The aim of this section is to compute the quenched free energy density, which is the average with respect to the training data $\mathcal{D}$ of the logarithm of the partition function $Z(\mathcal{D})$, i.e. the normalization of the posterior distribution $p(\boldsymbol{\theta} \mid \mathcal{D})$:

$$\beta f = -\lim_{d\to\infty} \mathbb{E}_{\mathcal{D}} \log Z(\mathcal{D}).$$

As stated in section 2, the computation of this quantity can be carried out using the replica method, commonly employed in statistical physics when dealing with systems that exhibits quenched disorder. Following this approach one can get rid of the logarithm considering that

$$Z^s = e^{s \log Z} = 1 + s \log Z + O(s^2),$$

hence

$$\mathbb{E} \log Z = \lim_{s\to 0^+} \frac{\mathbb{E} Z^s - 1}{s} \tag{9}$$

The main semplification of this method comes from the fact that the average of Z^s is performed assuming s to be an integer: the quantity corresponds physically to the partition function of s *replicas* of the original system. This is easier than computing the average of $\log Z$. Nonetheless one should keep in mind that once obtained the result of this average, the limit in (9) is done sending continuously $s \to 0^+$. This non-rigorousness at the core of the replica method has still never been formally resolved, but this approach mantain his popularity since in every case where its results can be compared with exact ones, they are in perfect agreement.

A.1 Generalized linear model with superstatistical covariates

The problem is now reduced to computing the averaged of the replicated partition function. Here we show the derivation of the replica results in the more general case of "incomplete information", *i.e.* without the assumption that the probability distributions used to generate the weights and the labels coincide with the ones used in the reconstruction. The results for *Bayes optimal* estimation and empirical risk minimization will be shown in A.2 and A.3. We start with:[13]

$$\begin{aligned}
\mathbb{E}_{\mathcal{D}} Z^s &= \mathbb{E}_{\mathcal{D}} \prod_{a=1}^{s} \int \prod_{l\in[d]} \mathrm{d}\theta_l^a \, p_\theta(\theta_l^a) \prod_{i\in[n]} p_{out}\left(y_i \mid \boldsymbol{\theta}^{a\top} \mathbf{x}_i\right) \\
&= \int \prod_{l\in[d]} \mathrm{d}\theta_{*,\,l} \, p_\theta^*(\theta_{*,\,l}) \int \prod_{l\in[d]} \prod_{a=1}^{s} \mathrm{d}\theta_l^a p_\theta(\theta_l^a) \times \\
&\times \int \prod_{i\in[n]} \mathrm{d}y_i \, \mathbb{E}_{\mathbf{X}} \left[p_{out}^*\left(y_i \mid \boldsymbol{\theta}_*^\top \mathbf{x}_i\right) \prod_{a=1}^{s} p_{out}\left(y_i \mid \boldsymbol{\theta}^{a\top} \mathbf{x}_i\right) \right]
\end{aligned} \tag{10}$$

Where we explicited $\mathbb{E}_{\mathcal{D}}$ as $\mathbb{E}_{\mathbf{X}}\mathbb{E}_{\mathbf{y}} = \mathbb{E}_{\mathbf{X}}\mathbb{E}_{\boldsymbol{\theta}_*}\mathbb{E}_{\mathbf{y}|\mathbf{X}\grave{}_*}$. We then manage to decouple the dependence on the labels and the weights in the previous expression introducing *local fields* $\{h_i^a\}$, $i \in [n]$, $a = 0, \ldots, s$. This is done considering that

$$1 = \prod_{i\in[n]} \left[\int \mathrm{d}h_i^0 \delta\left(h_i^0 - \boldsymbol{\theta}_*^\top \mathbf{x}_i\right) \prod_{a=1}^{s} \int \mathrm{d}h_i^a \delta\left(h_i^a - \boldsymbol{\theta}^{a\top} \mathbf{x}_i\right) \right]$$

[13] Every integral presented in this work is a definite integral, the notation $\int$ has to be intended as a convention for the integration over $(-\infty, +\infty)$, *e.g.* $\int \mathrm{d}s\mathrm{d}t f(s,t) = \int_{\mathbb{R}^2} \mathrm{d}s\mathrm{d}t f(s,t)$

and rewriting the expected value in (10) as

$$\int \mathrm{d}h_i^0 dp_{out}^*\left(y_i \mid h_i^0\right) \int \prod_{a=1}^{s} \mathrm{d}h_i^a p_{out}\left(y_i \mid h_i^a\right) \mathbb{E}_{\mathbf{X}}\left[\delta\left(h_i^0 - \boldsymbol{\theta}_*^\top \mathbf{x}_i\right) \prod_{a=1}^{s} \delta\left(h_i^a - \boldsymbol{\theta}^{a\top} \mathbf{x}_i\right)\right]. \tag{11}$$

The expected value in (11) defines the joint density over the local fields $\{h_i^a\}$, $i \in [n]$, $a = 0, \dots, s$. Considering for simpicity $\boldsymbol{\mu} = \mathbf{0}$ in (1), it is easy to verify that - at fixed Δ - these are Gaussian variables with zero mean and covariance given by the matrix $\boldsymbol{q}_\Delta$, whose elements q_Δ^{ab}, which we will also call (rescaled) *overlap parameters*, are defined as follows: [14]

$$\begin{aligned}\mathbb{E}_{\mathbf{X}|\boldsymbol{\Delta}} h_i^a h_j^b &= \mathbb{E}_{\mathbf{X}|\Delta}\left[\boldsymbol{\theta}^{a\top}\mathbf{x}_i \boldsymbol{\theta}^{b\top}\mathbf{x}_j\right] = \boldsymbol{\theta}^{a\top}\mathbb{E}_{\mathbf{X}|\Delta}\left[\mathbf{x}_i \mathbf{x}_j^\top\right]\boldsymbol{\theta}^b \\ &= \frac{\Delta}{d}\boldsymbol{\theta}^{a\top}\mathbb{I}_d\,\boldsymbol{\theta}^b = \frac{\Delta}{d}\boldsymbol{\theta}^{a\top}\boldsymbol{\theta}^b =: q_\Delta^{ab}.\end{aligned}$$

It is possible to see why the structure for the distribtion of the covariates $\mathbf{x}_i$ was chosen as in (1): conditioning on the value of Δ, it is possible to exploit the fact that $\mathbb{P}(\mathbf{x} \mid \Delta)$ is a Gaussian density function to simplify the computations by following a procedure similar to the case with Gaussian covariates. The only price to pay will be the (numerical) average over $\rho(\Delta)$ that remains explicit. At this point, using the notation $\mathbf{h}_i = (h_i^0, h_i^1, \dots, h_i^s)$ and the shorthand $\mathrm{d}\mathbf{q} = \prod_{0\le a\le b\le s} \mathrm{d}q^{ab}$ for any matrix $\mathbf{q} \in \mathbb{R}^{s\times s}$, we can rewrite (11) as

$$\begin{aligned}&\int \mathrm{d}h_i^0 dp_{out}^*\left(y_i \mid h_i^0\right) \int \prod_{a=1}^{s} \mathrm{d}h_i^a p_{out}\left(y_i \mid h_i^a\right) \mathcal{N}\left(\mathbf{h}_i \mid 0, \mathbf{q}_\Delta\right) \\ &= \int \mathrm{d}h_i^0 dp_{out}^*\left(y_i \mid h_i^0\right) \int \prod_{a=1}^{s} \mathrm{d}h_i^a p_{out}\left(y_i \mid h_i^a\right) \times \\ &\times \int \mathrm{d}\mathbf{q}_\Delta \prod_{0\le a\le b\le s} \delta\left(q_\Delta^{ab} - \frac{\Delta}{d}\boldsymbol{\theta}^{a\top}\boldsymbol{\theta}^b\right) \mathcal{N}\left(\mathbf{h}_i \mid 0, \mathbf{q}_\Delta\right),\end{aligned} \tag{12}$$

treating now the overlap parameters as variables and expliciting their connection to the weights through the Dirac's deltas. Using their Fourier representation

$$\delta\left(q_\Delta^{ab} - \frac{\Delta}{d}\boldsymbol{\theta}^{a\top}\boldsymbol{\theta}^b\right) = \int_{-\mathrm{i}\infty}^{\mathrm{i}\infty} \mathrm{d}\hat{q}^{ab} e^{\hat{q}^{ab}\left(\boldsymbol{\theta}^{a\top}\boldsymbol{\theta}^b - \frac{d}{\Delta}q_\Delta^{ab}\right)} = \int_{-\mathrm{i}\infty}^{\mathrm{i}\infty} \mathrm{d}\hat{q}^{ab} e^{-\frac{d}{\Delta}\hat{q}^{ab}q_\Delta^{ab} + \hat{q}^{ab}\boldsymbol{\theta}^{a\top}\boldsymbol{\theta}^b} \tag{13}$$

Putting together (12) and (13), (10) now becomes:

$$\begin{aligned}\mathbb{E}_{\mathcal{D}} Z^s &= \mathbb{E}_\Delta \int \mathrm{d}\mathbf{q}\mathrm{d}\hat{\mathbf{q}} \exp\left(-\frac{d}{\Delta}\operatorname{tr}\mathbf{q}_\Delta\hat{\mathbf{q}}^\top + \log I_\theta^{(d)}(\hat{\mathbf{q}}) + \log I_y^{(n)}(\mathbf{q})\right) \\ &= \mathbb{E}_\Delta \int \mathrm{d}\mathbf{q}\mathrm{d}\hat{\mathbf{q}} \exp\left(d\left(-\frac{\operatorname{tr}\mathbf{q}_\Delta\hat{\mathbf{q}}^\top}{\Delta} + \log I_\theta(\hat{\mathbf{q}}) + \alpha\log I_y(\mathbf{q})\right)\right) \\ &=: \mathbb{E}_\Delta \int \mathrm{d}\mathbf{q}\mathrm{d}\hat{\mathbf{q}}\, e^{d\Phi(\mathbf{q},\hat{\mathbf{q}})},\end{aligned} \tag{14}$$

where we have used $\alpha = n/d$ and defined

$$I_\theta^{(d)}(\hat{\mathbf{q}}) := \int \prod_{l\in[d]} \mathrm{d}\theta_{*,l}\, p_\theta^*(\theta_{*,l}) \int \prod_{l\in[d]}\prod_{a=1}^{s} \mathrm{d}\theta_l^a p_\theta(\theta_l^a)\, \exp \sum_{0\le a\le b\le s} \hat{q}^{ab}\boldsymbol{\theta}^{a\top}\boldsymbol{\theta}^b$$

[14] In order to simplify the notation in the following expessions, we will consider $\boldsymbol{\theta}^0$ as equivalent to $\boldsymbol{\theta}_*$

$$= \prod_{l\in[d]} \int \mathrm{d}\theta_{*,\,l}\, p^*_\theta(\theta_{*,\,l}) \int \prod_{a=1}^{s} \mathrm{d}\theta_l^a p_\theta(\theta_l^a)\, \exp \sum_{0\le a\le b\le s} \hat{q}^{ab} {\theta_l}^{a\top} \theta_l{}^b =: [I_\theta(\hat{\mathbf{q}})]^d$$

$$I_y^{(n)}(\mathbf{q}) := \int \prod_{i\in[n]} \mathrm{d}y_i \int \mathrm{d}h_i^0 dp^*_{out}\left(y_i \mid h_i^0\right) \int \prod_{i\in[n]} \prod_{a=1}^{s} \mathrm{d}h_i^a p_{out}\left(y_i \mid h_i^a\right) \mathcal{N}\left(\mathbf{h}_i \mid 0, \mathbf{q}_\Delta\right)$$

$$= \prod_{i\in[n]} \int \mathrm{d}y_i \int \mathrm{d}h_i^0 dp^*_{out}\left(y_i \mid h_i^0\right) \int \prod_{a=1}^{s} \mathrm{d}h_i^a p_{out}\left(y_i \mid h_i^a\right) \mathcal{N}\left(\mathbf{h}_i \mid 0, \mathbf{q}_\Delta\right) =: [I_y(\mathbf{q})]^n.$$

In eq. (14) we managed to write $\mathbb{E}_{\mathcal{D}} Z^s$ as a saddle point problem in the dimension d, which allow us to easily evaluate the integral. In fact, assuming

$$\lim_{d\to\infty} \lim_{s\to 0^+} \frac{\mathbb{E}_{\mathcal{D}} Z^s - 1}{sd} = \lim_{s\to 0^+} \lim_{d\to\infty} \frac{\mathbb{E}_{\mathcal{D}} Z^s - 1}{sd},$$

we have that, as $d \to \infty$ the integral in (14) is dominated by the configurations $(\mathbf{q},\, \hat{\mathbf{q}})$ that extremize the function Φ.

Replica symmetric ansatz

In order to find these extremizing configurations , the usual procedure is to restrict the search to the subset of *replica symmetric* solutions, *i.e.*

$$\begin{aligned}
&q_\Delta^{00} = r_\Delta^0 && \hat{q}^{00} = \hat{r}^0 && \\
&q_\Delta^{0a} = m_\Delta && \hat{q}^{0a} = \hat{m} && \text{for } 1 \le a \le s \\
&q_\Delta^{aa} = r_\Delta && \hat{q}^{aa} = -\frac{1}{2}\hat{r} && \text{for } 1 \le a \le s \\
&q_\Delta^{ab} = q_\Delta && \hat{q}^{ab} = \hat{q} && \text{for } 1 \le a < b \le s.
\end{aligned}$$

The factor $-1/2$ for the parametrization of $\hat{q}^{aa}$ is chosen for convenience, but it will not change the result, since we will extremize over $\hat{r}$.
This parametrization correspond to the assumption of invariance of the solution with respect to the permutation of repica indices. Moreover, in Bayes optimal setting we are sure it holds. In subsection B is shown that if $\mathbf{q}_\Delta$ is replica symmetryc, the same holds for $\mathbf{q}_\Delta^{-1}$. We will refer to its elements as:

$$\begin{aligned}
&\left(q_\Delta^{-1}\right)^{00} = \tilde{r}^0 && \\
&\left(q_\Delta^{-1}\right)^{0a} = \tilde{m} && \text{for } 1 \le a \le s \\
&\left(q_\Delta^{-1}\right)^{aa} = \tilde{r} && \text{for } 1 \le a \le s \\
&\left(q_\Delta^{-1}\right)^{ab} = \tilde{q} && \text{for } 1 \le a < b \le s.
\end{aligned}$$

On this ansatz, the trace term in $\Phi(\mathbf{q}, \hat{\mathbf{q}})$ becomes:

$$\sum_{0\le a\le b\le s} q^{ab}\hat{q}^{ab} = r_\Delta^0 \hat{r}^0 + s m_\Delta \hat{m} - \frac{1}{2} s r_\Delta \hat{r} + \frac{s(s-1)}{2} q_\Delta \hat{q}.$$

Similarly we have that the exponent in I_θ becomes:

$$\begin{aligned}\sum_{0\le a\le b\le s}\hat{q}^{ab}\theta^a\theta^b &= \hat{r}^0\theta_*^2+\hat{m}\theta_*\sum_{a=1}^{s}\theta^a-\frac{1}{2}\hat{r}\sum_{a=1}^{s}(\theta^a)^2+\hat{q}\sum_{1\le a<b\le s}\theta^a\theta^b\\ &= \hat{r}^0\theta_*^2+\hat{m}\theta_*\sum_{a=1}^{s}\theta^a-\frac{1}{2}(\hat{r}+\hat{q})\sum_{a=1}^{s}(\theta^a)^2+\frac{1}{2}\hat{q}\sum_{a,b=1}^{s}\theta^a\theta^b.\end{aligned}$$

As customary, we decouple different replica indices introducing an Hubbard-Stratonovich field $\epsilon\sim\mathcal{N}(\epsilon\mid 0,1)$, namely:

$$e^{\frac{\hat{q}}{2}\sum_{a,b=1}^{s}\theta^a\theta^b}=\mathbb{E}_\epsilon e^{\sqrt{\hat{q}}\epsilon\sum_{a=1}^{s}\theta^a}.$$

Putting together, we have that

$$\begin{aligned}I_\theta &= \int \mathrm{d}\theta_* p_\theta^*(\theta_*)\int\prod_{a=1}^{s}\mathrm{d}x^a p_\theta(\theta^a)e^{\sum_{0\le a\le b\le s}\hat{q}^{ab}\theta^a\theta^b}\\ &= \int \mathrm{d}\theta_* p_\theta^*(\theta_*)e^{\hat{r}^0\theta_*^2}\mathbb{E}_\epsilon\int\prod_{a=1}^{s}\mathrm{d}x^a p_\theta(\theta^a)e^{\hat{m}\theta_*\theta^a-\frac{1}{2}(\hat{r}+\hat{q})(\theta^a)^2+\sqrt{\hat{q}}\epsilon\theta^a}\\ &= \mathbb{E}_\epsilon\int \mathrm{d}\theta_* p_\theta^*(\theta_*)e^{\hat{r}^0\theta_*^2}\left(\int \mathrm{d}x\, p_\theta(\theta)e^{\hat{m}\theta_*\theta-\frac{1}{2}(\hat{r}+\hat{q})(\theta)^2+\sqrt{\hat{q}}\epsilon\theta}\right)^s,\end{aligned}$$

where we were able to remove the dependence of our expression on replica indices and explicit the dependance on s.
We can repeat a similar procedure for the term in I_y. Recalling that

$$\mathcal{N}(\mathbf{h}\mid 0,\mathbf{q}_\Delta)=\frac{e^{-\frac{1}{2}\sum_{0\le a\le b\le s}h^a\left(\mathbf{q}_\Delta^{-1}\right)^{ab}h^b}}{\sqrt{\det 2\pi\mathbf{q}_\Delta}},$$

it is easy to see that

$$\begin{aligned}-\frac{1}{2}\sum 0\le a\le b\le s h^a\left(\mathbf{q}_\Delta^{-1}\right)^{ab}h^b &= -\frac{1}{2}\tilde{r}^0(h^0)^2-\tilde{m}h^0\sum_{a=1}^{s}h^a-\frac{1}{2}\tilde{r}\sum_{a=1}^{s}(h^a)^2-\tilde{q}\sum_{1\le a<b\le s}h^ah^b\\ &= -\frac{1}{2}\tilde{r}^0(h^0)^2-\tilde{m}h^0\sum_{a=1}^{s}h^a-\frac{1}{2}(\tilde{r}-\tilde{q})\sum_{a=1}^{s}(h^a)^2-\frac{1}{2}\tilde{q}\sum_{a,b=1}^{s}h^ah^b.\end{aligned}$$

Introducing another Hubbard-Stratonovich field $\eta\sim\mathcal{N}(\eta\mid 0,1)$ in order to decouple replica indices in the last expression,

$$\begin{aligned}I_y &= \int \mathrm{d}y\int \mathrm{d}h^0 dp_{out}^*\left(y\mid h^0\right)\int\prod_{a=1}^{s}\mathrm{d}h^a p_{out}\left(y\mid h^a\right)\mathcal{N}\left(\mathbf{h}\mid 0,\mathbf{q}_\Delta\right)\\ &= A_s(r^0,m,r,q)\int \mathrm{d}h^0 dp_{out}^*\left(y\mid h^0\right)e^{-\frac{1}{2}\tilde{r}^0(h^0)^2}\mathbb{E}_\eta\int\prod_{a=1}^{s}\mathrm{d}h^a p_{out}\left(y\mid h^a\right)e^{-\tilde{m}h^0h^a-\frac{1}{2}(\tilde{r}-\tilde{q})(h^a)^2+\sqrt{-\tilde{q}}\eta h^a}\\ &= A_s(r^0,m,r,q)\mathbb{E}_\eta\int \mathrm{d}h^0 dp_{out}^*\left(y\mid h^0\right)e^{-\frac{1}{2}\tilde{r}^0(h^0)^2}\left(\int \mathrm{d}h p_{out}\left(y\mid h\right)e^{-\tilde{m}h^0h-\frac{1}{2}(\tilde{r}-\tilde{q})(h)^2+\sqrt{-\tilde{q}}\eta h}\right)^s.\end{aligned}$$

In the previous we have defined $A_s(r^0,m,r,q)=\left(\det 2\pi\mathbf{q}_\Delta\right)^{-1/2}$ computed on the replica symmetric ansatz. His explicit definition can be found in B.
In order to lighten the notation, the subscript Δ in $\mathbf{q}_\Delta$ and its element will not be written in the remaining part of this appendix. The dependence on Δ will be reminded and made explicit in the final result.

Taking the $s \to 0^+$ limit

At this point, to proceed with the replica approach, one needs to take the limit $s \to 0^+$ in (9). Before doing it, we should check that the function Φ in (14) does not contain any term $O(1)$ with respect to s, otherwise our quantity of interest will diverge.
We start considering

$$\begin{aligned}\lim_{s\to 0} \log A_s &= \lim_{n\to 0}\left[-\frac{s+1}{2}\log 2\pi - \frac{s-1}{2}\log(r-q) - \frac{1}{2}\log\left(rr^0 + (s-1)r^0 q - sm^2\right)\right] \\ &= -\frac{1}{2}\log 2\pi r^0,\end{aligned}$$

which implies

$$\begin{aligned}\lim_{s\to 0} \log I_y &= -\frac{1}{2}\log 2\pi r^0 + \log \int \mathrm{d}y \int \mathrm{D}\eta \int \mathrm{d}h^0 p^*_{\text{out}}\left(y \mid h^0\right) e^{-\frac{1}{2}\tilde{r}^0(s=0)\left(h^0\right)^2} \\ &= -\frac{1}{2}\log 2\pi r^0 + \log \int \mathrm{d}h^0 e^{-\frac{1}{2r^0}\left(h^0\right)^2} \\ &= -\frac{1}{2}\log 2\pi r^0 + \frac{1}{2}\log 2\pi r^0 = 0,\end{aligned} \tag{15}$$

where in the first equality we used the normalization $\mathbb{E}_\eta[1] = 1$ and $\int \mathrm{d}y p^*_{\text{out}}\left(y \mid h^0\right) = 1$ and on the second equality that $\tilde{r}^0 = 1/r^0$ at $s = 0$ (see appendix B). Likewise,

$$\lim_{s\to 0} \log I_\theta = \log \int \mathrm{d}\theta_* p^*_\theta\left(\theta_*\right) e^{\hat{r}^0(\theta_*)^2}.$$

Therefore

$$\lim_{s\to 0} \Phi = \frac{\hat{r}^0 r^0}{\Delta} + \log \int \mathrm{d}\theta_* p^*_\theta\left(\theta_*\right) e^{\hat{r}^0\left(x^0\right)^2}$$

which is only zero if we set $\hat{r}^0 = 0$. It is easy to check that this implies in particular that on the saddle-point we must have [15]

$$\mathbb{E}_\Delta \frac{r^0}{\Delta} = \mathbb{E}_{\theta_*}\left(\theta_*\right)^2. \tag{16}$$

Once fixed these consistency conditions, we can study the $O(s)$ terms. We could start by symmetrizing the integrals I_θ and I_y.
We let $\xi \to \xi + \hat{q}^{-1/2}\hat{m}x^0$, so that

$$I_\theta = \mathbb{E}_\xi \int \mathrm{d}\theta_* p^*_\theta\left(\theta_*\right) e^{-\frac{\hat{m}^2}{2\hat{q}}\theta_*^2 + \frac{\hat{m}}{\sqrt{\hat{q}}}\xi\theta_*}\left(\int \mathrm{d}\theta p_\theta(x) e^{-\frac{1}{2}(\hat{r}+\hat{q})\theta^2 + \sqrt{\hat{q}}\xi\theta}\right)^s,$$

therefore [16]

$$\begin{aligned}\lim_{s\to 0^+} \frac{1}{s}\log I_\theta &= \mathbb{E}_\xi \int \mathrm{d}\theta_* p^*_\theta\left(\theta_*\right) e^{-\frac{\hat{m}^2}{2\hat{q}}\theta_*^2 + \frac{\hat{m}}{\sqrt{\hat{q}}}\xi\theta_*} \log \int \mathrm{d}\theta p_\theta(\theta) e^{-\frac{1}{2}(\hat{r}+\hat{q})\theta^2 + \sqrt{\hat{q}}\xi\theta} \\ &=: \mathbb{E}_\xi\left[I^{(0)}_\theta(\xi)\log I^{(1)}_\theta(\xi)\right].\end{aligned}$$

[15] Making explicit the dependence of r^0 on Δ we notice that the first part of this equality does not depend on Δ

[16] In the following we are using $\lim_{s\to 0}\log \mathbb{E} f g^s = \lim_{s\to 0}\log \mathbb{E}(f + fs\log g)$, then we expand the external logarithm.

A similar but longer procedure can be applied to I_y. We perform the change of variable $\eta \to \eta - (-\tilde{q})^{-1/2}\tilde{m}h^0$, so that

$$\log I_Y = \log A_s + \underbrace{\log \mathbb{E}_\eta \int \mathrm{d}y g_0^{(s)}(y,\eta) g_1^{(s)}(y,\eta)^s}_{(\star)},$$

where we have defined

$$g_0^{(s)}(y,\eta) = \int \mathrm{d}h^0 p^*_{\text{out}}\ \left(y \mid h^0\right) e^{-\frac{1}{2}\left(\tilde{r}^0 - \frac{\tilde{m}^2}{\tilde{q}}\right)\left(h^0\right)^2 - \frac{\tilde{m}}{\sqrt{-\tilde{q}}}\eta h^0}$$
$$g_1^{(s)}(y,\eta) = \int \mathrm{d}h p_{\text{out}}\ (y \mid h) e^{-\frac{1}{2}(\tilde{r}-\tilde{q})h^2 + \sqrt{-\tilde{q}}\eta h}$$

We remind that one needs to be careful in taking the limit $s \to 0^+$ since $\left(\tilde{r}^0, \tilde{m}, \tilde{r}, \tilde{q}\right)$ all depend on s.

$$\begin{aligned}
(\star) &= \log \mathbb{E}_\eta \int \mathrm{d}y \left[g_0^{(0)}(y,\eta) + \left. s\partial_n g_0^{(s)}(y,\eta)\right|_{s=0} + s g_0^{(0)}(y,\eta) \log g_1^{(0)}(y,\eta) + O\left(n^2\right)\right] \\
&= \log \left[\sqrt{2\pi r^0} + s \int \mathrm{d}y \mathbb{E}_\eta \left(\left. \partial_n g_0^{(n)}(y,\eta)\right|_{s=0} + g_0^{(0)}(y,\eta) \log g_1^{(0)}(y,\eta)\right) + O\left(s^2\right)\right] \\
&= -\frac{1}{2}\log 2\pi r^0 + \frac{s}{\sqrt{2\pi r^0}} \int \mathrm{d}y \mathbb{E}_\eta \left[\left. \partial_s g_0^{(s)}(y,\eta)\right|_{s=0} + g_0^{(0)}(y,\eta) \log g_1^{(0)}(y,\eta)\right] + O\left(s^2\right),
\end{aligned}$$

where we have used the zeroth order result from (15). The first of the integrals is easy to evaluate. Using that $\int \mathrm{D}\eta \int \mathrm{d}y g_0^{(s)}(y,\eta) = \sqrt{\frac{2\pi}{\tilde{r}^0}}$ we can exchange derivative and integral arriving to

$$\begin{aligned}
\frac{1}{\sqrt{2\pi r^0}} \int \mathrm{d}y \int \mathbb{E}_\eta \left. \partial_s g_0^{(s)}(y,\eta)\right|_{s=0} &= \frac{1}{\sqrt{2\pi r^0}} \partial_s \left[\int \mathrm{d}y \mathbb{E}_\eta g_0^{(s)}(y,\eta)\right]_{s=0} = \left. \frac{1}{\sqrt{2\pi r^0}} \partial_s \sqrt{\frac{2\pi}{\tilde{r}^0}}\right|_{s=0} \\
&= -\frac{1}{2}\left[\frac{q}{r-q} - \frac{r^0 q - m^2}{r^0(r-q)}\right]
\end{aligned}$$

Putting all together we arrive to

$$\lim_{s\to 0^+} \frac{1}{s}\log I_Y = -\frac{1}{2}\log 2\pi(r-q) - \frac{1}{2}\frac{q}{r-q} + \frac{1}{\sqrt{2\pi r^0}} \int \mathrm{d}y \int \mathrm{D}\,\eta g_0(y,\eta) \log g_1(y,\eta)$$

where we have relabelled $g_0 = g_0^{(0)}$ and $g_1 = g_1^{(0)}$, given by

$$g_0(y,\eta) = \int \mathrm{d}h^0 p^{(0)}_{\text{out}}\ \left(y \mid h^0\right) e^{-\frac{1}{2}\frac{q}{r^0 q - m^2}\left(h^0\right)^2 + \sqrt{\frac{m^2}{r^0(r^0 q - m^2)}}\eta h^0}$$
$$g_1(y,\eta) = \int \mathrm{d}h p_{\text{out}}\ (y \mid h) e^{-\frac{1}{2}\frac{1}{r-q}h^2 + \sqrt{\frac{r^0 q - m^2}{r^0 (r-q)^2}}\eta h}$$

Note in particular that $g_0 = g_1$ in the Bayes-optimal case, when $r = r^0$ and $m = q$. We can still make some cosmetic changes in order to rewrite the above in a compact form. First, we can make

$$h \to \sqrt{r-q}h + \sqrt{\frac{r^0 q - m^2}{r^0}}\eta$$

such that, introducing the notation $\int \mathrm{D}h = (2\pi)^{-1/2}\int \mathrm{d}h e^{-\frac{1}{2}h^2}$

$$g_1(y,\eta) = \sqrt{2\pi(r-q)}e^{\frac{1}{2}\frac{r^0 q - m^2}{r^0(r-q)}\eta^2}\int \mathrm{D}h p_{\mathrm{out}}\left(y \mid \sqrt{r-q}h + \sqrt{\frac{r^0 q - m^2}{r^0}}\eta\right)$$

Therefore,

$$\begin{aligned}\int \mathrm{D}\eta \int \frac{\mathrm{d}y}{\sqrt{2\pi r^0}} g_0 \log g_1 =& \frac{1}{2}\log 2\pi(r-q)\int \mathrm{D}\eta \int \frac{\mathrm{d}y}{\sqrt{2\pi r^0}} g_0(y,\eta) + \frac{1}{2}\frac{r^0 q - m^2}{r^0(r-q)}\int \mathrm{D}\eta\eta^2 \int \frac{\mathrm{d}y}{\sqrt{2\pi r^0}} g_0(y,\eta) \\ &+ \log \int \mathrm{D}\eta \int \frac{\mathrm{d}y}{\sqrt{2\pi r^0}} \int \mathrm{D}h p_{\mathrm{out}}\left(y \mid \sqrt{r-q}h - \sqrt{\frac{r^0 q - m^2}{r^0}}\right) \\ =& \frac{1}{2}\log 2\pi(r-q) + \frac{1}{2}\frac{q}{r-q} \\ &+ \int \frac{\mathrm{d}y}{\sqrt{2\pi r^0}} \int \mathrm{D}\eta g_0(y,\eta) \log \int \mathrm{D}h p_{\mathrm{out}}\left(y \mid \sqrt{r-q}h - \sqrt{\frac{r^0 q - m^2}{r^0}}\eta\right)\end{aligned}$$

where we have used that

$$\int \mathrm{D}\eta \int \frac{\mathrm{d}y}{2\pi r^0}\eta^2 g_0(y,\eta) = \sqrt{\frac{r^0 q - m^2}{2\pi r^0 q}}\int \mathrm{d}\eta\eta^2 e^{-\frac{1}{2}\left(\frac{r^0 q - m^2}{r^0 q}\right)\eta^2} = \frac{r^0 q}{r^0 q - m^2}$$

putting together and further making a rescaling of the noise,

$$\eta \to \sqrt{\frac{qr^0}{r^0 q - m^2}}\eta$$

leads to

$$\lim_{s\to 0^+}\frac{1}{s}\log I_Y = \sqrt{\frac{q}{r^0 q - m^2}}\int \frac{\mathrm{d}y}{\sqrt{2\pi}}\int \frac{\mathrm{d}\eta}{\sqrt{2\pi}} e^{-\frac{1}{2}\frac{qr^0}{r^0 q - m^2}\eta^2} g_0\left(y, \sqrt{\frac{qr^0}{r^0 q - m^2}}\eta\right)\int \mathrm{D}h p_{\mathrm{out}}\left(y \mid \sqrt{r-q}h + \sqrt{q}\eta\right)$$

Focusing now on g_0,

$$g_0\left(y, \sqrt{\frac{qr^0}{r^0 q - m^2}}\eta\right) = \int \mathrm{d}h^0 p^*_{out}(y \mid h) e^{-\frac{1}{2}\frac{q}{r^0 q - m^2}\left(h^0\right)^2 + \sqrt{\frac{m^2 q}{(r^0 q - m^2)^2}}\eta h^0}$$

we can make a final change of variables to bring the measure over h^0 to a Gaussian form,

$$h^0 \to \sqrt{\frac{r^0 q - m^2}{q}} h^0 + \frac{m}{\sqrt{q}} \eta$$

leading to

$$g_0\left(y, \sqrt{\frac{r^0 q}{r^0 q - m^2}}\eta\right) = \sqrt{2\pi \frac{r^0 q - m^2}{q}} e^{-\frac{m^2}{2(r^0 q - m^2}\eta^2} \int \mathrm{D}\, h^0 p^*_{out}\left(y \mid \sqrt{\frac{r^0 q - m^2}{q}} h^0 + \frac{m}{\sqrt{q}}\eta\right)$$

Finally, wrapping it up

$$\begin{aligned}\lim_{n\to 0} \frac{1}{n} \log I_Y &= \int \mathrm{d}y \int \mathrm{D}\eta I_Y^{(0)}(y,\eta) \log I_Y^{(1)}(y,\eta) \\ &= \mathbb{E}_\eta \left[I_Y^{(0)} \log I_Y^{(1)} \right]\end{aligned}$$

for $\eta \sim \mathcal{N}(0,1)$ and

$$\begin{aligned} I_Y^{(0)}(y,\eta) &= \int \mathrm{D}h^0 p_{\text{out}}\left(y \mid \sqrt{\frac{r^0 q - m^2}{q}} h^0 + \frac{m}{\sqrt{q}}\eta\right) \\ I_Y^{(1)}(y,\eta) &= \int \mathrm{D}h p_{\text{out}}\left(y \mid \sqrt{r-q} h + \sqrt{q}\eta\right)\end{aligned}$$

Summary

As a final step we make the dependence of the overlap parameters on Δ explicit. Recalling that $q^{ab} := \Delta d^{-1} \boldsymbol{\theta}^{a\top} \boldsymbol{\theta}^b$ and that we have not written the subscript Δ in the previous section of this appendix, with a little abuse of notation we define (r^0, m, r, q) as

$$\begin{aligned} r^0_\Delta &:= \Delta r^0 & m_\Delta &:= \Delta m \\ r_\Delta &:= \Delta r & q_\Delta &:= \Delta q \end{aligned}$$

We can finally write the free energy density as the following extremization problem:

$$\begin{aligned} \beta f &= \operatorname*{extr}_{m,r,q,\hat{m},\hat{r},\hat{q}} \phi(m,r,q,\hat{m},\hat{r},\hat{q}) \qquad (17) \\ \phi &= -m\hat{m} + \frac{1}{2} r\hat{r} + \frac{1}{2} q\hat{q} + \mathbb{E}_\epsilon \left[I_\theta^{(0)}(\epsilon) \log I_\theta^{(1)}(\epsilon) \right] + \alpha \mathbb{E}_{\Delta,\eta} \left[\int \mathrm{d}y\, I_y^{(0)}(y,\Delta,\eta) \log I_y^{(1)}(y,\Delta,\eta) \right] \end{aligned}$$

with

$$
\begin{aligned}
I_X^{(0)}(\epsilon) &= \int \mathrm{d}\theta_* p_\theta^*(\theta_*) e^{-\frac{\hat{m}^2}{2\hat{q}}\theta_*^2 + \frac{\hat{m}}{\sqrt{\hat{q}}}\epsilon\theta_*} \\
I_X^{(1)}(\epsilon) &= \int \mathrm{d}\theta p_\theta(\theta) e^{-\frac{1}{2}(\hat{r}+\hat{q})\theta^2 + \sqrt{\hat{q}}\epsilon\theta} \\
I_Y^{(0)}(y,\eta) &= \int \mathrm{D}h^0 p_{\text{out}}\left(y \mid \sqrt{\Delta\frac{r^0 q - m^2}{q}}h^0 + \frac{\sqrt{\Delta}m}{\sqrt{q}}\eta\right) \\
I_Y^{(1)}(y,\eta) &= \int \mathrm{D}h p_{\text{out}}\left(y \mid \sqrt{\Delta(r-q)}h + \sqrt{\Delta q}\eta\right)
\end{aligned}
\tag{18}
$$

and

$$
r^0 = \mathbb{E}_{\theta *}\left[\theta_*^2\right]
$$

A.2 Bayes optimal setting

In this section we simplify the results of A.1 and compute the saddle point equations for the overlap parameters by considering the Bayes optimal setting, meaning $p_{out}(y|z) = p^*_{out}(y|z)$ and $p_\theta(\theta) = p^*_\theta(\theta)$. This implies in particular $r^0 = r$, $m = q$, $\hat{r}^0 = \hat{r} = 0$ and $\hat{m} = \hat{q}$; as a consequence $I_\theta^{(0)} = I_\theta^{(1)}$ and $I_y^{(0)} = I_y^{(1)}$.
All cases considered in this works assume Gaussian prior $p^*_\theta(\theta) = \mathcal{N}(\theta \mid 0, 1)$, which means

$$
I_\theta^{(0)} = \frac{1}{\sqrt{2\pi}}\int \mathrm{d}\theta \exp\left(-\frac{1}{2}(1+\hat{q})\theta^2 + \sqrt{\hat{q}}\epsilon\theta\right) = \frac{1}{\sqrt{1+\hat{q}}}\exp\frac{\hat{q}\epsilon^2}{2(1+\hat{q})}.
$$

Therefore

$$
\mathbb{E}_\epsilon\left[I_\theta^{(0)}(\epsilon)\log I_\theta^{(0)}(\epsilon)\right] = \frac{1}{\sqrt{2\pi(1+\hat{q})}}\int \mathrm{d}\epsilon e^{-\frac{1}{2}(1-\frac{\hat{q}}{1+\hat{q}})\epsilon^2}\left[\frac{\hat{q}\epsilon^2}{2(1+\hat{q})} - \frac{1}{2}\log(1+\hat{q})\right] = \frac{\hat{q} - \log(1+\hat{q})}{2}
$$

Hence, the function ϕ in (17) becomes

$$
\phi(q,\hat{q}) = -\frac{1}{2}q\hat{q} + \frac{\hat{q} - \log(1+\hat{q})}{2} + \alpha\mathbb{E}_{\Delta,\eta}\left[\int \mathrm{d}y\, I_y^{(0)}(y,\eta)\log I_y^{(0)}(y,\eta)\right], \tag{19}
$$

and the extremization with respect to $\hat{q}$ is given by the following *saddle point equation*

$$
\frac{\partial}{\partial\hat{q}}\phi = -\frac{1}{2}\left(q - 1 + \frac{1}{1+\hat{q}}\right) \overset{!}{=} 0 \implies q = \frac{\hat{q}}{1+\hat{q}}. \tag{20}
$$

This form for the prior also implies

$$
r^0 = r = \mathbb{E}_{\theta *}[\theta_*^2] = 1. \tag{21}
$$

In order to perform the extremization with respect to q, we need to specify $p^*_{out}(y|z)$.

Linear channel

The linear output channel corresponds to

$$y_i = \boldsymbol{\theta}_*^\top \mathbf{x}_i + \sigma\boldsymbol{\xi}, \quad \boldsymbol{\xi} \sim \mathcal{N}\left(\boldsymbol{\xi} \mid \mathbf{0}, \mathbb{I}_n\right).$$

In this case $p^*_{out}(y \mid z) = \mathcal{N}(y \mid z, \sigma^2)$. Plugging it into (18) we have

$$\begin{aligned} I_y^{(0)} &= \frac{1}{2\pi\sigma}\int \mathrm{d}h \exp\left(-\frac{1}{2}h^2 - \frac{1}{2\sigma^2}\left(y - \sqrt{\Delta(1-q)}h - \sqrt{\Delta q}\eta\right)^2\right) \\ &= \frac{1}{\sqrt{2\pi(\sigma^2 + \Delta(1-q))}} \exp\frac{-(y-\sqrt{\Delta q}\eta)^2}{2(\sigma^2+\Delta(1-q))}. \end{aligned}$$

Therefore, using $\mathbb{E}_\eta[k] = k, \forall k \in \mathbb{R}$ and the following

$$\begin{aligned} \int \mathrm{d}y\, I_y^{(0)}(y,\eta)\log I_y^{(0)}(y,\eta) &= \int \mathrm{d}y\, \mathcal{N}\left(y \mid \sqrt{\Delta q}\eta, \sigma^2 + \Delta(1-q)\right)\frac{-(y-\sqrt{\Delta q}\eta)^2}{2(\sigma^2+\Delta(1-q))} - \frac{1}{2}\log(\sigma^2+\Delta(1-q)) \\ &= -\frac{1+\log(\sigma^2+\Delta(1-q))}{2}, \end{aligned}$$

we can write (19) as

$$\phi(q,\hat{q}) = -\frac{1}{2}q\hat{q} + \frac{\hat{q} - \log(1+\hat{q})}{2} - \frac{\alpha}{2} - \frac{\alpha}{2}\mathbb{E}_\Delta \log(\sigma^2 + \Delta(1-q))$$

Then the second saddle point equation for the linear channel case is given by:

$$\frac{\partial}{\partial q}\phi = -\frac{1}{2}\left(\hat{q} - \alpha\mathbb{E}_\Delta \frac{\Delta}{\sigma^2+\Delta(1-q)}\right) \overset{!}{=} 0 \implies \hat{q} = \alpha\mathbb{E}_\Delta\frac{\Delta}{\sigma^2+\Delta(1-q)} \tag{22}$$

Probit channel

The *probit* channel corresponds to

$$y_i = \mathrm{sign}\left(\boldsymbol{\theta}_*^\top \mathbf{x}_i + \sigma\boldsymbol{\xi}\right), \quad \boldsymbol{\xi} \sim \mathcal{N}\left(\boldsymbol{\xi} \mid \mathbf{0}, \mathbb{I}_n\right).$$

In this case $p^*_{out}(y|z) = \mathrm{erfc}\left(\frac{-yz}{\sqrt{2}\sigma}\right)/2,\ y \in \{-1,1\}$. [17] Notice that in this classification case the integral $\int \mathrm{d}y$ in (17) should be replaced with $\sum_{y=\pm1}$. Plugging this into (18) we have that

$$I_y^{(0)} = \frac{1}{2\sqrt{2\pi}}\int \mathrm{d}h e^{-\frac{1}{2}h^2}\,\mathrm{erfc}\left(-\sqrt{\Delta}y\frac{\sqrt{1-q}h + \sqrt{q}\eta}{\sqrt{2}\sigma}\right) = \frac{1}{2}\,\mathrm{erfc}\left(-y\eta\sqrt{\frac{\Delta q}{2(\sigma^2+\Delta(1-q))}}\right)$$

where we have used $y^2 = 1$ and the following:

$$\begin{aligned} &\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}\mathrm{d}h e^{-\frac{1}{2}h^2}\,\mathrm{erfc}\left(\frac{Ah+B}{\sqrt{2}}\right) = \frac{\sqrt{2}}{\pi}\int_{-\infty}^{\infty}\mathrm{d}h\int_{\frac{Ah+B}{\sqrt{2}}}^{\infty}\mathrm{d}t e^{-\frac{1}{2}h^2 - t^2} \\ &= \frac{1}{\pi}\int_{-\infty}^{\infty}\mathrm{d}h\int_0^{\infty}\mathrm{d}t e^{-\frac{1}{2}(h^2 - (Ah+B+t)^2)} = \sqrt{\frac{2}{\pi(1+A^2)}}\int_0^{\infty}\mathrm{d}t e^{-\frac{(B+t)^2}{2(1+A^2)}} \end{aligned} \tag{23}$$

[17] $\mathrm{erfc}(x) = \frac{2}{\pi}\int_x^\infty \mathrm{d}t e^{-t^2}$ is the *complementary error function*.

$$= \frac{2}{\sqrt{\pi}} \int_{\frac{B}{\sqrt{2(1+A^2)}}}^{\infty} \mathrm{d}t e^{-t^2} = \mathrm{erfc}\left(\frac{B}{\sqrt{2(1+A^2)}}\right)$$

(in the second line we performed the change of variable $t \to (Ah + B + t)/\sqrt{2}$, while in the third line $t \to \sqrt{2(1+A^2)}t - B$).

Therefore, after considering that, since $I_y^{(0)}(y, \Delta, \eta)$ is a function of the product $y\eta$,

$$\begin{aligned}\mathbb{E}_\eta[I_y^{(0)}(y,\Delta,\eta)] &= \frac{1}{\sqrt{2\pi}} \int \mathrm{d}\eta e^{-\frac{1}{2}\eta^2} I_y^{(0)}(y,\Delta,\eta) \\ &\overset{\eta\to-\eta}{=} \frac{1}{\sqrt{2\pi}} \int \mathrm{d}\eta e^{-\frac{1}{2}\eta^2} I_y^{(0)}(y,\Delta,-\eta) = \mathbb{E}_\eta[I_y^{(0)}(-y,\Delta,\eta)],\end{aligned}$$

we have that

$$\mathbb{E}_\eta \sum_{y=\pm 1} I_y^{(0)}(y,\Delta,\eta) = 2\mathbb{E}_\eta I_y^{(0)}(-1,\Delta,\eta)$$

and, defining $k_q(\Delta) := \sqrt{\Delta q}/\sqrt{\sigma^2 + \Delta(1-q)}$, (19) becomes [18]

$$\phi(q,\hat{q}) = -\frac{1}{2}q\hat{q} + \frac{\hat{q} - \log(1+\hat{q})}{2} + \alpha \mathbb{E}_{\Delta,\eta}\left[\mathrm{erfc}\left(\eta\frac{k_q(\Delta)}{\sqrt{2}}\right)\log\mathrm{erfc}\left(\eta\frac{k_q(\Delta)}{\sqrt{2}}\right)\right] - \alpha\log 2 \quad (24)$$

It is possible now to derive the saddle point equation from the extremization with respect to q [19, 20]:

$$\begin{aligned}\hat{q} &= -2\alpha\frac{\partial}{\partial q}\mathbb{E}_{\Delta,\eta}\left[\mathrm{erfc}\left(\eta\frac{k_q(\Delta)}{\sqrt{2}}\right)\log\mathrm{erfc}\left(\eta\frac{k_q(\Delta)}{\sqrt{2}}\right)\right] \\ &= -2\alpha\mathbb{E}_{\Delta,\eta}\left[\frac{\partial}{\partial q}\left(\mathrm{erfc}\left(\eta\frac{k_q(\Delta)}{\sqrt{2}}\right)\right)\left(\log\mathrm{erfc}\left(\eta\frac{k_q(\Delta)}{\sqrt{2}}\right)+1\right)\right] \\ &= \frac{\sqrt{2}}{\sqrt{\pi}}\alpha\mathbb{E}_{\Delta,\eta}\left[e^{-\frac{k_q(\Delta)^2}{2}\eta^2}\eta\frac{\partial k_q(\Delta)}{\partial q}\left(\log\mathrm{erfc}\left(\eta\frac{k_q(\Delta)}{\sqrt{2}}\right)+1\right)\right] \\ &= \frac{\sqrt{2}\alpha}{\sqrt{\pi}}\mathbb{E}_{\Delta,\eta}\left[\frac{k_q(\Delta)}{2q}(1+k_q(\Delta)^2)e^{-\frac{k_q(\Delta)^2}{2}\eta^2}\eta\log\mathrm{erfc}\left(\eta\frac{k_q(\Delta)}{\sqrt{2}}\right)\right] \\ &= \frac{\alpha}{q\sqrt{2\pi}}\mathbb{E}_{\Delta,\eta}\left[k_q(\Delta)(1+k_q(\Delta)^2)e^{-\frac{k_q(\Delta)^2}{2}\eta^2}\eta\log\mathrm{erfc}\left(\eta\frac{k_q(\Delta)}{\sqrt{2}}\right)\right]\end{aligned} \quad (25)$$

The previous equation appears more complicated than the corresponding one for the linear channel (22), nonetheless we were able to reduce our high-dimensional problem to a numerically tractable one.

A.3 Empirical risk minimization

In this section we simplify the results of A.1 and compute the saddle point equation for the overlap parameters in the case of empirical risk minimization. As we have seen in (5), we can define distributions p_{out} and p_θ that play the role of likelihood and prior in the computations performed in A.1. Given their association with the selection of loss and regularization functions, we will address the cases separately, considering various forms that can be employed for these functions. One difference here is the presence of the parameter β and the additional limit β to infinity.

[18] We have used the identity found in (23) to say that $\mathbb{E}_\eta \mathrm{erfc}(\eta\frac{k_q(\Delta)}{\sqrt{2}})\log\frac{1}{2} = -\log 2$

[19] $\frac{\mathrm{d}}{\mathrm{d}x}\mathrm{erfc}(x) = -2\exp(-x^2/2)/\sqrt{\pi}$

[20] $\frac{\partial}{\partial q}k_q = \frac{1}{2k_q}\left(\frac{\Delta}{\sigma^2+\Delta(1-q)} + \frac{\Delta^2 q}{(\sigma^2+\Delta(1-q))^2}\right) = \frac{k_q}{2q}(1+k_q^2)$

Ansatz for the overlaps

In all the cases we are going to consider we are going to assume the following ansatz, that defines the quantities $(\chi, \hat{M}, \hat{R}, \hat{\chi})$ independent of β:

$$\chi = \beta(r-q) \qquad \hat{m} = \beta\hat{M} \qquad \hat{r} = \beta\hat{R} - \beta^2\hat{\chi} \qquad \hat{q} = \beta^2\hat{\chi}. \tag{26}$$

The necessity of these decision will be clear in the next subsections. [21]
With this ansatz the trace term becomes

$$\lim_{\beta\to\infty} \frac{1}{\beta}\left(-m\hat{m} + \frac{1}{2}r\hat{r} + \frac{1}{2}q\hat{q}\right) = \lim_{\beta\to\infty}\left(-m\hat{M} + \frac{1}{2}r\hat{R} - \frac{\beta}{2}\hat{\chi}(r-q)\right) \tag{27}$$

$$= -m\hat{M} + \frac{1}{2}r\hat{R} - \frac{1}{2}\chi\hat{\chi}, \tag{28}$$

and, remembering also that $r^0 = 1$ (see (21)), the expression for the MSE in (8) can be written as

$$\varepsilon_{\text{est}} = r^0 - 2m + q = 1 - 2m + r - \frac{\chi}{\beta} \overset{\beta\to\infty}{=} 1 - 2m + r. \tag{29}$$

L_2 regularization

A popular choice for the regularization term is $r(\theta) = \theta^2/2$ [22], that corresponds to

$$p_\theta(\theta) = \exp\left(-\frac{\beta\lambda}{2}\theta^2\right),$$

as defined in (5). Plugging it into the expression for $I_\theta^{(1)}$ in (18) we obtain

$$I_\theta^{(1)} = \int \mathrm{d}\theta \exp\left(-\frac{\beta\lambda}{2}\theta^2 - \frac{\hat{r}+\hat{q}}{2}\theta^2 + \sqrt{\hat{q}}\epsilon\theta\right) = \sqrt{\frac{2\pi}{\hat{r}+\hat{q}+\beta\lambda}}\exp\left(\frac{\hat{q}\epsilon^2}{2(\hat{r}+\hat{q}+\beta\lambda)}\right) \tag{30}$$

and

$$I_\theta^{(0)} = \frac{1}{\sqrt{2\pi}}\int \mathrm{d}\theta_* \exp\left(-\frac{1}{2}\theta_*^2 - \frac{\hat{m}^2}{2\hat{q}}\theta_*^2 + \frac{\hat{m}}{\sqrt{\hat{q}}}\epsilon\theta_*\right) = \sqrt{\frac{\hat{q}}{\hat{m}^2+\hat{q}}}\exp\left(\frac{\hat{m}^2\epsilon^2}{2(\hat{m}^2+\hat{q})}\right). \tag{31}$$

As a consequence, using the ansatz defined in (26)

$$\mathbb{E}_\epsilon I_\theta^{(0)}\log I_\theta^{(1)} = \sqrt{\frac{\hat{q}}{2\pi(\hat{m}^2+\hat{q})}}\int \mathrm{d}\epsilon \exp\left(-\frac{1}{2}\left(1 - \frac{\hat{m}^2}{\hat{m}^2+\hat{q}}\right)\epsilon^2\right)\frac{\hat{q}\epsilon^2}{2(\hat{r}+\hat{q}+\beta\lambda)} + o(\beta)$$

$$= \int \mathrm{d}\epsilon \mathcal{N}\left(\epsilon \mid 0, \frac{\hat{m}^2+\hat{q}}{\hat{q}}\right)\frac{\hat{q}\epsilon^2}{2(\hat{r}+\hat{q}+\beta\lambda} + o(\beta) = \frac{\beta^2(\hat{M}^2+\hat{\chi})}{2\beta(\hat{R}+\lambda)} + o(\beta)$$

Finally

$$\lim_{\beta\to\infty}\frac{1}{\beta}\mathbb{E}_\epsilon I_\theta^{(0)}\log I_\theta^{(1)} = \frac{\hat{M}^2+\hat{\chi}}{2(\hat{R}+\lambda)}. \tag{32}$$

[21] One should note that the extremization in (17) should be now performed with respect to $(m, r, \chi, \hat{M}, \hat{R}, \hat{\chi})$
[22] The factor 1/2 is conventional and does not affect the result of the estimation through ERM.

The extremization of ϕ in (17) with respect to $(\hat{M}, \hat{R}, \hat{\chi})$, considering also (27), results in the following self consistent equations:

$$\begin{aligned}
\lim_{\beta\to\infty} \frac{1}{\beta}\frac{\partial}{\partial \hat{M}}\phi &= -m + \frac{\hat{M}}{\hat{R}+\lambda} \overset{!}{=} 0 \quad &\implies \quad m &= \frac{\hat{M}}{\hat{R}+\lambda} \\
\lim_{\beta\to\infty} \frac{1}{\beta}\frac{\partial}{\partial \hat{R}}\phi &= \frac{r}{2} - \frac{\hat{M}^2+\hat{\chi}}{2(\hat{R}+\lambda)^2} \overset{!}{=} 0 \quad &\implies \quad r &= \frac{\hat{M}^2+\hat{\chi}}{(\hat{R}+\lambda)^2} \\
\lim_{\beta\to\infty} \frac{1}{\beta}\frac{\partial}{\partial \hat{\chi}}\phi &= -\frac{\chi}{2} + \frac{1}{2(\hat{R}+\lambda)} \overset{!}{=} 0 \quad &\implies \quad \chi &= \frac{1}{\hat{R}+\lambda}
\end{aligned} \tag{33}$$

L_1 regularization

Another frequent choice for the regularization is $r(\theta) = |\theta|$. This choice corresponds to

$$p_\theta(\theta) = \exp\left(-\beta\lambda|\theta|\right),$$

as defined in (5). Plugging it into the expression for $I_\theta^{(1)}$ in (18) and using the ansatz defined in (26) we obtain:

$$I_\theta^{(1)} = \int \mathrm{d}\theta \exp\left(-\beta\lambda|\theta| - \frac{\hat{r}+\hat{q}}{2}\theta^2 + \sqrt{\hat{q}}\epsilon\theta\right) = \int \mathrm{d}\theta \exp\left(-\beta\lambda|\theta| - \frac{\beta\hat{R}}{2}\theta^2 + \beta\sqrt{\hat{\chi}}\epsilon\theta\right)$$

As $\beta \to \infty$, we can perform a saddle-point (S.P.) evaluation of the last integral:

$$I_\theta^{(1)} \overset{\beta\gg 1}{\approx} \exp\left(-\beta \min_\theta \left(\lambda|\theta| + \frac{\hat{R}}{2}\theta^2 - \sqrt{\hat{\chi}}\epsilon\theta\right)\right) = \begin{cases} \exp\left(\frac{\beta}{2\hat{R}}(\sqrt{\hat{\chi}}\epsilon - \lambda)^2\right), & \epsilon > \frac{\lambda}{\sqrt{\hat{\chi}}} \\ \exp\left(\frac{\beta}{2\hat{R}}(\sqrt{\hat{\chi}}\epsilon + \lambda)^2\right), & \epsilon < -\frac{\lambda}{\sqrt{\hat{\chi}}} \\ 0, \quad \text{otherwise} \end{cases}$$

The expression for $I_\theta^{(0)}$ is again (31). Therefore, the term we shall compute is [23]

$$\begin{aligned}
\lim_{\beta\to\infty} \frac{1}{\beta}\mathbb{E}_\epsilon I_\theta^{(0)} \log I_\theta^{(1)} &= \frac{\sqrt{\hat{\chi}}}{\hat{R}\sqrt{2\pi(\hat{M}^2+\hat{\chi})}} \int_{\lambda/\sqrt{\hat{\chi}}}^{\infty} \mathrm{d}\epsilon \exp\left(-\frac{\hat{\chi}}{2(\hat{M}^2+\hat{\chi})}\epsilon^2\right)(\sqrt{\hat{\chi}}\epsilon - \lambda)^2 \\
&= \frac{\hat{M}^2+\hat{\chi}+\lambda^2}{2\hat{R}} \operatorname{erfc}\left(\frac{\lambda}{\sqrt{2(\hat{M}^2+\hat{\chi})}}\right) - \frac{\lambda\sqrt{\hat{M}^2+\hat{\chi}}}{\hat{R}\sqrt{2\pi}} \exp\left(-\frac{\lambda^2}{2(\hat{M}^2+\hat{\chi})}\right)
\end{aligned}$$

In order to simplify the notation in the following expressions, we define:

$$\phi_{\mathrm{erfc}} := \operatorname{erfc}\left(\frac{\lambda}{\sqrt{2(\hat{M}^2+\hat{\chi})}}\right), \quad \phi_\theta := \left(\hat{M}^2+\hat{\chi}+\lambda^2\right)\phi_{\mathrm{erfc}} - \lambda\sqrt{\frac{2\left(\hat{M}^2+\hat{\chi}\right)}{\pi}} e^{-\frac{\lambda^2}{2(\hat{M}^2+\hat{\chi})}},$$

where we kept implicit the dependance of the two functions on the overlap parameters and λ. The extremization of ϕ in (17) with respect to $(\hat{M}, \hat{R}, \hat{\chi})$, considering also (27), results in the

[23] The following computations can be easiy done using trivial changes of variable and using integrations by parts.

following self consistent equations:

$$\begin{aligned}
\lim_{\beta\to\infty}\frac{1}{\beta}\frac{\partial}{\partial\hat{M}}\phi &= -m+\frac{\hat{M}}{\hat{R}}\phi_{\text{erfc}}\overset{!}{=}0 \quad\implies\quad m=\frac{\hat{M}}{\hat{R}}\phi_{\text{erfc}}\\
\lim_{\beta\to\infty}\frac{1}{\beta}\frac{\partial}{\partial\hat{R}}\phi &= \frac{r}{2}-\frac{1}{2\hat{R}^2}\phi_{\theta}\overset{!}{=}0 \quad\implies\quad r=\frac{1}{\hat{R}^2}\phi_{\theta}\\
\lim_{\beta\to\infty}\frac{1}{\beta}\frac{\partial}{\partial\hat{\chi}}\phi &= -\frac{\chi}{2}+\frac{1}{2\hat{R}}\phi_{\text{erfc}}\overset{!}{=}0 \quad\implies\quad \chi=\frac{1}{\hat{R}}\phi_{\text{erfc}}
\end{aligned} \tag{34}$$

Quadratic loss

One of the most common choices for the loss function, when the labels are $y_i\in\mathbb{R}$, $i\in[n]$ and not dicotomized, is the quadratic loss $\ell(y,z)=(y-z)^2/2$ [24], that corresponds to

$$p_{out}(y\mid z)=\exp\left(-\frac{\beta}{2}(y-z)^2\right),$$

as defined in (5). Plugging it into the expression for $I_y^{(1)}$ in (18) we obtain

$$\begin{aligned}
I_y^{(1)} &= \frac{1}{\sqrt{2\pi}}\int \mathrm{d}h^0\exp\left(-\frac{(h^0)^2}{2}-\frac{\beta}{2}(y-\sqrt{\Delta(r-q)}h^0-\sqrt{\Delta q})^2\right)\\
&= \frac{1}{\sqrt{1+\beta\Delta(r-q)}}\exp\left(-\frac{\beta}{2}\frac{(y-\sqrt{\Delta q}\eta)^2}{1+\beta\Delta(r-q)}\right).
\end{aligned}$$

In order to compare the results of ERM with the Bayes optimal setting, we consider the linear case with noise variance σ^2 for the teacher likelihood (the same used in A.2), as it is the only likelihood for real non-dicotomized labels that we have presented in this work.
Hence, recalling that $r^0=1$,

$$\begin{aligned}
I_y^{(0)} &= \frac{1}{2\pi\sigma}\int \mathrm{d}h\exp\left(-\frac{h^2}{2}-\frac{1}{2\sigma^2}\left(y-\sqrt{\Delta\frac{q-m^2}{q}}h-\frac{\sqrt{\Delta}m}{\sqrt{q}}\eta\right)^2\right)\\
&= \frac{1}{\sqrt{2\pi\left(\sigma^2+\Delta\frac{q-m^2}{q}\right)}}\exp\left(-\frac{1}{2\left(\sigma^2+\Delta\frac{q-m^2}{q}\right)}\left(y-\frac{\sqrt{\Delta}m}{\sqrt{q}}\eta\right)^2\right)\\
&= \mathcal{N}\left(y\mid\frac{\sqrt{\Delta}m}{\sqrt{q}}\eta,\sigma^2+\Delta\frac{q-m^2}{q}\right).
\end{aligned}$$

Introducing the notation $\mathrm{E}_y[(\cdot)]=\int \mathrm{d}y\mathcal{N}\left(y\mid\frac{\sqrt{\Delta}m}{\sqrt{q}}\eta,\sigma^2+\Delta\frac{q-m^2}{q}\right)(\cdot)$ and the ansatz defined in (26) [25],this leads to [26]

$$\begin{aligned}
\lim_{\beta\to\infty}\frac{1}{\beta}\mathbb{E}_{\Delta,\eta}\int \mathrm{d}y I_y^{(0)}\log I_y^{(1)} &= -\mathbb{E}_{\Delta}\frac{1}{2(1+\Delta\chi)}\mathbb{E}_{\eta,y}\left[y^2+\Delta r\eta^2-2\sqrt{\Delta r}y\eta\right]\\
&= -\mathbb{E}_{\Delta}\frac{1}{2(1+\Delta\chi)}\left(\sigma^2+\Delta\frac{r-m^2}{r}+\frac{\Delta m^2}{r}+\Delta r-2\Delta m\right)
\end{aligned}$$

[24] The factor 1/2 is conventional and does not affect the result of the estimation through ERM.
[25] We recall that one of the implications of this ansatz is $\lim_{\beta\to\infty}q=r$.
[26] We are using $\mathbb{E}_\eta[\eta^2]=1$, $\mathbb{E}_y[y]=\frac{\sqrt{\Delta}m}{\sqrt{q}}\eta$ and $\mathbb{E}_y[y^2]=\sigma^2+\Delta\frac{q-m^2}{q}+(\mathbb{E}_y[y])^2$.

$$= -\mathbb{E}_\Delta \frac{\sigma^2 + \Delta(1-2m+r)}{2(1+\Delta\chi)}$$

In this ansatz, the extremization of ϕ in (17) with respect to (m, r, χ) results in the following self consistent equations:

$$\begin{aligned}
&\lim_{\beta\to\infty} \frac{1}{\beta}\frac{\partial}{\partial m}\phi = -\hat{M} + \alpha\mathbb{E}_\Delta \frac{\Delta}{1+\Delta\chi} \overset{!}{=} 0 && \implies \quad \hat{M} = \alpha\mathbb{E}_\Delta \frac{\Delta}{1+\Delta\chi} \\
&\lim_{\beta\to\infty} \frac{1}{\beta}\frac{\partial}{\partial r}\phi = \frac{\hat{R}}{2} - \frac{\alpha}{2}\mathbb{E}_\Delta \frac{\Delta}{1+\Delta\chi} \overset{!}{=} 0 && \implies \quad \hat{R} = \alpha\mathbb{E}_\Delta \frac{\Delta}{1+\Delta\chi} \\
&\lim_{\beta\to\infty} \frac{1}{\beta}\frac{\partial}{\partial \hat{\chi}}\phi = -\frac{\hat{\chi}}{2} + \frac{\alpha}{2}\mathbb{E}_\Delta \frac{\Delta(\sigma^2+\Delta(1-2m+r))}{(1+\Delta\chi)^2} \overset{!}{=} 0 && \implies \quad \hat{\chi} = \alpha\mathbb{E}_\Delta \frac{\Delta(\sigma^2+\Delta(1-2m+r))}{(1+\Delta\chi)^2}
\end{aligned} \tag{35}$$

B Inverse and determinant of a replica symmetrix matrix

Inverse of replica symmetric matrix

Here we aim to compute the elements of $\mathbf{q}^{-1}$. It is easy to see that they can be parametrized exactly in the same way as $\mathbf{q}$,

$$\begin{aligned}
&\left(q^{-1}\right)^{00} = \tilde{r}^0 && \\
&\left(q^{-1}\right)^{a0} = \tilde{m} && \text{for } 1 \le a \le s \\
&\left(q^{-1}\right)^{aa} = \tilde{r} && \text{for } 1 \le a \le s \\
&\left(q^{-1}\right)^{ab} = \tilde{q} && \text{for } 1 \le a < b \le s
\end{aligned}$$

We know that this satisfy $\mathbf{q}^{-1}\mathbf{q} = \mathbb{I}_s$, which in components read

$$\sum_{c=0}^{s} \left(q^{-1}\right)^{ac} q^{cb} = \left(q^{-1}\right)^{a0} q^{0b} + \sum_{c=1}^{s} \left(q^{-1}\right)^{ac} q^{cb} = \delta^{ab},$$

where δ^{ab} is the Kronecker delta.[27]Separating in components,

$$\begin{cases}
\tilde{r}^0 r^0 + s\tilde{m}m = 1 & (a = b = 0) \\
\tilde{r}^0 m + \tilde{m}r + (s-1)\tilde{m}q = 0 & (a = 0, b > 0) \\
\tilde{m}r^0 + \tilde{r}m + (s-1)\tilde{q}m = 0 & (a > 1, b = 0) \\
m\tilde{m} + r\tilde{r} + (s-1)q\tilde{q} = 1 & (a = b > 0) \\
m\tilde{m} + \tilde{r}q + \tilde{q}r + (s-2)q\tilde{q} = 0 & (1 \le a < b \le s)
\end{cases}$$

The solution for this system is given by

[27] $\delta^{ab} = 1$ iif $a = b$, otherwise $\delta^{ab} = 0$.

$$\left(q^{-1}\right)^{00} = \tilde{r}^0 = \frac{r+(s-1)q}{r^0(r+(s-1)q)-sm^2}, \quad \left(q^{-1}\right)^{aa} = \tilde{r} = \frac{r^0 r+(s-2)r^0 q-(s-1)m^2}{(r-q)(r^0 r+(s-1)r^0 q-sm^2)},$$
$$\left(q^{-1}\right)^{a0} = \tilde{m} = \frac{m}{sm^2-r^0 r-(s-1)r^0 q}, \quad \left(q^{-1}\right)^{ab} = \tilde{q} = \frac{m^2-r^0 q}{(r-q)(r^0 r+(s-1)r^0 q-sm^2)}.$$

In the $s \to 0^+$ limit of the above,

$$\lim_{s\to 0^+} \tilde{r}^0 = \frac{1}{r^0} \qquad \lim_{s\to 0^+} \tilde{r} = \frac{m^2+(r-2q)r^0}{r^0(r-q)^2} \tag{36}$$

$$\lim_{s\to 0^+} \tilde{m} = -\frac{m}{r^0(r-q)} \qquad \lim_{s\to 0^+} \tilde{q} = \frac{m^2-r^0 q}{r^0(r-q)^2} \tag{37}$$

In particular, note that this satisfy

$$\lim_{s\to 0^+} (\tilde{r}-\tilde{q}) = \frac{1}{r-q}.$$

Determinant of a replica symmetric matrix

The replica symmetric overlap matrix is given by

$$\mathbf{q} = \begin{pmatrix} r^0 & m & m & \dots & m \\ m & r & q & \dots & q \\ m & q & r & \dots & q \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ m & q & q & \dots & r \end{pmatrix} \in \mathbb{R}^{(s+1)\times(s+1)}$$

In order to compute its determinant, we will attempt to guess its eigenvector. We first try with:

$$\begin{pmatrix} r^0 & m & m & \dots & m \\ m & r & q & \dots & q \\ m & q & r & \dots & q \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ m & q & q & \dots & r \end{pmatrix} \begin{pmatrix} x \\ 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} = \begin{pmatrix} r^0 xsm \\ mx+r+(s-1)q \\ mx+r+(s-1)q \\ \vdots \\ mx+r+(s-1)q \end{pmatrix} \stackrel{!}{=} \lambda \begin{pmatrix} x \\ 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix},$$

giving two linear equations for the two unknowns (x, λ),

$$\begin{cases} \lambda x = r^0 x + sm \\ \lambda = mx + r + (s-1)q \end{cases}$$

Inserting the second equation into the first give a quadratic equation for x,

$$mx^2 + \left[\left(r-r^0\right) + (s-1)q\right] x - sm = 0.$$

The solutions of the system are therefore,

$$\begin{cases} x_\pm = -\frac{1}{2m}\left[(r-r^0)+(s-1)q\right] \pm \frac{1}{2m}\sqrt{4sm^2+\left[(r-r^0)+(s-1)q\right]^2} \\ \lambda_\pm = \frac{1}{2}\left[(r+r^0+(s-1)q)\right] \pm \frac{1}{2}\sqrt{4sm^2+\left[(r-r^0)+(s-1)q\right]^2} \end{cases}$$

Note that the product of eigenvalues simplifies to

$$\lambda_+\lambda_- = r^0 r + (s-1)r^0 q - sm^2$$

The other $s-1$ eigenvalues can be easily found by checking that the vector $\mathbf{v}^{(i)} \in \mathbb{R}^{s+1}$ whose elements are defined as $v_j^{(i)} = \delta^{i,j} - \delta^{i+1,j}$ is an eigenvector. In fact:

$$\begin{pmatrix} r^0 & m & m & \dots & m \\ m & r & q & \cdots & q \\ m & q & r & \cdots & q \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ m & q & q & \cdots & r \end{pmatrix} \begin{pmatrix} 0 \\ 1 \\ -1 \\ \vdots \\ 0 \end{pmatrix} = (r-q) \begin{pmatrix} 0 \\ 1 \\ -1 \\ \vdots \\ 0 \end{pmatrix}$$

There are $s-1$ such independent eigenvectors. Therefore the determinant, which is the products of eigenvalues, is given by

$$\det \mathbf{q} = (r-q)^{s-1}\left(rr^0 + (s-1)r^0 q - sm^2\right)$$

In particular, we have the following useful asymptotic:

$$\begin{aligned} \log\det \mathbf{q} &= (s-1)\log(r-q) + \log\left(rr^0 + (s-1)r^0 q - sm^2\right) \\ &= \log r^0 + s\left[\log(r-q) + \frac{r^0 q - m^2}{r^0(r-q)}\right] + O\left(s^2\right). \end{aligned}$$

C Bayes optimal setting: some technicalities

The Bayes optimal setting, as defined in 2, is the supervised learning setting in which the teacher and the student distributions coincide. Given our choice of using the mean square error (MSE) as the estimation error, this setting is *optimal* in the sense that it allows to achieve the minimal error, meaning that the (minimal mean square error) estimator is

$$\hat{\boldsymbol{\theta}}(\mathcal{D}) := \arg\min_{\boldsymbol{\theta}} \mathbb{E}_{\boldsymbol{\theta}_*|\mathcal{D}} ||\boldsymbol{\theta} - \boldsymbol{\theta}_*||^2, \tag{38}$$

where in general the expectation $\mathbb{E}_{\boldsymbol{\theta}_*|\mathcal{D}}$ refers to the teacher posterior distribution

$$p^*(\boldsymbol{\theta}_* \mid \mathcal{D}) = \frac{1}{Z(\mathcal{D})} \prod_{l\in[d]} p_\theta^*(\theta_{l,*}) \prod_{i\in[n]} p_{out}^*\left(y_i \mid \boldsymbol{\theta}_*^\top \mathbf{x}_i\right).$$

The latter coincides with (3) only in Bayes optimal setting.

C.1 Optimality of the mean posterior estimator

The mean posterior (MP) estimator is defined as $\mathbb{E}_{\boldsymbol{\theta}|\mathcal{D}}[\boldsymbol{\theta}]$, where the expectation is done with respect to the (student) posterior (3).
In this section we show that the MP and the MMSE estimators coincides in Bayes optimal setting. The minimization of the (averaged) MSE in (38) implies

$$\begin{aligned} 0 \overset{!}{=} \boldsymbol{\nabla}_{\boldsymbol{\theta}} \mathbb{E}_{\boldsymbol{\theta}_*|\mathcal{D}} ||\boldsymbol{\theta} - \boldsymbol{\theta}_*||^2 &= \boldsymbol{\nabla}_{\boldsymbol{\theta}} \int \mathrm{d}\boldsymbol{\theta}_* p^*(\boldsymbol{\theta}_* \mid \mathcal{D}) ||\boldsymbol{\theta} - \boldsymbol{\theta}_*||^2 \\ &= 2\left(\boldsymbol{\theta} - \int \mathrm{d}\boldsymbol{\theta}_* p^*(\boldsymbol{\theta}_* \mid \mathcal{D}) \boldsymbol{\theta}_*\right) \end{aligned}$$

Therefore, $\hat{\boldsymbol{\theta}}(\mathcal{D}) = \mathbb{E}_{\boldsymbol{\theta}_*|\mathcal{D}}[\boldsymbol{\theta}_*]$, which is the MP estimator only considering the Bayes optimal setting.

C.2 Nishimori identity

Proposition (Nishimori identity): *given a couple of random variable (X, Y) drawn from the joint distribution $P(X, Y)$ and conditional distribution $P(X \mid Y)$. Let $k \geq 1$ and $x^{(1)}, \dots, x^{(k)}$ (the replicas) be i.i.d. samples from the conditional $P(X = \cdot \mid Y)$. Let us denote $\langle \cdot \rangle_k$ the expectation with respect to the conditional distribution $P(x^{(1)} \mid Y) \dots P(x^{(k)} \mid Y)$ and $\mathbb{E}$ the expectation with respect to the joint distribution $P(X \mid Y)$. Then, for any continuous bounded function g,*

$$\mathbb{E}\left\langle g\left(Y, x^{(1)}, \dots, x^{(k)}\right)\right\rangle_k = \mathbb{E}\left\langle g\left(Y, X, x^{(1)}, \dots, x^{(k-1)}\right)\right\rangle_{k-1}$$

Nishimori identity is a direct consequence of Bayes' formula. In fact, sampling (X, Y) from $P(X, Y)$ is equivalent to sampling Y from its marginal distribution $P(Y) = \int \mathrm{d}x P(x, Y)$ and then sampling X from $P(X \mid Y)$. Therefore

$$\begin{aligned} \mathbb{E}\left\langle g\left(Y, X, x^{(1)}, \dots, x^{(k-1)}\right)\right\rangle_{k-1} &= \mathbb{E}_Y \int \mathrm{d}X P(X \mid Y) \int \prod_{i=1}^{k-1} \mathrm{d}x_i P(x_i \mid Y) g\left(Y, X, x^{(1)}, \dots, x^{(k-1)}\right) \\ &= \mathbb{E}_Y \int \prod_{i=1}^{k} \mathrm{d}x_i P(x_i \mid Y) g\left(Y, x^{(1)}, \dots, x^{(k)}\right) \\ &= \mathbb{E}\left\langle g\left(Y, x^{(1)}, \dots, x^{(k)}\right)\right\rangle_k, \end{aligned}$$

where the second equality is just a change of the name of the integration variable X to $x^{(k)}$.
An useful consequence of this identity concerns the overlaps in Bayes optimal setting, more accurately their typical value.
In fact [28]

$$q = d^{-1}\mathbb{E}_{\mathcal{D}}\left\langle \boldsymbol{\theta}^{(1)\top}\boldsymbol{\theta}^{(2)}\right\rangle_2 = d^{-1}\mathbb{E}_{\mathcal{D}}\left\langle \boldsymbol{\theta}_*^{\top}\boldsymbol{\theta}^{(1)}\right\rangle_1 = m$$

and similarly $r^0 = r$, $\hat{q} = \hat{m}$ and $\hat{r}^0 = \hat{r}$. Exploiting the Nishimori identity it is also possible to prove that the overlap matrix is replica symmetric in Bayes optimal setting [7, 5].

[28] The notation $\langle \cdot \rangle_k$, consistently to the one used in the **Proposition**, refers to the expectation with respect to $p\left(\boldsymbol{\theta}^{(1)} \mid \mathcal{D}\right) \dots p\left(\boldsymbol{\theta}^{(k)} \mid \mathcal{D}\right)$. Since $\mathbb{E}_{\mathcal{D}}$ is an expectation with respect to the teacher distributions, it is easy to see that Nishimori identity can be applied only in Bayes optimal setting.

C.3 Generalized Aprroximate Message Passing algorithm

Bayes optimal estimation could be performed by sampling from the posterior (3), which is costly in high-dimensions in general and finding an algorithm that achieves the MMSE in polynomial time is a hard problem. For this specific setting, an algorithm inspired by statistical physics called (generalized) Approximate Message Passing can be used to compute marginals of the posterior and achieve the optimal estimation in polynomial time (with respect to d) [6]. Therefore, we use it to perform the numerical experiments of section 4. This formulation of the algorithm is the one presented in [12]. Note that this algorithm assume Bayes optimal setting, hence the output channel and the prior distributions are the same for the teacher and the student.

Algorithm 1 GAMP

Input: Data $\mathbf{X} \in \mathbb{R}^{n\times d}$, $\mathbf{y} \in \mathcal{Y}^n$
  Define $X^2 \leftarrow \mathbf{X} \odot \mathbf{X}$
  Initialize $\hat{\boldsymbol{w}}^{t=0} = \mathbf{0} \in \mathbb{R}^d$, $\hat{\boldsymbol{c}}^{t=0} = \mathbf{1} \in \mathbb{R}^d$, $\boldsymbol{g}^{t=0} = \mathbf{0} \in \mathbb{R}^n$
  while $t < t_{\max}$ **do**:
    $\boldsymbol{V}^t = X^2\hat{\boldsymbol{c}}^t$; $\boldsymbol{\omega}^t = \boldsymbol{X}\hat{\boldsymbol{w}}^{t-1} - \boldsymbol{V}^t \odot \boldsymbol{g}^{t-1}$;
    $\boldsymbol{g}^t = f_{out}(\mathbf{y}, \boldsymbol{\omega}^t, \boldsymbol{V}^t)$; $\partial\boldsymbol{g}^t = \partial_\omega f_{out}(\mathbf{y}, \boldsymbol{\omega}^t, \boldsymbol{V}^t)$;
    $\boldsymbol{A}^t = -{X^2}^\top \partial\boldsymbol{g}$; $\boldsymbol{b}^t = \boldsymbol{A}^t \odot \hat{\boldsymbol{w}}^t + X^\top \boldsymbol{g}^t$;
    $\hat{\boldsymbol{w}}^{t+1} = f_\theta(\boldsymbol{b}^t, \boldsymbol{A}^t)$, $\hat{\boldsymbol{c}}^t = \partial_{\boldsymbol{b}} f_\theta(\boldsymbol{b}^t, \boldsymbol{A}^t)$
  end while
Return: $\hat{\boldsymbol{w}}^{t_{\max}}$, $\hat{\boldsymbol{c}}^{t_{\max}}$

where the auxiliary functions $f_{out}(y, \omega, V) = \partial_\omega \log \mathcal{Z}_{out}(y, \omega, V)$ and $f_\theta(b, A) = \partial_b \log \mathcal{Z}_\theta(b, A)$ are scalar functions acting component-wise that depend on the output channel and priors:

$$\mathcal{Z}_{out}(y, \omega, V) = \int \frac{\mathrm{d}z}{\sqrt{2\pi V}} e^{-\frac{(z-\omega)^2}{2V}} p_{out}(y|z), \qquad \mathcal{Z}_\theta(b, A) = \int \mathrm{d}\theta \; e^{-\frac{A}{2}\theta^2 + b\theta} p_\theta(\theta) \tag{39}$$

Note that $\hat{\boldsymbol{w}}^{t_{\max}}$ is an estimator of $\boldsymbol{\theta}_*$.
Our choice of gaussian prior implies

$$f_\theta(b, A) = \frac{b}{1+A}.$$

For the linear channel defined in 3.1.1, with noise variance σ^2, we have that

$$f_{out}(y, \omega, V) = \frac{y - \omega}{V + \sigma^2}.$$

For the probit channel defined in 3.1.2, with noise variance σ^2, we have that

$$f_{out}(y, \omega, V) = \sqrt{\frac{2}{\pi(V+\sigma^2)}} y e^{-\frac{\omega^2}{2(V+\sigma^2)}} \left[\operatorname{erfc}\left(-\frac{y\omega}{\sqrt{2(V+\sigma^2)}}\right)\right]^{-1}.$$

D Superstatistical model with inverse Gamma distribution

Consider the inverse Gamma distribution with shape parameter $a > 0$ and scale parameter $b > 0$, namely $\rho(\Delta \mid a, b) = b^a (1/\Delta)^{a+1} \exp(-b/\Delta)/\Gamma(a)$, then (1) becomes

$$\begin{aligned}
\mathbb{P}(\mathbf{x}) &:= \int_0^\infty \mathcal{N}\left(\boldsymbol{x} \mid \boldsymbol{\mu}, \frac{\Delta}{d}\mathbb{I}_d\right) \rho(\Delta \mid a, b)\, \mathrm{d}\Delta \\
&= \frac{b^a}{\Gamma(a)} \int_0^\infty \frac{1}{\Delta^{a+1}\sqrt{\det(2\pi d^{-1}\Delta \mathbb{I}_d)}} \exp\left(-\frac{\frac{1}{2}d(\mathbf{x}-\boldsymbol{\mu})^\top(\mathbf{x}-\boldsymbol{\mu})+b}{\Delta}\right) \mathrm{d}\Delta \\
&= \frac{b^a d^{d/2}}{(2\pi)^{d/2}\Gamma(a)} \int_0^\infty \frac{1}{\Delta^{a+1+\frac{d}{2}}} \exp\left(-\frac{\frac{1}{2}d||\mathbf{x}-\boldsymbol{\mu}||^2)+b}{\Delta}\right) \mathrm{d}\Delta \\
&= \frac{b^a \Gamma(A) d^{d/2}}{(2\pi)^{d/2} B^A \Gamma(a)} \underbrace{\int_0^\infty \rho\left(\Delta \mid A = a + \frac{d}{2}, B = \frac{1}{2}d||\mathbf{x}-\boldsymbol{\mu}||^2 + b\right) \mathrm{d}\Delta}_{=1} \\
&= \frac{(2b)^a \Gamma(a+\frac{d}{2}) d^{d/2}}{\pi^{d/2}\Gamma(a)} \left(2b + ||\sqrt{d}(\mathbf{x}-\boldsymbol{\mu})||^2\right)^{-a-\frac{d}{2}}.
\end{aligned} \tag{40}$$

Different choices of the parameters a and b allows to explore various regimes for the covariates distribution. In fact, for $a > 1$ and considering that the average (rescaled) covariance matrix $d\mathbb{E}_\Delta[(\mathbf{x}-\boldsymbol{\mu})^\top(\mathbf{x}-\boldsymbol{\mu})] = \mathbb{E}_\Delta[\Delta] = b/(a-1) =: \overline{\Delta}$, by keeping $\overline{\Delta}$ fixed and sending $a \to \infty$ we have that $\mathbb{P}(\mathbf{x}) \to \mathcal{N}(\mathbf{x} \mid \boldsymbol{\mu}, \frac{\overline{\Delta}}{d}\mathbb{I}_d$, the known Gaussian case, while by approaching the limit $a \to 1^+$, the distribution $\mathbb{P}(\mathbf{x})$ get heavier tails. In our simulations we consider the scaling $b = a - 1$ so that $\overline{\Delta} = 1$ indipendently of a. Instead, for $a \in (0, 1]$ the quantity $\mathbb{E}_\Delta[(\mathbf{x}-\boldsymbol{\mu})^\top(\mathbf{x}-\boldsymbol{\mu})]$ is not finite, hence the covariates' distribution has infinite covariance.
Moreover, the choice of this distribution finds another motivation in previous works that adopted it to describe non-Gaussian data in quantitative finance [15], [26] and econometrics [37].