Machine Learning: Science and Technology

Theoretical characterization of uncertainty in high-dimensional linear classification

IOP Publishing Mach. Learn.: Sci. Technol. 4 (2023) 025029 https://doi.org/10.1088/2632-2153/acd749

MACHINE LEARNING
Science and Technology

OPEN ACCESS

RECEIVED
27 February 2023

REVISED
25 April 2023

ACCEPTED FOR PUBLICATION
4 May 2023

PUBLISHED
8 June 2023

PAPER

Theoretical characterization of uncertainty in high-dimensional linear classification

Lucas Clarté[1,*], **Bruno Loureiro**[2], **Florent Krzakala**[3] **and Lenka Zdeborová**[1]

[1] École Polytechnique Fédérale de Lausanne (EPFL), Statistical Physics of Computation lab. CH-1015 Lausanne, Switzerland
[2] Département d'Informatique, École Normale Supérieure—PSL & CNRS, 45 rue d'Ulm, F-75230 Paris cedex 05, France
[3] École Polytechnique Fédérale de Lausanne (EPFL) Information, Learning and Physics lab. CH-1015 Lausanne, Switzerland
[*] Author to whom any correspondence should be addressed.

E-mail: lucas.clarte@epfl.ch

Keywords: high-dimensional statistics, uncertainty quantification, Bayesian inference, approximate message passing, state evolution

Abstract

Being able to reliably assess not only the *accuracy* but also the *uncertainty* of models' predictions is an important endeavor in modern machine learning. Even if the model generating the data and labels is known, computing the intrinsic uncertainty after learning the model from a limited number of samples amounts to sampling the corresponding posterior probability measure. Such sampling is computationally challenging in high-dimensional problems and theoretical results on heuristic uncertainty estimators in high-dimensions are thus scarce. In this manuscript, we characterize uncertainty for learning from a limited number of samples of high-dimensional Gaussian input data and labels generated by the probit model. In this setting, the Bayesian uncertainty (i.e. the posterior marginals) can be asymptotically obtained by the approximate message passing algorithm, bypassing the canonical but costly Monte Carlo sampling of the posterior. We then provide a closed-form formula for the joint statistics between the logistic classifier, the uncertainty of the statistically optimal Bayesian classifier and the ground-truth probit uncertainty. The formula allows us to investigate the calibration of the logistic classifier learning from a limited amount of samples. We discuss how over-confidence can be mitigated by appropriately regularizing.

1. Introduction

An important part of statistics is concerned with assessing the *uncertainty* associated with a prediction based on data. Indeed, in many sensitive fields where statistical methods are widely used, trustworthiness can be as important as accuracy. The same holds true for modern applications of machine learning where liability is important, e.g. self-driving cars and facial recognition. Yet, assessing the uncertainty of machine learning methods comes with many questions. Measuring uncertainty in complex architectures such as deep neural networks is a challenging problem, with a rich literature proposing different strategies, e.g. [1, 21, 27, 28, 34, 35, 57, 65].

On the side of theoretical control of the uncertainty estimators there is an extended work in the context of Gaussian processes [29, 40, 59] that offer Bayesian estimates of uncertainties based on a Gaussian approximation over the predictor class [18, 41, 57]. Essentially when the posterior measure is a high-dimensional Gaussian then computation of the marginals is possible and well controlled. Beyond the setting of Gaussian posterior measures, well-established mathematical guarantees fall short in the high-dimensional regime where the number of data samples is of the same order as the number of dimensions even for the simplest models [61]. Sharp theoretical results on uncertainty quantification in high-dimensional models where posterior distributions are not Gaussian are consequently scarce.

In this manuscript we provide an exact characterization of uncertainty for high-dimensional classification of data with Gaussian covariates and probit labels. There are two main sources of uncertainty in

this model—the more explicit is the noise level parameterizing the probit function, then there is the uncertainty coming from the fact that learning is done from a limited number of samples. Uncertainty estimation in classification problems aims to compute the probability that a given new sample has one of the labels. The most likely label is then typically chosen for the prediction of the new labels, but the probability itself is of our interest here. We stress that we are interested in the uncertainty sample-wise, i.e. for every given sample, not on average. We address questions such as: (a) How does the uncertainty of the logistic classifier compare with the actual Bayesian uncertainty when learning with a limited amount of data? (b) How do these two uncertainty measures compare with the intrinsic model uncertainty due to the noise in the data-generating process?

The key player in our analysis will be the Bayesian estimator of uncertainty corresponding to the probabilities of labels for new samples computed by averaging over the posterior distribution. Although in general computing the Bayesian estimator from posterior sampling can be prohibitively computationally costly in high dimensions, we show that in the present model, it can be efficiently done using a tailored generalized approximate message passing (GAMP) algorithm [13, 56]. Leveraging tools from the GAMP literature and its state evolution, we provide an asymptotic characterization of the joint statistics between the minimizer of the logistic loss, the optimal Bayesian estimator over the data and the oracle estimator. This allows us to provide quantitative answers to questions a) & b) above, and to study how uncertainty estimation depends on the parameters of the model, such as the regularization, size of the training set and noise.

1.1. Main contributions

The main contributions in this manuscript are:

- It is well known that the optimal Bayesian classifier for a data model with Gaussian i.i.d. covariates and probit labels is well approximated by the GAMP algorithm [11, 31]. We extend these results by showing that GAMP also provides an exact sample-wise estimation of the Bayesian uncertainty when $d \to \infty$.
- We provide an exact asymptotic description of the joint statistics between the uncertainty of the oracle, and the one estimated by the Bayes-optimal (BO) and logistic classifiers for the aforementioned data model. This allows us to compare these uncertainties to each other. Comparing the oracle and Bayes optimal we quantify the uncertainty coming from the limited size of the dataset. Comparing Bayesian and logistic classifiers allows us to quantify the under- or overconfidence of the latter.
- We derive an asymptotic expression of the calibration for the Bayesian and logistic classifiers. In particular, we show that the Bayesian estimator is calibrated. For the logistic classifier, our expression allows us to characterize the influence of various parameters on under- or overconfidence of the logistic classifier.
- We quantify the role played by the ℓ_2-regularization on uncertainty estimation. In particular, we compare cross-validation with respect to the optimization loss (logistic) with cross-validation with respect to the $0/1$ error.

1.2. Related work

1.2.1. Measures of uncertainty

Measuring uncertainty in neural networks is a challenging problem with a vast literature proposing both frequentist and Bayesian approaches [1]. On the frequentist side, various algorithms have been introduced to evaluate and improve the calibration of machine learning models. Some of them, such as isotonic regression [67], histogram binning [66], Platt scaling [54] or temperature scaling [27] are applied to previously trained models. Other approaches aim to calibrate models during training, using well-chosen metrics [37, 53], through data augmentation [64] or using the iterates of the optimizer [42]. Alternatively, different authors have proposed uncertainty measures based on Bayesian estimates [47, 65]. This includes popular methods such as Bayesian dropout [21, 33], deep ensembles [35, 37, 44] and variational inference [55], Laplace approximation [18, 34] and tempered posteriors [2, 3, 32] to cite a few. Finally, some works based on conformal inference [60] are concerned with providing non-asymptotic and distribution-free guarantees for the uncertainty [4, 28].

1.2.2. Exact asymptotics

Our theoretical analysis builds on series of developments on the study of exact asymptotics in high-dimensions. The GAMP algorithm and the corresponding state evolution equations appeared in [31, 56]. Exact asymptotics for Bayesian estimation in generalized linear models was rigorously established in [11]. On the empirical risk minimization side, exact asymptotics based on different techniques, such as Convex Gaussian min-max theorem [5, 16, 19, 36, 38, 51, 52, 62], Random Matrix Theory [43], GAMP [23, 39] and first order expansions [14] have been used to study high-dimensional logistic regression and max-margin estimation.

1.2.3. Uncertainty & exact asymptotics
An early discussion on the variance of high-dimensional Bayesian linear regression has been appeared in [15, 45, 46]. Calibration has been studied in the context of high-dimensional unregularized logistic regression in [9], where it was shown that the logistic classifier is systematically overconfident in the regime where number of samples is proportional to the dimension. An equivalent result for regression was discussed in [10], where it was shown that quantile regression suffers from an under-coverage bias in high-dimensions. While [9] is the closest to the present paper, we differ from their setting in three major ways. First, they consider the behavior of unpenalized logistic regression, while we study the effect of ℓ_2 regularization on uncertainty. Second, we compute the full joint distribution of the prediction for the oracle, the empirical risk minimizer and the Bayes optimal estimator, while [9] focus the discussion on the calibration of the empirical risk minimizer with respect to the oracle only. Lastly (and less importantly), [9] considers logit data, while we consider a probit data model.

1.3. Notation
Vectors are denoted in bold. $\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu},\Sigma)$ denotes the Gaussian density. $\odot$ denotes the (component-wise) Hadamard product. $1(A)$ denotes the indicator on the set A. For any differentiable function $f:\mathbb{R}\to\mathbb{R}$, f' will denote its first derivative.

2. Setting

2.1. The data model
Consider a binary classification problem where n samples $(\boldsymbol{x}^\mu, y^\mu)\in\mathbb{R}^d\times\{-1,1\}$, $\mu=1,\dots,n$ are independently drawn from the following probit model:

$$f_\star(\boldsymbol{x}) := \mathbb{P}(y^\mu=1|\boldsymbol{x}^\mu) = \sigma_\star\left(\frac{\boldsymbol{w}_\star^\top\boldsymbol{x}^\mu}{\tau}\right), \tag{1}$$

$$\boldsymbol{x}^\mu\sim\mathcal{N}(\boldsymbol{0},1/d\mathrm{I}_d), \quad \boldsymbol{w}_\star\sim\mathcal{N}(\boldsymbol{0},\mathrm{I}_d) \tag{2}$$

where $\sigma_\star(x)=1/2\,\mathrm{erfc}(-x/\sqrt{2})$ and $\tau\geqslant 0$ parameterizes the noise level. Note that the probit model is equivalent to generating the labels via $y^\mu=f_0(\boldsymbol{w}_\star^\top\boldsymbol{x}^\mu+\tau\xi^\mu)$ with $\xi^\mu\sim\mathcal{N}(0,1)$ and $f_0(x):=\mathrm{sign}(x)$. In the following we will be referring to the function $f_\star(\boldsymbol{x})$ or to its parameters $\boldsymbol{w}_\star$ as the *teacher*, having in mind the teacher-student setting from neural networks. We will refer to $f_\star(\boldsymbol{x})$ as the *oracle uncertainty* as it takes into account only the noise in the label-generating process, but it does not take into account uncertainty coming from the limited size of the training dataset.

Note that our discussion could be straightforwardly generalized to a generic prior distribution $\boldsymbol{w}_\star\sim P_{\boldsymbol{w}_\star}$. However, our goal in this work is to provide a fair comparison between Bayesian estimation and empirical risk minimization (ERM). Indeed, ERM does not assume any information on the components of $\boldsymbol{w}_\star$, and a fair comparison is to consider the maximum entropy Gaussian prior.

Given the training data $\mathcal{D}=\{(\boldsymbol{x}^\mu,y^\mu)\}_{\mu=1}^n$ and a test sample $\boldsymbol{x}\sim\mathcal{N}(\boldsymbol{0},1/d\mathrm{I}_d)$, the goal is to find a (probabilistic) classifier $\boldsymbol{x}\mapsto\hat{y}(\boldsymbol{x})$ minimizing the $0/1$ test error

$$\varepsilon_g = \mathbb{E}_{(\boldsymbol{x},y)}\mathbb{P}\left(\hat{y}(\boldsymbol{x})\neq y\right). \tag{3}$$

When different estimators are compared, we will note $\varepsilon_g^{\mathrm{t}}$ the error of the estimator t to remove ambiguity.

2.2. Considered classifiers
We will focus on comparing two probabilistic classifiers $\hat{f}(\boldsymbol{x})=\mathbb{P}(y=1|\boldsymbol{x})$. The first is the widely used logistic classifier: $\hat{f}_{\mathrm{erm}}(\boldsymbol{x})=\sigma(\hat{\boldsymbol{w}}_{\mathrm{erm}}^\top\boldsymbol{x})$ where $\sigma(x)=(1+e^{-x})^{-1}$ is the logistic function and the weights $\hat{\boldsymbol{w}}\in\mathbb{R}^d$ are obtained by minimizing the following (regularized) empirical risk:

$$\hat{\mathcal{R}}_n(\boldsymbol{w}) = \frac{1}{n}\sum_{\mu=1}^n\log\left(1+e^{-y^\mu\boldsymbol{w}^\top\boldsymbol{x}^\mu}\right)+\frac{\lambda}{2}||\boldsymbol{w}||_2^2. \tag{4}$$

Using $\hat{f}_{\mathrm{erm}}(\boldsymbol{x})$ as a measure of uncertainty is not considered very principled. Nevertheless, it is arguably the most commonly used measure to give a rough idea of how confident is the neural network prediction for a given sample.

The second estimator we investigate is the statistically optimal Bayesian estimator for the problem, which is given by:

$$\hat{f}_{\rm bo}(\boldsymbol{x}) = \mathbb{P}_{\rm BO}(y=1|\boldsymbol{x}) = \int_{\mathbb{R}^d} \mathrm{d}\boldsymbol{w}\, P(y=1|\boldsymbol{x}^\top \boldsymbol{w}) P(\boldsymbol{w}|\mathcal{D}) \\ = \int_{\mathbb{R}^d} \mathrm{d}\boldsymbol{w}\, \sigma_\star\left(\frac{\boldsymbol{w}^\top \boldsymbol{x}}{\tau}\right) P(\boldsymbol{w}|\mathcal{D})\,, \tag{5}$$

where the posterior distribution $P(\boldsymbol{w}|\mathcal{D})$ given the training data $\mathcal{D} = \{(\boldsymbol{x}^\mu, y^\mu)\}_{\mu=1}^n$ is explicitly given by:

$$P(\boldsymbol{w}|\mathcal{D}) = \frac{1}{\mathcal{Z}(\tau)} \prod_{\mu=1}^{n} \sigma_\star\left(y^\mu \frac{\boldsymbol{w}^\top \boldsymbol{x}^\mu}{\tau}\right) \mathcal{N}(\boldsymbol{w}|\boldsymbol{0}, \mathrm{I}_d), \tag{6}$$

for a normalization constant $\mathcal{Z}(\tau) \in \mathbb{R}$. The BO estimator $\hat{f}_{\rm bo}(\boldsymbol{x})$ provides the perfect measure of uncertainty that takes into account both the noise in the data generation and the finite number of samples in the training set. The traditional drawback of course is that it assumes the knowledge of the value τ and other details of the data-generating model.

2.3. Uncertainty and calibration

The main purpose of this manuscript is to characterize how the intrinsic uncertainty of the probit model compares to both the Bayesian and logistic confidences/uncertainties in the high-dimensional setting where the number of samples n is comparable to the dimension d. In this case, the limited number of samples is a source of uncertainty comparable in magnitude to the noise level τ. To define what is uncertainty in our context, note that the *confidence functions* $\hat{f}(\boldsymbol{x}) = \mathbb{P}(y=1|\boldsymbol{x})$ defined above give the probability that the label is $y=1$ (with the label prediction commonly given by thresholding this function). In mathematical terms, we aim at characterizing the correlation between the oracle, Bayesian and logistic confidences, as parameterized by the joint probability density:

$$\rho(a,b,c) = \mathbb{P}_{\mathcal{D},\boldsymbol{x}}\big(f_\star(\boldsymbol{x}) = a, \hat{f}_{\rm bo}(\boldsymbol{x}) = b, \hat{f}_{\rm erm}(\boldsymbol{x}) = c\big)\,. \tag{7}$$

Similarly, we will note $\rho_{\star,\rm erm}(a,c) = \mathbb{P}(f_\star = a, \hat{f}_{\rm erm} = c)$, $\rho_{\rm bo,erm}(b,c) = \mathbb{P}(\hat{f}_{\rm bo} = b, \hat{f}_{\rm erm} = c)$ and $\rho_{\star,\rm bo}(a,b) = \mathbb{P}(f_\star = a, \hat{f}_{\rm bo} = b)$. These densities correspond to ρ summed over $\hat{f}_{\rm bo}$, $f_\star$ and $\hat{f}_{\rm erm}$ respectively. Here the sample $\boldsymbol{x}$ is understood as any sample from the test set, on which the confidence/uncertainty is evaluated. It is important that equation (7) is defined for the same sample $\boldsymbol{x}$ in all the 3 arguments. Note that $\rho_{*,\rm erm}$ allows to compare the ERM uncertainty with the oracle uncertainty (the best we could do if we had infinite data), while $\rho_{\rm bo,erm}$ quantifies the ERM uncertainty with respect to the best statistical estimate under a finite amount of data.

In the next section, we provide a characterization of this joint density in the high-dimensional limit where $n, d \to \infty$ with fixed sample complexity $\alpha = n/d$, as a function of the noise level τ and regularization λ. To obtain this result we leverage recent works on approximate message passing (AMP) algorithms and their state evolution.

Some of our results will be conveniently formulated in terms of so-called calibration of a probabilistic classifier $\hat{f}: \mathbb{R}^d \to [0,1]$ defined as:

$$\Delta_p(\hat{f}) := p - \mathbb{E}_{\boldsymbol{x},y^*}(f_\star(\boldsymbol{x})|\hat{f}(\boldsymbol{x}) = p) \tag{8}$$

where $\hat{f}$ can be the logistic classifier or the BO one. Intuitively, the calibration quantifies how well the predictor assigns probabilities to events. If $\Delta_p = 0$ the predictor is said to be *calibrated at level p*. Instead, if for $p > 1/2, \Delta_p > 0$ (respectively $\Delta_p < 0$), then the predictor is said to be *overconfident* (respectively *underconfident*) Note, however, that the calibration is an average notion, while the above joint probability distribution (7) captures more detailed information about the point-wise confidence and its reliability. In this work, we will also consider the calibration of ERM with respect to Bayes

$$\tilde{\Delta}_p := p - \mathbb{E}_{\boldsymbol{x},y^*}(\hat{f}_{\rm bo}(\boldsymbol{x})|\hat{f}_{\rm erm}(\boldsymbol{x}) = p). \tag{9}$$

Finally, while our discussion focuses in the calibration for concreteness, note that many other uncertainty metrics could be studied from the joint density equation (7).

Algorithm 1. GAMP.

Input: Data $\mathrm{X} \in \mathbb{R}^{n\times d}$, $\boldsymbol{y} \in \{-1,1\}^n$
Define $\mathrm{X}^2 = \mathrm{X}\odot\mathrm{X} \in \mathbb{R}^{n\times d}$ and Initialize $\hat{\boldsymbol{w}}^{t=0} = \mathcal{N}(\boldsymbol{0}, \sigma_w^2 \mathrm{I}_d)$, $\hat{\boldsymbol{c}}^{t=0} = \boldsymbol{1}_d$, $\boldsymbol{g}^{t=0} = \boldsymbol{0}_n$.
for $t \leqslant t_{\max}$ **do**
$\boldsymbol{V}^t = \mathrm{X}^2\hat{\boldsymbol{c}}^t$; $\boldsymbol{\omega}^t = \mathrm{X}\hat{\boldsymbol{w}}^t - \boldsymbol{V}^t \odot \boldsymbol{g}^{t-1}$; /* *Update channel mean and variance* */
$\boldsymbol{g}^t = f_{\text{out}}(\boldsymbol{y}, \boldsymbol{w}^t, \boldsymbol{V}^t)$; $\partial \boldsymbol{g}^t = \partial_\omega f_{\text{out}}(\boldsymbol{y}, \boldsymbol{w}^t, \boldsymbol{V}^t)$; /* *Update channel* */
$\boldsymbol{A}^t = -{\mathrm{X}^2}^\top \partial \boldsymbol{g}^t$; $\boldsymbol{b}^t = \mathrm{X}^\top \boldsymbol{g}^t + \boldsymbol{A}^t \odot \hat{\boldsymbol{w}}^t$; /* *Update prior mean and variance* */
/* *Update marginals* */
$\hat{\boldsymbol{w}}^{t+1} = f_w(\boldsymbol{b}^t, \boldsymbol{A}^t) := (\mathrm{I}_d + \boldsymbol{A}^t)^{-1}\boldsymbol{b}^t$; $\hat{\boldsymbol{c}}^{t+1} = \partial_b f_w(\boldsymbol{b}^t, \boldsymbol{A}^t) := (\mathrm{I}_d + \boldsymbol{A}^t)^{-1}$
end for
Return: Estimators $\hat{\boldsymbol{w}}_{\text{amp}}, \hat{\boldsymbol{c}}_{\text{amp}} \in \mathbb{R}^d$

3. Technical theorems

Our first technical result is the existence of an efficient algorithm (Algorithm 1), called GAMP) [31, 56] that is able to accurately estimate $\hat{f}_{\text{bo}}(\boldsymbol{x})$ in high-dimensions. The asymptotic accuracy of GAMP for the BO average (over the samples) test error is know from [11]. In order to formulate our results we also need to prove that the probabilities estimated by GAMP are also accurate *sample-wise*, this relatively straightforward extension of the results of [11] is covered by the following lemma:

Lemma 3.1 (Sample-wise GAMP-Optimality). *For a sequence of problems given by equation (2), and given the estimator $\hat{\boldsymbol{w}}_{\text{amp}}$ from algorithm 1, the predictor*

$$\hat{f}_{\text{amp}}(\boldsymbol{x}) = \mathbb{P}(y=1|\boldsymbol{x}) = \sigma_\star\left(\frac{\hat{\boldsymbol{w}}_{\text{amp}}^\top \boldsymbol{x}}{\sqrt{\tau^2 + \hat{\boldsymbol{c}}_{\text{amp}}^\top(\boldsymbol{x}\odot\boldsymbol{x})}}\right) \tag{10}$$

is such that, with high probability over a new sample $\boldsymbol{x}$ the classifier above is asymptotically equal to the Bayesian estimator $\hat{f}_{\text{bo}}(\boldsymbol{x}) = \mathbb{P}(y=1|\boldsymbol{x}) = \hat{f}_{\text{amp}}(\boldsymbol{x})$ in equation (5). More precisely:

$$\forall \varepsilon > 0, \lim_{d\to\infty} \mathbb{P}_{\boldsymbol{x},\mathcal{D}}\left(|\hat{f}_{\text{amp}}(\boldsymbol{x}) - \hat{f}_{\text{bo}}(\boldsymbol{x})|^2 \leqslant \varepsilon\right) \to 1. \tag{11}$$

In particular, the predictor $\hat{f}_{\text{amp}}$ asymptotically achieves the best possible test performance (the one achieved by the BO estimator)

The proof of lemma 3.1 is provided in appendix B. As mentioned above, the lemma does not require the prior on $\boldsymbol{w}_\star$ to be Gaussian. Changing the prior of $\boldsymbol{w}_\star$ amount to changing the denoising functions $(f_w, \partial_b f_w)$ in Algorithm 1. Similarly, the probit likelihood defined in equation (2) is not required for our analysis. In fact, the equations hold for any probabilistic generalized linear model, and in particular for the logit data model studied in [9], reproduced in appendix D. This choice of likelihood function only changes the denoising *channel* functions $(f_{\text{out}}, \partial_\omega f_{\text{out}})$. The motivation behind the use of the GAMP Algorithm is twofold. First, it allows us to characterize the posterior mean needed to express the probability $\hat{f}_{\text{amp}}(\boldsymbol{x})$ for a given new sample $\boldsymbol{x}$ in polynomial time in d. Indeed, each iteration of the loop in Algorithm 1 is $O(d^2)$. Second, the asymptotic performance of GAMP is conveniently tracked by low-dimensional *state evolution* equations which can be easily solved in a computer.

Our second technical result is a formula for the joint distribution of the teacher label, its Bayes estimate, and the estimate from empirical risk minimization defined in equation (7), described in the following theorem:

Theorem 3.2. *Consider training data $\mathcal{D} = \{(\boldsymbol{x}^\mu, y^\mu)\}_{\mu=1}^n$ sampled from the model defined in equation (2). Let $\hat{\boldsymbol{w}}_{\text{erm}} \in \mathbb{R}^d$ be the solution of the empirical risk minimization (4) and $\hat{\boldsymbol{w}}_{\text{amp}}$ denote the estimator returned by running algorithm 1 on the data $\mathcal{D}$. Then in the high-dimensional limit where $n, d \to \infty$ with $\alpha = n/d$ fixed, the asymptotic joint density (7) is given by:*

$$\rho(a,b,c) = \tau'\tau \frac{\mathcal{N}\left(\begin{bmatrix}\tau\cdot\sigma_\star^{-1}(a)\\ \tau'\cdot\sigma_\star^{-1}(b)\\ \sigma^{-1}(c)\end{bmatrix}\middle|\, 0_3, \Sigma\right)}{|\sigma_\star'(\sigma_\star^{-1}(a))||\sigma_\star'(\sigma_\star^{-1}(b))||\sigma'(\sigma^{-1}(c))|} \tag{12}$$

where we noted

$$\tau'^2 = \tau^2 + 1 - q_{\text{bo}}, \qquad \Sigma = \begin{bmatrix} 1 & q_{\text{bo}} & m \\ q_{\text{bo}} & q_{\text{bo}} & m \\ m & m & q_{\text{erm}} \end{bmatrix} \tag{13}$$

and the so-called overlaps:

$$q_{\text{bo}} = \frac{1}{d}\hat{\boldsymbol{w}}_{\text{amp}}^{\top}\boldsymbol{w}_\star = \frac{1}{d}||\hat{\boldsymbol{w}}_{\text{amp}}||_2^2 \tag{14}$$

$$m = \frac{1}{d}\hat{\boldsymbol{w}}_{\text{erm}}^{\top}\boldsymbol{w}_\star, \quad q_{\text{erm}} = \frac{1}{d}||\hat{\boldsymbol{w}}_{\text{erm}}||_2^2 \tag{15}$$

solve the following set of self-consistent equations:

$$\frac{1}{q_{\text{bo}}} = 1 + \alpha\, \mathbb{E}_{(z,\eta),\xi}\left[f_{\text{out}}(f_0(z+\tau\xi),\eta, 1-q_{\text{bo}})^2\right], \tag{16}$$

and

$$V = \frac{1}{\lambda + \hat{V}}, \qquad q_{\text{erm}} = \frac{\hat{m}^2 + \hat{q}}{(\lambda+\hat{V})^2}, \qquad m = \frac{\hat{m}}{\lambda+\hat{V}}. \tag{17}$$

$$\begin{cases} \hat{V} &= -\alpha \mathbb{E}_{(z,\omega),\xi}\left[\partial_\omega f_{\text{erm}}(f_0(z+\tau\xi),\omega,V)\right] \\ \hat{q} &= \alpha \mathbb{E}_{(z,\omega),\xi}\left[f_{\text{erm}}(f_0(z+\tau\xi),\omega,V)^2\right] \\ \hat{m} &= \alpha \mathbb{E}_{(z,\omega),\xi}\left[f_{\text{erm}}(f_0(z+\tau\xi),\omega,V)\right] \end{cases} \tag{18}$$

where $(z,\eta,\omega) \sim \mathcal{N}(0_3,\Sigma)$, $\xi \sim \mathcal{N}(0,1)$ *and the thresholding functions are defined as*

$$\begin{aligned} f_{\text{out}}(y,\omega,V) &= \frac{2y\,\mathcal{N}(\omega y|0,V+\tau^2)}{\text{erfc}\left(-\frac{y\omega}{\sqrt{2(\tau^2+V)}}\right)} \\ f_{\text{erm}}(y,w,V) &= V^{-1}\left(\text{prox}_{Vl(y,.)}(w) - w\right) \end{aligned} \tag{19}$$

with $\text{prox}_{\tau f}(x) = argmin_z\left(1/2\tau\|z-x\|_2^2 + f(z)\right)$ *being the proximal operator.*

In appendix A we show how this result can be deduced directly from the heuristic cavity method, and the analysis of the GAMP state evolution to compute the overlaps of ERM and BO estimators. To compute the correlation between the ERM and BO estimators, we use the Nishimori identity [30, 68]. More details, as well as the formal proof, are given in appendix B.

Our third theorem is an asymptotic expression for the calibration error.

Theorem 3.3. *The analytical expression of the joint density ρ yields the following expression for the calibration Δ_p:*

$$\Delta_p(\hat{f}_{\text{erm}}) = p - \sigma_\star\left(\frac{m/q_{\text{erm}} \times \sigma^{-1}(p)}{\sqrt{1 - m^2/q_{\text{erm}} + \tau^2}}\right). \tag{20}$$

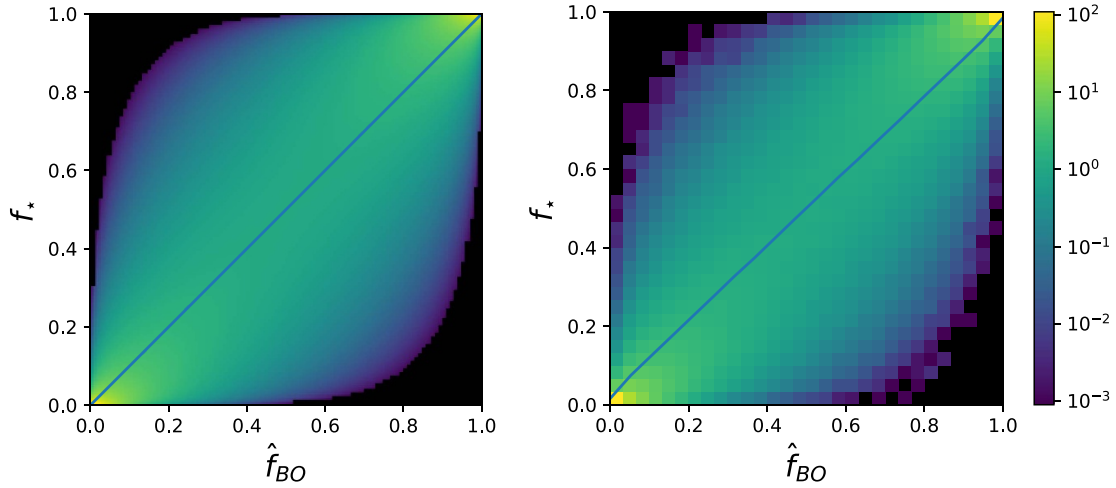
Moreover, the Bayesian classifier is always well calibrated with respect to the teacher, meaning:

$$\forall p \in [0,1], \quad \Delta_p(\hat{f}_{\text{bo}}) = 0\,. \tag{21}$$

Additionally, the calibration of ERM with respect to the Bayesian classifier and the oracle are equal:

$$\forall p \in [0,1], \quad \Delta_p(\hat{f}_{\text{erm}}) = \tilde{\Delta}_p\,. \tag{22}$$

The proof of theorem 3.3 is provided in appendix B.3. Equation (20) shows the different factors that influence Δ_p: the aleatoric uncertainty represented by the noise τ^2, the finiteness of data that appears through m/q_{erm} and m^2/q_{erm}, and the mismatch in the model with the activations $\sigma_\star, \sigma$. Moreover, equation (22) provides a recipe to compute the calibration Δ_p in the high-dimensional limit from the knowledge of the data model (2) only, but without knowing the specific realization of the weights $\boldsymbol{w}_\star$. This is because the quantities q_{bo}, q_{erm} and m self-average as $n, d \to \infty$, we then obtain the calibration Δ_p without knowing the realization of $\boldsymbol{w}_\star$.



(a) (b)

Figure 1. Theoretical prediction (left panel) and numerical estimation (right panel) of the joint density $\rho_{\mathrm{bo},\star}$ at $\alpha = 10$ and noise level $\tau = 0.5$. Numerical plot was done by running GAMP at dimension $d = 1000$, computing $(f_\star, \hat{f}_{\mathrm{bo}})$ on $n_{\mathrm{test}} = 10^7$ test samples. The blue curve is the mean of $f_\star$ given $\hat{f}_{\mathrm{bo}}$. For these parameters, the test error of Bayes is $\varepsilon_g^{\mathrm{bo}} = 0.173$, the oracle test error $\varepsilon_g^\star = 0.148$.

4. Results for uncertainty estimation

4.1. Bayes versus oracle uncertainty

We now discuss the consequences of the theorems from section 3. Figure 1 left panel depicts the theoretical prediction of the joint density $\rho_{\mathrm{bo},\star}$, between the Bayes posterior confidence/uncertainty $\hat{f}_{\mathrm{bo}}$ (x-axes) and the oracle confidence/uncertainty $f_\star$ (y-axes). The theoretically derived density (figure 1 left panel) is compared to its numerical estimation in figure 1 right panel, computed numerically using the GAMP algorithm. To estimate the numerical density in the right panel, we proceed as follows: after fixing the dimension d and the number of training samples $n = \alpha d$, GAMP is run on the training set. Once GAMP estimators have been obtained, n_{test} test samples are drawn and for each of them, we compute the confidence of the oracle/teacher $f_\star(\boldsymbol{x})$ from equation (2), and the Bayesian confidence $\hat{f}_{\mathrm{bo}}(\boldsymbol{x}) = \hat{f}_{\mathrm{amp}}(\boldsymbol{x})$ from theorem 3.1. Finally, we plot the histogram of the thus obtained joint density $\rho_{\mathrm{bo},\star}$ over the test samples. As the figure shows, there is a good agreement between theory and finite instance simulations.

We see that the density is positive on the whole support, it peaks around $(0, 0)$ and $(1, 1)$, but has a notable weight around the diagonal as well. The relatively large spread of the joint density is a consequence of the fact that on top of the intrinsic uncertainty of the teacher, the learning is only done with $n = \alpha d$ samples which brings an additional source of uncertainty captured in the Bayes estimator. Figure 1 thus quantifies this additional uncertainty due to finite α. We are not aware of something like this being done analytically in previous literature.

The blue curve is the mean of $f_\star$ conditioned on the values of $\hat{f}_{\mathrm{bo}}$. The difference between this and the diagonal is the calibration Δ_p defined in equation (8). We see that the figure illustrates $\Delta_p(\hat{f}_{\mathrm{bo}}) = 0$, i.e. the Bayesian prediction is well calibrated, as predicted by theorem 3.3.

Figure 2 then depicts the same densities as figure 1 for several different values of the sample complexity α and noise τ. The corresponding test error is given for information. We see, for instance, that at small α the BO confidence is low, close to 0.5, because not much can be learned from very few samples. The oracle confidence does not depend on α, and is low for growing τ. At large α, on the other hand, the BO confidence is getting well correlated with the oracle one. At larger α and small noise the BO test error is getting smaller and the corresponding confidence is close to 1 or 0 (depending on the label). The trends seen in this figure are expected, but again here we quantify them in an analytic form of equation (12) which as far as we know has not been done previously.

4.2. Logistic regression uncertainty and calibration

Having explicit access to the Bayesian confidence/uncertainty in a high-dimensional setting is a unique occasion to quantify the quality of the logistic classifier, which has its own natural measure of confidence

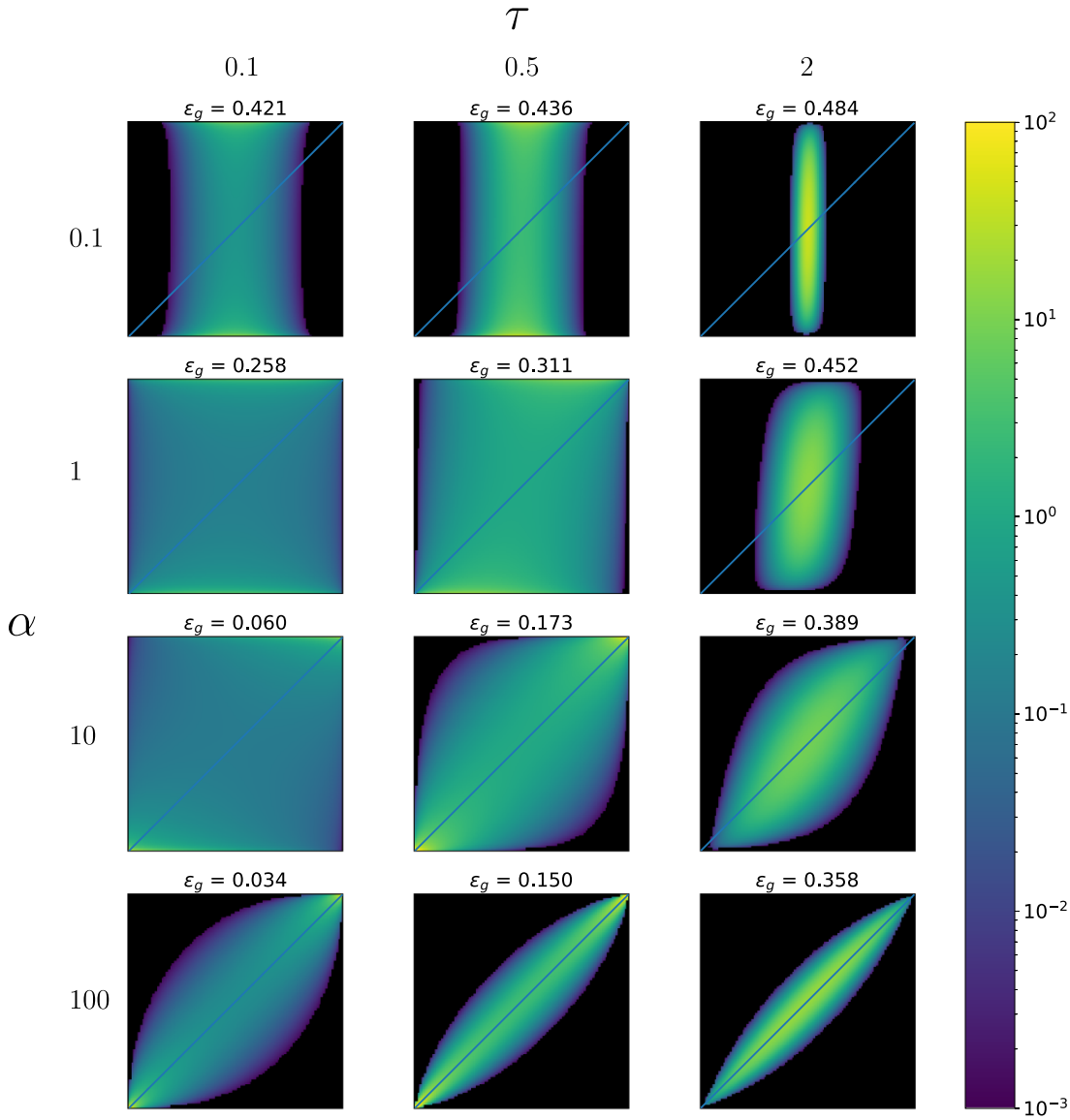


Figure 2. Density between Bayes confidence $\hat{f}_{\text{bo}}$ (x-axis) and the oracle confidence $f_\star$ (y-axis) for multiple values of α, τ: the rows correspond respectively to $\alpha = 0.1, 1, 10, 100$ from top to bottom, and the columns correspond respectively to $\tau = 0.1, 0.5, 2$. The blue curve is the mean of $f_\star$ given $\hat{f}_{\text{bo}}$. Generalization errors of the Bayes estimator are in written on top of the corresponding plot. The best possible generalization errors, achieved if the teacher weights are known, for $\tau = 0.1, 0.5, 2$ are respectively $\varepsilon_g^\star = 0.032, 0.148, 0.352$.

induced by the logit. How accurate is this measure? We start with the logistic classifier at zero regularization and then move to the regularized case in the next section.

Figure 3 compares the joint density of $(\hat{f}_{\text{erm}}, f_\star)$ (left panel), and $(\hat{f}_{\text{erm}}, \hat{f}_{\text{bo}})$ (right panel) with the same noise and number of samples as used in figure 1. The blue curves are the means of $f_\star$ (respectively $\hat{f}_{\text{bo}}$) conditioned on $\hat{f}_{\text{erm}}$, their shape is demonstrating that the (non-regularized) logistic classifier is on average overconfident, as is well known in practice.

The equality between these two blue curves illustrates theorem 3.3, equation (22): $\Delta_p(\hat{f}_{\text{erm}}) = \tilde{\Delta}_p$. Note, however, that while the calibrations of the ERM with respect to the oracle or the BO are equal, the conditional variances of $f_\star$ and $\hat{f}_{\text{bo}}$ are very different. This shows how the calibration is only a very partial fix of the confidence estimation for ERM: when $\hat{f}_{\text{erm}} = p$, both Bayes and the oracle's predictions will be $p - \Delta_p$ *on average*, but for the considered parameters the predictions of the oracle are much more spread around this value than those of Bayes estimator. This means that the ERM still captures rather well some part of the uncertainty coming from the limited number of samples. Figure 7 in the appendix C complements figure 3 by showing other values of α and τ.

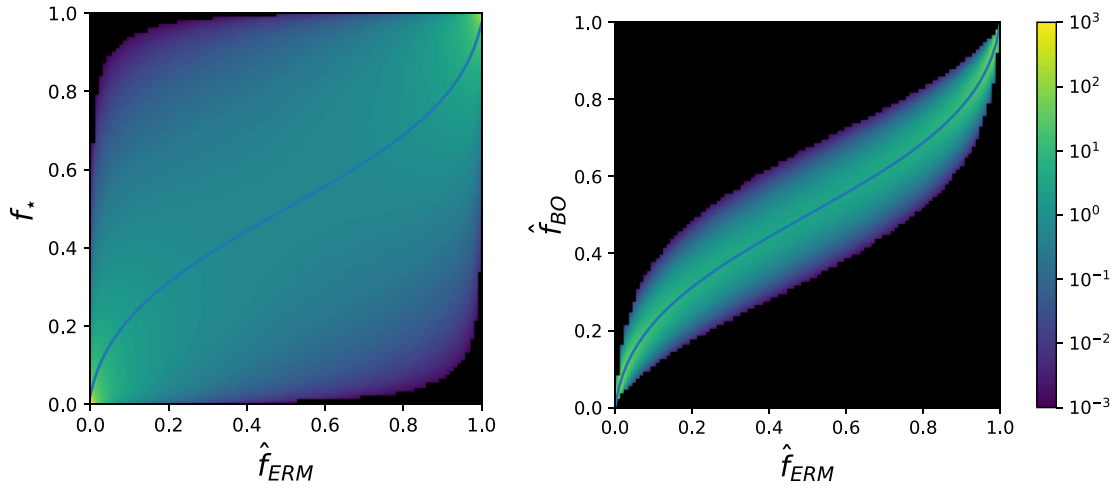


Figure 3. The probability density $\rho_{\mathrm{erm},\star}$ (left panel) and $\rho_{\mathrm{erm,bo}}$ (right panel), at $\alpha = 10$, $\tau = 0.5$ and $\lambda = 0^+$. The blue curves are the mean of the marginal distribution of $f_\star$ and $\hat{f}_{\mathrm{bo}}$ respectively under fixed $\hat{f}_{\mathrm{erm}}$, which are equal to $p - \Delta_p$ and $p - \tilde{\Delta}_p$. We observe overconfidence of the logistic classifier for these parameters. Test error of ERM is here $\varepsilon_g^{\mathrm{erm}} = 0.174$, very close to the one of BO $\varepsilon_g^{\mathrm{bo}} = 0.173$.

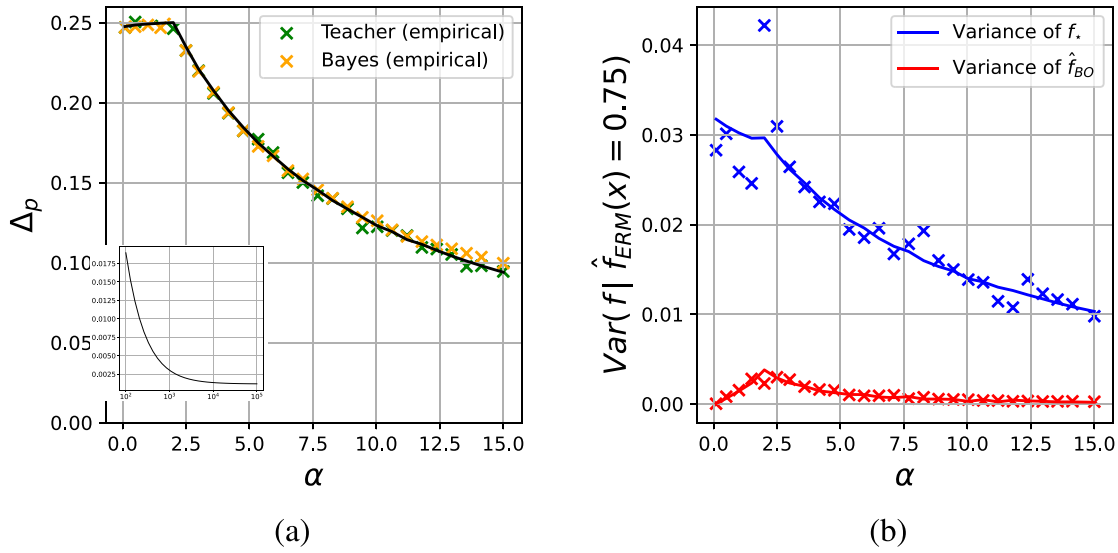


Figure 4. (a) Calibration of the logistic regression with $\lambda = 0^+, \tau = 2, p = 0.75$. Orange (respectively green) crosses are the numerical estimations of $\tilde{\Delta}_p$ (respectively Δ_p). Numerical values are obtained by averaging the calibration over 10 test sets of size $n_{\mathrm{test}} = 10^5$, at $d = 300$. Inset depicts the larger α behavior. (b) Variance of $f_\star$ and $\hat{f}_{\mathrm{bo}}$ conditioned on $\hat{f}_{\mathrm{ERM}} = p = 0.75$. Crosses are numerical values with the same parameters as figure (a). Though both $f_\star$ and $\hat{f}_{BO}$ have the same mean, their variance is significantly different.

We now investigate the calibration as a function of the sample complexity α. The plot (a) of figure 4 shows the curve Δ_p at $\lambda = 0^+$ computed using the analytical expression (20). The curve is compared to the numerical estimation of Δ_p (green crosses) and $\tilde{\Delta}_p$ (orange crosses). For a small dp, If we define $I_{p,\mathrm{d}p} = \{1 \leqslant i \leqslant n_{\mathrm{test}} | \hat{f}_{\mathrm{erm}}(x_i) \in [p, p + \mathrm{d}p]\}$, Δ_p and $\tilde{\Delta}_p$ are estimated experimentally with the formulas

$$\Delta_p \simeq p - \frac{\sum_{i \in I_{p,\mathrm{d}p}} f_\star(x_i)}{|I_{p,\mathrm{d}p}|}, \tilde{\Delta}_p \simeq p - \frac{\sum_{i \in I_{p,\mathrm{d}p}} \hat{f}_{\mathrm{bo}}(x_i)}{|I_{p,\mathrm{d}p}|}. \tag{23}$$

The calibrations Δ_p and $\tilde{\Delta}_p$ are both equal to the theoretical curve, further confirming the results of equation (22). Note the transition at $\alpha_c \sim 2.4$: for $\alpha < \alpha_c$, the training data is linearly separable. Since $\lambda = 0^+$, the empirical risk has no minimum and the estimator $\boldsymbol{w}_{\mathrm{erm}}$ diverges in norm. As a consequence, $\Delta_p \to p - 0.5$, as we observe on the plot. In the inset of figure 4 (left) we depict the theoretical curve

evaluated up to larger values of α. We see a saturation at about $\Delta_p \simeq 0.0011 \neq 0$. We note that in the work of [9] (partly reproduced in appendix D) the calibration was observed to go to 0 as $1/\alpha$. This difference is due to the mismatch between the function producing the data (probit) and the estimator (logit) in our case (whereas [9] used logit for both) which will generically be present in real data and thus the decay to zero observed in [9] is not expected to be seen generically.

Right panel of figure 4 displays the variance of $f_\star$ and $\hat{f}_{\rm bo}$ at fixed $\hat{f}_{\rm erm}$ as a function of α. This plot illustrates that the conditional variance of $f_\star$ is significantly higher than that of $\hat{f}_{\rm bo}$, as was previously noted in figure 3. This shows that the (non-regularized) logistic uncertainty captures rather decently the uncertainty due to the limited number of samples.

4.3. Effect of regularization on uncertainty and calibration

Logistic regression is rarely used in practice without regularization. In figures 8 and 10 in appendix C we depict the role of regularization on the density $\rho_{\rm erm,bo}$. As one would anticipate as the regularization strength grows the overconfidence of the logistic classifier at small λ becomes under-confidence at large λ.

One usually optimizes the strength λ of the ℓ_2 penalty through cross-validation. Ideally, we would choose λ that gives a low validation error but also that yields a well-calibrated estimator. The two main ways to choose λ is to minimize the validation $0/1$ classification error or the validation logistic loss. In our teacher-student setting, the classification error and logistic loss on test data can be computed exactly in the high-dimensional limit, using our state-evolution equations. We will thus define $\lambda_{\rm error}$ (respectively $\lambda_{\rm loss}$) as the minimizer of the expected $0/1$ classification error (respectively the logistic loss) for a new test sample. More precisely :

$$\begin{cases} \lambda_{\rm error} &= \arg\min_\lambda \mathbb{P}_{\vec{x},y}\left[y \neq \text{sign}(\hat{\vec{w}}(\lambda)^{\rm T}\vec{x})\right] \\ \lambda_{\rm loss} &= \arg\min_\lambda \mathbb{E}_{\vec{x},y}\left[-\log\sigma\left(y \times \hat{\vec{w}}(\lambda)^{\rm T}\vec{x}\right)\right] \end{cases} \tag{24}$$

where $\hat{\vec{w}}(\lambda)$ minimizes the empirical risk with regularization strength λ. Note that cross-validating λ on a validation set would induce fluctuations due to the finiteness of validation data. These fluctuations are not present when defining λ with equation (24). In the setting of the present paper, these two values of regularization lead to a very close test error/loss. In other words, choosing one or another of these λ seems to have little effect on the test performance of logistic regression.

Figure 5 plots the calibration Δ_p in the noiseless (left panel) and noisy (right panel) settings. We observe that for most parameters ERM with $\lambda_{\rm loss}$ is significantly less overconfident than with $\lambda_{\rm error}$. However, for larger values of α and τ we observe the opposite.

We also note that for small α the logistic regression at $\lambda_{\rm loss}$ even gets mildly underconfident, $\Delta_p < 0$. The bottom panels of the figure depict the corresponding variance. Interestingly we see that in both cases, despite a better calibration, $\lambda_{\rm loss}$ yields a higher variance than $\lambda_{\rm error}$ hence its point-wise estimates of uncertainty are not necessarily better.

Figure 6 shows $\rho_{\rm bo,erm}$ evaluated at $\lambda_{\rm error}$ and $\lambda_{\rm loss}$. Comparing the upper panels to figure 3 (at $\lambda = 0$), it is clear that choosing λ to optimize the error (and the loss) improves calibration. In the lower panels of figure 6 we can also see that the calibration at $\lambda_{\rm loss}$ (right panel) is better, i.e. the blue line is closer to $y = x$, than the one at $\lambda_{\rm error}$ (left panel). We conclude that using optimal regularization is clearly advantageous to obtain better-calibrated classification. However, we also note that the interplay between the mean of the distribution (the calibration) and its variance is subtle and more investigation is needed into designing a model-agnostic method where both are optimal simultaneously.

5. Discussion

This paper leverages the properties of the GAMP algorithm and associated closed-form control of the posterior marginals to provide a detailed theoretical analysis of uncertainty in a simple probit model. We investigate the relations between the respective uncertainties of the oracle, Bayes and regularized logistic regression. We see this as a grounding step for a line of future work that will leverage recent extensions of the GAMP algorithm and its associated analysis to multi-layer neural networks [7, 24], learning with random features and kernels [20, 22, 48], estimation under generative priors [6, 8], classification on more realistic models of data [25, 26, 58], etc. The present methodology is not restricted to classification and can be used for a more thorough study of confidence intervals in high-dimensional regression, extending [10]. This is left for further studies. The code of this project is available on Github [69].

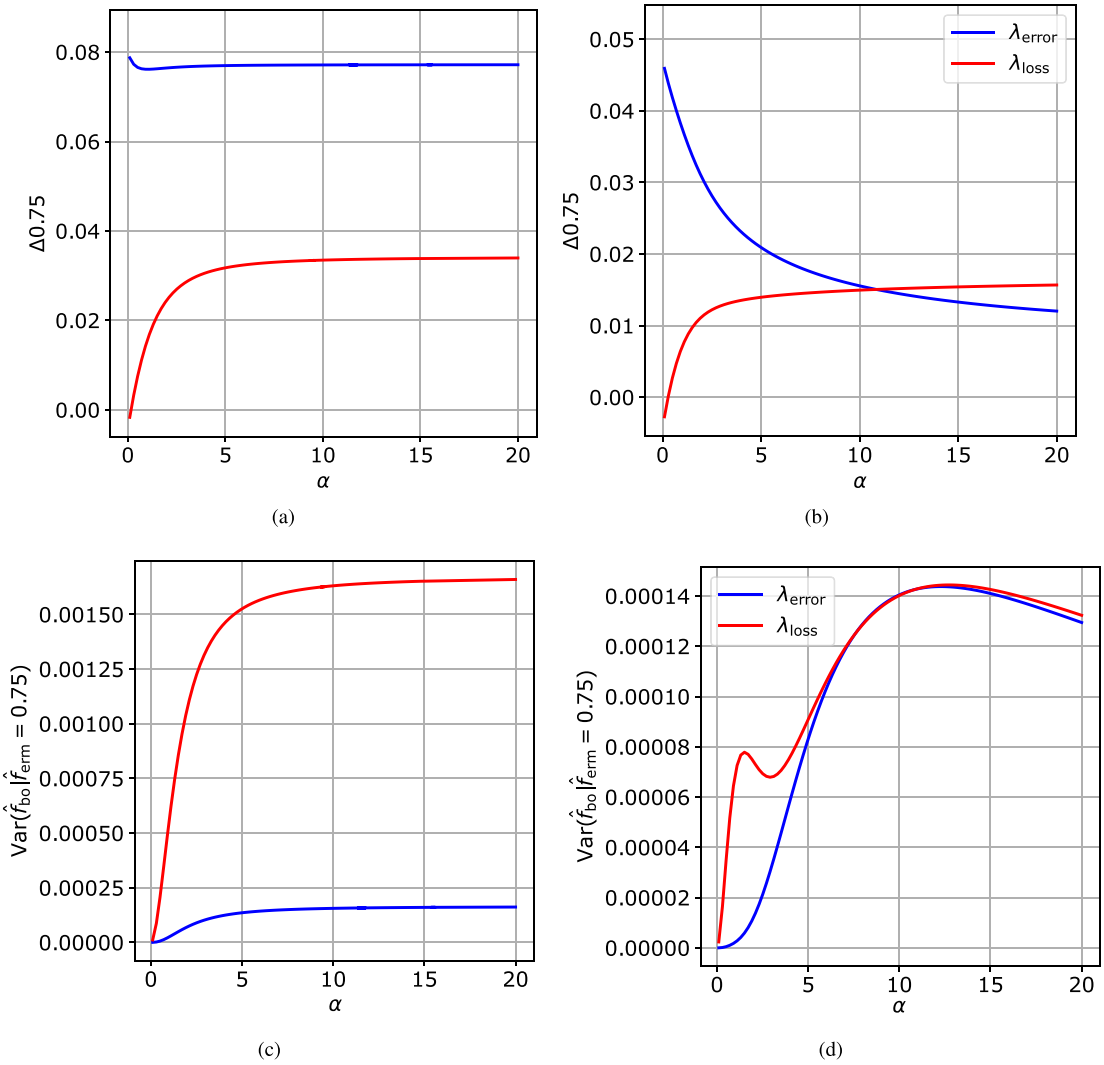


Figure 5. (Top) The calibration $\Delta_{0.75}(\hat{f}_{\text{erm}})$ as a function of α with $\lambda = \lambda_{\text{error}}(\alpha, \tau)$ (blue curve) and $\lambda = \lambda_{\text{loss}}(\alpha, \tau)$ (red curve)s. (Bottom) Variance of $\hat{f}_{\text{bo}}$ conditioned on $\hat{f}_{\text{erm}}(\boldsymbol{x}) = 0.75$ with λ_{error} and λ_{loss}. In (a) and (c), $\tau = 0$; in (b) and (d), $\tau = 0.5$.

Data availability statement

All data that support the findings of this study are included within the article (and any supplementary files).

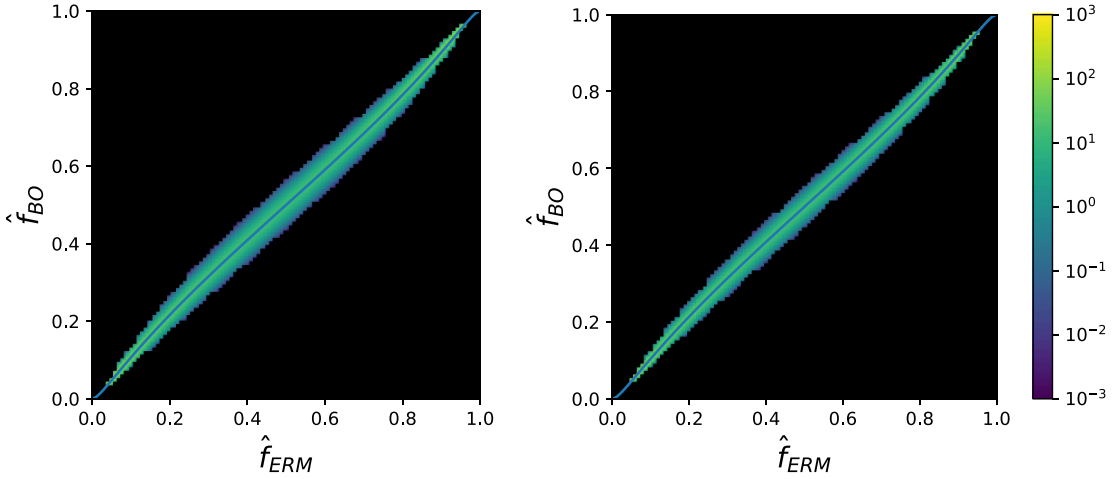
Acknowledgments

We thank Cédric Gerbelot for useful discussion and Benjamin Aubin for his help on the numerical experiments. We acknowledge funding from the ERC under the European Union's Horizon 2020 Research and Innovation Programme Grant Agreement 714608-SMiLe.

Appendix A. Cavity derivation of the analytical results

In this appendix, we sketch how the self-consistent equations (16) and (18) characterizing the sufficient statistics $(q_{\text{bo}}, m, q_{\text{erm}})$ can actually be derived via the heuristic cavity method [49, 50] from statistical physics.

We shall use the notation of Rangan's GAMP algorithm [56] and present our results as a derivation of GAMP algorithm from cavity, or belief propagation, as in [68]. This allows to connect all our results as well as the state evolution equations of the GAMP 1 algorithm in a single framework. Note that in its most general form, GAMP can be used both as an algorithm for estimating the marginals of the posterior distribution $\boldsymbol{w}_{\text{amp}} = \mathbb{E}[w|\mathcal{D}]$ or to minimize the empirical risk in (4)—the only difference between the two being the choice of denoising functions (f_{out}, f_w).



(a) $\lambda_{\text{error}} = 0.0976$, $\varepsilon_g = 0.1732$

(b) $\lambda_{\text{loss}} = 0.0980$, $\varepsilon_g = 0.1734$

(c) $\lambda_{\text{error}} = 0.0039$, $\varepsilon_g = 0.0843$

(d) $\lambda_{\text{loss}} = 0.0096$, $\varepsilon_g = 0.0847$

Figure 6. Density $\rho_{\text{erm,bo}}$ for different α, τ. Top row: $\alpha = 10, \tau = 0.5$. Bayes test error is $\varepsilon_g^{\text{bo}} = 0.1731$. Plot (a) (respectively (b)) is done at $\lambda = \lambda_{\text{error}}$ (respectively $\lambda = \lambda_{\text{loss}}$). Bottom row: $\alpha = 5, \tau = 0$, $\varepsilon_g^{\text{bo}} = 0.0839$. Plot (c) (respectively (d)) is done at $\lambda = \lambda_{\text{error}}$ (respectively $\lambda = \lambda_{\text{loss}}$). On the bottom row, we can clearly see that the calibration is better for λ_{loss}. Generalization errors of ERM as well as the values of the regularizations are indicated below the plots.

The novelty of our approach consists of running two GAMP algorithms in parallel *on the same instance* of data $\mathcal{D} = \{(\boldsymbol{x}^\mu, y^\mu)\}_{\mu=1}^n$ drawn from the probit model introduced in equation (2). Although we run the two versions of GAMP independently, they are correlated through the data $\mathcal{D}$—and our goal is to characterize exactly their joint distribution.

A.1. Joint state evolution

Consider we are running two AMPs in parallel, one for BO estimation and one for ERM. To distinguish both messages, we will denote ERM messages with a tilde: $\tilde{\omega}^t, \tilde{V}^t$, etc. To derive the asymptotic distribution of the estimators $(\hat{\boldsymbol{w}}_{\text{amp}}, \hat{\boldsymbol{w}}_{\text{erm}})$, it is more convenient to start from a close cousin of AMP: the reduced Belief Propagation equations (rBP). Note that in the high-dimensional limit that we are interested in this manuscript, rBP is equivalent to AMP, see for instance [7] or [8] for a detailed derivation. Written in coordinates, the rBP equations are given by:

$$\begin{cases} \omega^t_{\mu\to i} = \sum_{j\neq i} x_j^\mu \hat{w}^t_{j\to\mu} \\ V^t_{\mu\to i} = \sum_{j\neq i} (x_j^\mu)^2 \hat{c}^t_{j\to\mu} \end{cases}, \qquad \begin{cases} g^t_{\mu\to i} = f_{\text{out}}(y^\mu, \omega^t_{\mu\to i}, V^t_{\mu\to i}) \\ \partial g^t_{\mu\to i} = \partial_\omega f_{\text{out}}(y^\mu, \omega^t_{\mu\to i}, V^t_{\mu\to i}) \end{cases} \tag{25}$$

$$\begin{cases} b^t_{\mu\to i} = \sum_{\nu\neq\mu} x^\nu_i g^t_{\nu\to i} \\ A^t_{\mu\to i} = -\sum_{\nu\neq\mu} (x^\nu_i)^2 \partial g^t_{\nu\to i} \end{cases}, \qquad \begin{cases} \hat{w}^{t+1}_{i\to\mu} f_w(b^t_{i\to\mu}, A^t_{i\to\mu}) \\ \hat{c}^{t+1}_{i\to\mu} \partial_b f_w(b^t_{\mu\to i}, A^t_{\mu\to i}) \end{cases} \tag{26}$$

where (f_{out}, f_w) denote the denoising functions that could be associated either to BO or ERM estimation, and that can be generically written in terms of an estimation likelihood P_{out} and prior P_w as:

$$\begin{cases} f_{\text{out}}(y,\omega,V) &= \partial_\omega \log \mathcal{Z}_{\text{out}}(y,\omega,V) \\ \mathcal{Z}_{\text{out}}(y,\omega,V) &= \int_{\mathbb{R}} \frac{\mathrm{dx}}{\sqrt{2\pi V}} e^{-\frac{(x-\omega)^2}{2V}} P_{\text{out}}(y|x) \end{cases}, \qquad \begin{cases} f_w(b,A) &= \partial_b \log \mathcal{Z}_w(b,A) \\ \mathcal{Z}_w(b,A) &= \int_{\mathbb{R}} \mathrm{d}w\, P_w(w) e^{-\frac{1}{2}Aw^2 + bw} \end{cases}. \tag{27}$$

By assumption, the rBP messages are independent from each other, and since we are running both BO and ERM independently, they are only coupled to each other through the data, which has been generated by the same data model:

$$y^\mu \sim P_0(\cdot|\boldsymbol{w}_\star^\top \boldsymbol{x}^\mu), \qquad \boldsymbol{x}^\mu \sim \mathcal{N}(0, 1/d\mathrm{I}_d), \qquad \boldsymbol{w}_\star \sim \prod_{i=1}^d P_0(w_{\star i}). \tag{28}$$

Note that here we work in a more general setting than the one in the main manuscript (2). Indeed, the derivation presented here work for *any* factorized distribution of teacher weights $\boldsymbol{w}_\star$ and any likelihood P_0 (of which the probit is a particular case). For convenience, define the so-called *teacher local field*:

$$z_\mu = \sum_{j=1}^d x^\mu_j w_{\star j}. \tag{29}$$

Step 1: Asymptotic joint distribution of $(z_\mu, \omega^t_{\mu\to i}, \tilde{\omega}^t_{\mu\to i})$

Note that $(z_\mu, \omega^t_{\mu\to i}, \tilde{\omega}^t_{\mu\to i})$ are given by a sum of independent random variables with variance $d^{-1/2}$, and therefore by the Central Limit Theorem in the limit $d\to\infty$ they are asymptotically Gaussian. Therefore we only need to compute their means, variances and cross correlation. The means are straightforward, since x^μ_i have mean zero and therefore they will also have mean zero. The variances are given by:

$$\begin{aligned} \mathbb{E}\left[z^2_\mu\right] &= \mathbb{E}\left[\sum_{i=1}^d \sum_{j=1}^d x^\mu_i x^\mu_j w_{\star i} w_{\star j}\right] = \sum_{i=1}^d \sum_{j=1}^d \mathbb{E}\left[x^\mu_i x^\mu_j\right] w_{\star i} w_{\star j} = \frac{1}{d}\sum_{i=1}^d \sum_{j=1}^d \delta_{ij} w_{\star i} w_{\star j} \\ &= \frac{||\boldsymbol{w}_\star||_2^2}{d} \underset{d\to\infty}{\to} \rho \end{aligned} \tag{30}$$

$$\begin{aligned} \mathbb{E}\left[\left(\omega^t_{\mu\to i}\right)^2\right] &= \mathbb{E}\left[\sum_{j\neq i}^d \sum_{k\neq i}^d x^\mu_j x^\mu_k \hat{w}^t_{j\to\mu} \hat{w}^t_{k\to\mu}\right] = \sum_{j\neq i}^d \sum_{k\neq i}^d \mathbb{E}\left[x^\mu_j x^\mu_k\right] \hat{w}^t_{j\to\mu} \hat{w}^t_{k\to\mu} \\ &= \frac{1}{d}\sum_{j\neq i}^d \sum_{k\neq i}^d \delta_{jk} \hat{w}^t_{j\to\mu} \hat{w}^t_{k\to\mu} = \frac{1}{d}\sum_{j\neq i}^d \left(\hat{w}^t_{j\to\mu}\right)^2 = \frac{||\hat{\boldsymbol{w}}^t||_2^2}{d} - \frac{1}{d}\left(\hat{w}^t_{i\to\mu}\right)^2 \underset{d\to\infty}{\to} q^t \end{aligned} \tag{31}$$

$$\begin{aligned} \mathbb{E}\left[z_\mu \omega^t_{\mu\to i}\right] &= \mathbb{E}\left[\sum_{j\neq i}^d \sum_{k=1}^d x^\mu_j x^\mu_k \hat{w}^t_{j\to\mu} w_{\star k}\right] = \sum_{j\neq i}^d \sum_{k=1}^d \mathbb{E}\left[x^\mu_j x^\mu_k\right] \hat{w}^t_{j\to\mu} w_{\star k} \\ &= \frac{1}{d}\sum_{j\neq i}^d \sum_{k=1}^d \delta_{jk} \hat{w}^t_{j\to\mu} w_{\star k} = \frac{1}{d}\sum_{j\neq i}^d \hat{w}^t_{j\to\mu} w_{\star j} = \frac{\hat{\boldsymbol{w}}^t \cdot \boldsymbol{w}_\star}{d} - \frac{1}{d}\hat{w}^t_{i\to\mu} w_{\star i} \underset{d\to\infty}{\to} m^t \end{aligned} \tag{32}$$

$$\begin{aligned} \mathbb{E}\left[\omega^t_{\mu\to i} \tilde{\omega}^t_{\mu\to i}\right] &= \mathbb{E}\left[\sum_{j\neq i}^d \sum_{k\neq i}^d x^\mu_j x^\mu_k \hat{w}^t_{j\to\mu} \tilde{\hat{w}}^t_{k\to\mu}\right] \\ &= \sum_{j\neq i}^d \sum_{k\neq i}^d \mathbb{E}\left[x^\mu_j x^\mu_k\right] \hat{w}^t_{j\to\mu} \tilde{\hat{w}}^t_{k\to\mu} = \frac{1}{d}\sum_{j\neq i}^d \sum_{k\neq i}^d \delta_{jk} \hat{w}^t_{j\to\mu} \tilde{\hat{w}}^t_{k\to\mu} \\ &= \frac{1}{d}\sum_{j\neq i}^d \hat{w}^t_{j\to\mu} \tilde{\hat{w}}^t_{j\to\mu} = \frac{\hat{\boldsymbol{w}}^t \cdot \tilde{\hat{\boldsymbol{w}}}^t}{d} - \frac{1}{d}\hat{w}^t_{i\to\mu} \tilde{\hat{w}}^t_{i\to\mu} \underset{d\to\infty}{\to} Q^t \end{aligned} \tag{33}$$

where we have used that $\hat{w}^t_{i\to\mu} = O(d^{-1/2})$ to simplify the sums at large d. Summarizing our findings:

$$(z_\mu, \omega^t_{\mu\to i}, \tilde{\omega}^t_{\mu\to i}) \sim \mathcal{N}\left(\mathbf{0}_3, \begin{bmatrix} \rho & m^t & \tilde{m}^t \\ m^t & q^t & Q^t \\ \tilde{m}^t & Q^t & \tilde{q}^t \end{bmatrix}\right) \tag{34}$$

with:

$$\begin{aligned} \rho &\equiv \frac{1}{d}||\boldsymbol{w}_\star||^2, & q^t &\equiv \frac{1}{d}||\hat{\boldsymbol{w}}^t_{\mathrm{BO}}||^2, & \tilde{q}^t &\equiv \frac{1}{d}||\hat{\boldsymbol{w}}^t_{\mathrm{ERM}}||^2 \\ m^t &\equiv \frac{1}{d}\hat{\boldsymbol{w}}_{\mathrm{BO}}\cdot\boldsymbol{w}_\star, & \tilde{m}^t &\equiv \frac{1}{d}\hat{\boldsymbol{w}}_{\mathrm{ERM}}\cdot\boldsymbol{w}_\star, & Q^t &\equiv \frac{1}{d}\hat{\boldsymbol{w}}_{\mathrm{BO}}\cdot\hat{\boldsymbol{w}}_{\mathrm{ERM}}. \end{aligned} \tag{35}$$

Step 2: Concentration of variances $V^t_{\mu\to i}$, $\tilde{V}^t_{\mu\to i}$

Since the variances $V^t_{\mu\to i}, \tilde{V}^t_{\mu\to i}$ depend on $(x^\mu_i)^2$, in the asymptotic limit $d\to\infty$ they concentrate around their means:

$$\mathbb{E}\left[V^t_{\mu\to i}\right] = \sum_{j\neq i}\mathbb{E}\left[(x^\mu_i)^2\right]\hat{c}^t_{j\to\mu} = \frac{1}{d}\sum_{j\neq i}\hat{c}^t_{j\to\mu} = \frac{1}{d}\sum_{j=1}^{d}\hat{c}^t_{j\to\mu} - \frac{1}{d}\hat{c}^t_{i\to\mu} \underset{d\to\infty}{\to} V^t \equiv \frac{1}{d}\sum_{j=1}^{d}\hat{c}^t_j \tag{36}$$

where we have defined the variance overlap V^t. The same argument can be used for $\tilde{V}^t_{\mu\to\mathrm{in}}$. Summarizing, asymptotically we have:

$$V^t_{\mu\to i}\to V^t, \qquad \tilde{V}^t_{\mu\to i}\to\tilde{V}^t. \tag{37}$$

Step 3: Distribution of $b^t_{\mu\to i}$, $\tilde{b}^t_{\mu\to i}$

By definition, we have

$$b^t_{\mu\to i} = \sum_{\nu\neq\mu} x^\nu_i g^t_{\nu\to i} = \sum_{\nu\neq\mu} x^\nu_i f_{\mathrm{out}}(y^\mu, \omega^t_{\nu\to i}, V^t_{\nu\to i}) = \sum_{\nu\neq\mu} x^\nu_i f_{\mathrm{out}}(f_0(z_\nu+\tau\xi_\nu), \omega^t_{\nu\to i}, V^t_{\nu\to i}) \tag{38}$$

Note that in the sum $z_\nu = \sum_{j=1}^{d} x^\nu_j w_{\star j}$ there is a term $i=j$, and therefore z_μ is correlated with x^ν_i. To make this explicit, we split the teacher local field:

$$z_\mu = \sum_{j=1}^{d} x^\mu_j w_{\star j} = \underbrace{\sum_{j\neq i} x^\mu_j w_{\star j}}_{z_{\mu\to i}} + x^\mu_i w_{\star i} \tag{39}$$

and note that $z_{\mu\to i} = O(1)$ is independent from x^ν_i. Since $x^\mu_i w_{\star i} = O(d^{-1/2})$, to take the average at leading order, we can expand the denoising function:

$$\begin{aligned} f_{\mathrm{out}}(f_0(z_\mu+\tau\xi_\mu), \omega^t_{\nu\to i}, V^t_{\nu\to i}) = {} & f_{\mathrm{out}}(f_0(z_{\nu\to i}+\tau\xi_\nu), \omega^t_{\nu\to i}, V^t_{\nu\to i}) \\ & + \partial_z f_{\mathrm{out}}(f_0(z_{\nu\to i}+\tau\xi_\nu), \omega^t_{\nu\to i}, V^t_{\nu\to i}) x^\nu_i w_{\star i} + O(d^{-1}). \end{aligned} \tag{40}$$

Inserting in the expression for $b^t_{\mu\to i}$,

$$\begin{aligned} b^t_{\mu\to i} = {} & \sum_{\nu\neq\mu} x^\nu_i f_{\mathrm{out}}(f_0(z_{\nu\to i}+\tau\xi_\nu), \omega^t_{\nu\to i}, V^t_{\nu\to i}) \\ & + \sum_{\nu\neq\mu}(x^\nu_i)^2 \partial_z f_{\mathrm{out}}(f_0(z_{\nu\to i}+\tau\xi_\nu), \omega^t_{\nu\to i}, V^t_{\nu\to i}) w_{\star i} + O(d^{-3/2}). \end{aligned} \tag{41}$$

Therefore:

$$\begin{aligned} \mathbb{E}\left[b^t_{\mu\to i}\right] &= \frac{w_{\star i}}{d}\sum_{\nu\neq\mu}\partial_z f_{\mathrm{out}}(f_0(z_{\nu\to i}+\tau\xi_\nu), \omega^t_{\nu\to i}, V^t_{\nu\to i}) + O(d^{-3/2}) \\ &= \frac{w_{\star i}}{d}\sum_{\nu=1}^{n}\partial_z f_{\mathrm{out}}(f_0(z_{\nu\to i}+\tau\xi_\nu), \omega^t_{\nu\to i}, V^t_{\nu\to i}) + O(d^{-3/2}). \end{aligned} \tag{42}$$

Note that as $d \to \infty$, for fixed t and for all ν, the fields $(z_{\nu\to i}, \omega^t_{\nu\to i})$ are identically distributed according to average in equation (34). Therefore,

$$\frac{1}{d}\sum_{\nu=1}^{n} \partial_z f_{\text{out}}(f_0(z_{\nu\to i}+\tau\xi_\nu), \omega^t_{\nu\to i}, V^t_{\nu\to i}) \underset{d\to\infty}{\to} \alpha\, \mathbb{E}_{(\omega,z),\xi}\left[\partial_z f_{\text{out}}(f_0(z+\tau\xi), \omega, V^t)\right] \equiv \hat{m}^t \tag{43}$$

so:

$$\mathbb{E}\left[b^t_{\mu\to i}\right] \underset{d\to\infty}{\to} w_{\star i}\hat{m}^t. \tag{44}$$

Similarly, the variance is given by:

$$\begin{aligned}
&\text{Var}\left[b^t_{\mu\to i}\right] && (45)\\
&= \sum_{\nu\neq\mu}\sum_{\kappa\neq\mu} \mathbb{E}\left[x^\nu_i x^\kappa_i\right] f_{\text{out}}(f_0(z_{\nu\to i}+\tau\xi_\nu), \omega^t_{\nu\to i}, V^t_{\nu\to i}) f_{\text{out}}(f_0(z_{\kappa\to i}+\tau\xi_\kappa), \omega^t_{\kappa\to i}, V^t_{\kappa\to i}) + O(d^{-2})\\
&= \frac{1}{d}\sum_{\nu\neq\mu} f_{\text{out}}(f_0(z_{\nu\to i}+\tau\xi_\nu), \omega^t_{\nu\to i}, V^t_{\nu\to i})^2 + O(d^{-2})\\
&= \frac{1}{d}\sum_{\nu=1}^{n} f_{\text{out}}(f_0(z_{\nu\to i}+\tau\xi_\nu), \omega^t_{\nu\to i}, V^t_{\nu\to i})^2 + O(d^{-2})\\
&\underset{d\to\infty}{\to} \alpha\, \mathbb{E}_{(z,\omega),\xi}\left[f_{\text{out}}(f_0(z+\tau\xi), \omega, V^t)^2\right] \equiv \hat{q}^t. && (46)
\end{aligned}$$

The same discussion holds for the ERM. We now just need to compute the correlation between both fields:

$$\begin{aligned}
&\text{Cov}\left[b^t_{\mu\to i}, \tilde{b}^t_{\mu\to i}\right] && (47)\\
&= \sum_{\nu\neq\mu}\sum_{\kappa\neq\mu} \mathbb{E}\left[x^\nu_i x^\kappa_i\right] f_{\text{out}}(f_0(z_{\nu\to i}+\tau\xi_\nu), \omega^t_{\nu\to i}, V^t_{\nu\to i}) \tilde{f}_{\text{out}}(f_0(z_{\kappa\to i}+\tau\xi_\kappa), \tilde{\omega}^t_{\kappa\to i}, \tilde{V}^t_{\kappa\to i}) + O(d^{-2})\\
&= \frac{1}{d}\sum_{\nu=1}^{n} f_{\text{out}}(f_0(z_{\nu\to i}+\tau\xi_\nu), \omega^t_{\nu\to i}, V^t_{\nu\to i}) \tilde{f}_{\text{out}}(f_0(z_{\nu\to i}+\tau\xi_\nu), \tilde{\omega}^t_{\nu\to i}, \tilde{V}^t_{\nu\to i}) + O(d^{-2})\\
&\underset{d\to\infty}{\to} \alpha\, \mathbb{E}_{(z,\omega,\tilde{\omega}),\xi}\left[f_{\text{out}}(f_0(z+\tau\xi), \omega, V^t) \tilde{f}_{\text{out}}(f_0(z+\tau\xi), \tilde{\omega}, \tilde{V}^t)\right] \equiv \hat{Q}^t. && (48)
\end{aligned}$$

To summarize, we have:

$$(b^t_{\mu\to i}, \tilde{b}^t_{\mu\to i}) \sim \mathcal{N}\left(w_{\star i}\begin{bmatrix}\hat{m}^t\\ \tilde{\hat{m}}^t\end{bmatrix}, \begin{bmatrix}\hat{q}^t & \hat{Q}^t\\ \hat{Q}^t & \tilde{\hat{q}}^t\end{bmatrix}\right). \tag{49}$$

Step 4: Concentration of $A^t_{\mu\to i}, \tilde{A}^t_{\mu\to i}$

The only missing piece is to determine the distribution of the prior variances $A^t_{\mu\to i}, \tilde{A}^t_{\mu\to i}$. Similar to the previous variance, they concentrate:

$$\begin{aligned}
A^t_{\mu\to i} &= -\sum_{\nu\neq\mu} (x^\nu_i)^2 \partial_\omega f_{\text{out}}(y^\nu, \omega^t_{\nu\to i}, V^t_{\nu\to i}) && (50)\\
&= -\sum_{\nu\neq\mu} (x^\nu_i)^2 \partial_\omega f_{\text{out}}(f_0(z_{\nu\to i}+\tau\xi_\nu), \omega^t_{\nu\to i}, V^t_{\nu\to i}) + O(d^{-3/2})\\
&= -\frac{1}{d}\sum_{\nu=1} \partial_\omega f_{\text{out}}(f_0(z_{\nu\to i}+\tau\xi_\nu), \omega^t_{\nu\to i}, V^t_{\nu\to i}) + O(d^{-3/2})\\
&\underset{d\to\infty}{\to} -\alpha\, \mathbb{E}_{(z,\omega),\xi}\left[\partial_\omega f_{\text{out}}(f_0(z+\tau\xi), \omega, V^t)\right] \equiv \hat{V}^t. && (51)
\end{aligned}$$

A.1.1. Summary

We now have all the ingredients we need to characterize the asymptotic distribution of the estimators:

$$\hat{\boldsymbol{w}}_{\mathrm{BO}} \sim f_{\mathrm{out}}\left(\boldsymbol{w}_\star \hat{m}^t + \sqrt{\hat{q}^t}\boldsymbol{\xi}, \hat{V}^t\right) \tag{52}$$

$$\hat{\boldsymbol{w}}_{\mathrm{ERM}} \sim \tilde{f}_{\mathrm{out}}\left(\boldsymbol{w}_\star \tilde{\hat{m}}^t + \sqrt{\tilde{\hat{q}}^t}\boldsymbol{\eta}, \tilde{\hat{V}}^t\right) \tag{53}$$

where $\boldsymbol{\eta}, \boldsymbol{\xi} \sim \mathcal{N}(\mathbf{0}, \mathrm{I}_d)$ are independent Gaussian variables. From that, we can recover the usual GAMP state evolution equations for the overlaps:

$$\begin{cases} V^{t+1} = \mathbb{E}_{(w_\star,\xi)}\left[\partial_b f_w(\hat{m}^t w_\star + \sqrt{\hat{q}^t}\xi, \hat{V}^t)\right] \\ q^{t+1} = \mathbb{E}_{(w_\star,\xi)}\left[f_w(\hat{m}^t w_\star + \sqrt{\hat{q}^t}\xi, \hat{V}^t)^2\right] \\ m^{t+1} = \mathbb{E}_{(w_\star,\xi)}\left[f_w(\hat{m}^t w_\star + \sqrt{\hat{q}^t}\xi, \hat{V}^t) w_{\star i}\right] \end{cases}, \quad \begin{cases} \hat{V}^t = -\alpha\mathbb{E}_{(z,\omega),\xi}\left[\partial_\omega f_{\mathrm{out}}(f_0(z+\tau\xi), \omega, V^t)\right] \\ \hat{q}^t = \alpha\mathbb{E}_{(z,\omega),\xi}\left[f_{\mathrm{out}}(f_0(z+\tau\xi), \omega, V^t)^2\right] \\ \hat{m}^t = \alpha\mathbb{E}_{(z,\omega),\xi}\left[f_{\mathrm{out}}(f_0(z+\tau\xi), \omega, V^t)\right] \end{cases} \tag{54}$$

which is also valid for the tilde variables. But we can also get a set of equations for the correlations:

$$\begin{cases} Q^t = \mathbb{E}_{w_\star,(b,\tilde{b})}\left[f_w(b, \hat{V}^t)\tilde{f}_w\left(\tilde{b}, \tilde{\hat{V}}^t\right)\right] \\ \hat{Q}^t = \alpha\mathbb{E}_{(z,\omega,\tilde{\omega}),\xi}\left[f_{\mathrm{out}}(f_0(z+\tau\xi), \omega, V^t)\tilde{f}_{\mathrm{out}}(f_0(z+\tau\xi), \tilde{\omega}, \tilde{V}^t)\right] \end{cases}. \tag{55}$$

A.2. Simplifications

A.2.1. Simplifying BO state evolution

State evolution of BO can be reduced to two equations. First, note that asymptotically

$$m := \frac{1}{d}\hat{\boldsymbol{w}}_{\mathrm{bo}} \cdot \boldsymbol{w}_\star = \frac{1}{d}\mathbb{E}_{\boldsymbol{w}_\star,\mathcal{D}}\left[\hat{\boldsymbol{w}}_{\mathrm{bo}} \cdot \boldsymbol{w}_\star\right]$$

with high probability. By Nishimori identity, the vector $\boldsymbol{w}_\star$ in the expectation can be replaced by an independent copy of the Bayesian posterior. This yields:

$$\frac{1}{d}\mathbb{E}_{\boldsymbol{w}_\star,\mathcal{D}}\left[\hat{\boldsymbol{w}}_{\mathrm{bo}} \cdot \boldsymbol{w}_\star\right] = \frac{1}{d}\mathbb{E}_{\mathcal{D}}\left[\hat{\boldsymbol{w}}_{\mathrm{bo}}\right] = q.$$

Hence $m = q$. Similarly, noting $\langle\cdot\rangle$ the average over the posterior distribution:

$$V = \frac{1}{d}\langle\|\boldsymbol{w} - \hat{\boldsymbol{w}}_{\mathrm{bo}}\|^2\rangle = \frac{1}{d}\mathbb{E}_{\mathcal{D}}\left[\langle\|\boldsymbol{w} - \hat{\boldsymbol{w}}_{\mathrm{bo}}\|^2\rangle\right] = \frac{1}{d}\mathbb{E}_{\mathcal{D}}\left[\langle\|\boldsymbol{w}\|^2\rangle\right] - \frac{1}{d}\mathbb{E}_{\mathcal{D}}\left[\hat{\boldsymbol{w}}_{\mathrm{bo}} \cdot \hat{\boldsymbol{w}}_{\mathrm{bo}}\right].$$

Like before, we used the fact that in asymptotically, $\langle\|\boldsymbol{w} - \hat{\boldsymbol{w}}_{\mathrm{bo}}\|^2\rangle$ concentrates around its mean. Using Nishimori, the first term is equal to $\mathbb{E}_{\boldsymbol{w}_\star}\left[\|\boldsymbol{w}_\star\|^2\right] = 1$. By definition, the second term is equal to q, thus $V = 1 - q$.

Using similar arguments, $\hat{m} = \hat{q} = \hat{V}$. Thus, the state evolution can be reduced to two equations on q and $\hat{q}$.

A.2.2. Simplifying the $Q, \hat{Q}$ equations

In fact, the Nishimori property also allow us to show that the cross-correlation $Q, \hat{Q}$ are the same as the overlaps $\tilde{m}, \tilde{\hat{m}}$, in a similar way to appendix A.2.1. Indeed,

$$Q = \frac{1}{d}\hat{\boldsymbol{w}}_{\mathrm{bo}} \cdot \hat{\boldsymbol{w}}_{\mathrm{erm}} = \frac{1}{d}\mathbb{E}_{\mathcal{D}}\left[\hat{\boldsymbol{w}}_{\mathrm{bo}} \cdot \hat{\boldsymbol{w}}_{\mathrm{erm}}\right] = \frac{1}{d}\mathbb{E}_{\boldsymbol{w}_\star,\mathcal{D}}\left[\boldsymbol{w}_\star \cdot \hat{\boldsymbol{w}}_{\mathrm{erm}}\right] = \tilde{m}. \tag{56}$$

Alternatively, we can also prove that directly showing that the iterations for Q^t are a stable orbit of $\tilde{m}^t$. Indeed, assume that at time step t we have $Q^t = \tilde{m}^t$ and $\hat{Q}^t = \tilde{\hat{m}}^t$. Then, focusing at our specific setting, at time $t+1$ we have:

$$Q^{t+1} = \mathbb{E}_{w_\star,b,\tilde{b}}[f_w(b, \hat{V})f_w(b, \hat{V})] = \mathbb{E}_{w_\star}\left[\frac{b}{\hat{V}+1}\frac{\tilde{b}}{\tilde{\hat{V}}+\lambda}\right] = \mathbb{E}_{w_\star}\left[\frac{\hat{Q} + \tilde{\hat{m}}\hat{m}}{(\hat{V}+1)(\tilde{\hat{V}}+\lambda)}\right]$$

$$= \mathbb{E}_{w_\star}\left[\frac{\tilde{\hat{m}}}{\tilde{\hat{V}}+\lambda}\right].$$

Because as we have shown above $\hat{m} = \hat{q}$ and $\hat{Q}^t = \tilde{\hat{m}}^t$. This is precisely the equation for $\tilde{m}$.

A.3. Evaluating the equations

A.3.1. BO

In BO estimation, the estimation likelihood P_{out} and prior P_w match exactly that of the generating model for data, which for the model (2) is:

$$P_{\text{out}}(y|x) = \frac{1}{2}\text{erfc}\left(-\frac{y\omega}{\sqrt{2\Delta}}\right), \qquad P_w(w) = \mathcal{N}(0,1). \tag{57}$$

Therefore, it is easy to show that:

$$\mathcal{Z}_{\text{out}}(y,\omega,V) = \frac{1}{2}\text{erfc}\left(-\frac{y\omega}{\sqrt{2(\tau^2+V)}}\right), \qquad Z_w(b,A) = \frac{e^{\frac{b^2}{1+A}}}{1+A} \tag{58}$$

and therefore:

$$f_{\text{out}}(y,\omega,V) = \frac{2y\,\mathcal{N}(\omega y|0,V+\tau^2)}{\text{erfc}\left(-\frac{y\omega}{\sqrt{2(\tau^2+V)}}\right)}, \qquad f_w(b,A) = \frac{b}{1+A}. \tag{59}$$

This form of the prior allow us to simplify some of the equations considerably:

$$q_{\text{bo}}^{t+1} = \mathbb{E}_{(w_\star,\xi)}\left[f_w(\hat{q}^t w_\star + \sqrt{\hat{q}^t}\xi, \hat{q}^t)^2\right] = \frac{1}{1+\hat{q}_{\text{bo}}^t} \tag{60}$$

which is the equation found in theorem 3.2. The other equation cannot be closed analytically, however it can be considerably simplified:

$$\hat{q}_{\text{bo}} = -\alpha\mathbb{E}_{(z,\omega),\xi}\left[\partial_\omega f_{\text{out}}(f_0(z+\tau\xi),\omega,V^t)\right] \tag{61}$$

$$= \frac{2}{\pi}\frac{\alpha}{1+\tau^2-q_{\text{bo}}^t}\int_{\mathbb{R}} \text{d}z\,\mathcal{N}\left(z\middle|0,\frac{q_{\text{bo}}^t}{2(1+\tau^2-q_{\text{bo}}^t)}\right)\frac{e^{-2z^2}}{\text{erfc}(z)\text{erfc}(-z)}. \tag{62}$$

A.4. ERM estimation

For ERM, the estimation likelihood P_{out} and prior P_w are related to the loss and penalty functions:

$$P_{\text{out}}(y|x) = e^{-\beta\ell(y,x)}, \qquad P_w(w) = e^{-\beta r(w)}. \tag{63}$$

where the parameter $\beta > 0$ is introduced for convenience, and should be taken to infinity. Focusing on the regularization part and redefining $(b,A) \to (\beta b, \beta A)$

$$\mathcal{Z}_w(b,A) = \int_{\mathbb{R}} \text{d}w\, e^{-\beta(\frac{A}{2}w^2 - bw + r(w))} \underset{\beta\to\infty}{\asymp} e^{\beta\left[\frac{b^2}{2A} - \mathcal{M}_{A^{-1}r}(A^{-1}b)\right]} \tag{64}$$

where we have used Laplace's method and defined the *Moreau envelope*:

$$\mathcal{M}_{\tau f}(x) = \min_{z\in\mathbb{R}}\left[\frac{1}{2\tau}(x-z)^2 + f(z)\right]. \tag{65}$$

Therefore,

$$f_w(b,A) = \lim_{\beta\to\infty}\frac{1}{\beta}\partial_b \log Z_w(b,A) = \text{prox}_{A^{-1}r}(A^{-1}b) \tag{66}$$

where we have defined the *proximal operator*:

$$\text{prox}_{\tau f}(x) = \underset{z\in\mathbb{R}}{\text{argmin}}\left[\frac{1}{2\tau}(x-z)^2 + f(z)\right] \tag{67}$$

and used the well-known property $\partial_x \mathcal{M}_{\tau f}(x) = -\frac{1}{\tau}\left(\text{prox}_{\tau f}(x) - x\right)$. In particular, for the ℓ_2-penalty $r(w) = \lambda/2w^2$, we have:

$$\text{prox}_{\lambda/2(\cdot)^2}(x) = \frac{x}{1+\lambda} \qquad \Leftrightarrow \qquad f_w(b,A) = \frac{b}{\lambda+A}. \tag{68}$$

The simple form of the regularization allow us to simplify the state evolution equations considerably:

$$\begin{cases} \tilde{V}^{t+1} &= \mathbb{E}_{(w_\star,\xi)}\left[\partial_b f_w(\hat{\tilde{m}}^t w_\star + \sqrt{\hat{\tilde{q}}^t}\xi, \hat{\tilde{V}}^t)\right] = \frac{1}{\lambda+\hat{\tilde{V}}} \\ \tilde{q}^{t+1} &= \mathbb{E}_{(w_\star,\xi)}\left[f_w(\hat{\tilde{m}}^t w_\star + \sqrt{\hat{\tilde{q}}^t}\xi, \hat{\tilde{V}}^t)^2\right] = \frac{\hat{m}^2+\hat{\tilde{q}}}{(\lambda+\hat{\tilde{V}})^2} \\ \tilde{m}^{t+1} &= \mathbb{E}_{(w_\star,\xi)}\left[f_w(\hat{\tilde{m}}^t w_\star + \sqrt{\hat{\tilde{q}}^t}\xi, \hat{\tilde{V}}^t) w_{\star i}\right] = \frac{\hat{m}}{\lambda+\hat{\tilde{V}}} \end{cases} \tag{69}$$

which are the equations found in theorem 3.2. A similar discussion can be carried for the loss term, and yields in general:

$$f_{\text{out}}(y,\omega,V) = V^{-1}\left(\text{prox}_{\tau\ell(y,\cdot)}(x) - x\right). \tag{70}$$

Unfortunately, the logistic loss $\ell(y,x) = \log(1+e^{-yx})$ does not admit a closed-form solution for the proximal, and therefore for a given (y,ω,V) we need to compute it numerically.

Appendix B. Proof of theorems

A possible route for proving our result is to give a rigorous proof of the cavity equations. Instead, we shall use a shortcut, and leverage on recent progress for both the ERM cavity results [5, 16, 38, 52, 61, 63]), the Bayes performances [11, 12], as well as on the performance of GAMP [24, 31, 56].

B.1. GAMP optimality

The optimally of GAMP is a direct consequence of the generic results concerning its performance (the state evolution in [31, 56]) and the characterization of the Bayes performance in [11]. G-a works, one considers a sequence of inference problems indexed by the dimension d, with data $\mathcal{D}_d$ (which are defined in section 2 for our purpose). As d increases, both GAMP performances and Bayes errors converge with high probability to the same deterministic limit given by the so-called 'replica', or 'state evolution' equations.

To simplify the notation, all our statements involving the asymptotic limit $d \to \infty$ are implicitly defined for such sequences, and the convergence is assumed to be in terms of probability.

Let us prove that, indeed, GAMP estimates for posterior probability are asymptotically exact with high probability. First, we note that the estimation of the Bayes posterior probability for the signs corresponds to finding the estimators that minimize the mean square error (MSE). Indeed consider, for fixed data (this remains true averaging over data), the mean squared error for an estimator $\hat{Y}(\mathbf{X})$:

$$\text{MSE}(\hat{Y}(\mathbf{X})) = \mathbb{E}_{Y,\mathbf{X}}\left[(Y-\hat{Y}(\mathbf{X}))^2\right] = \mathbb{E}_{\mathbf{X}}\mathbb{E}_{Y|\mathbf{X}}\left[(Y-\hat{Y}(\mathbf{X}))^2\right]. \tag{71}$$

The mean square error is given by using the posterior mean [17], as can be seen immediately differentiating with respect to $\hat{Y}$ (for a given $\mathbf{x}$), so that:

$$\hat{Y}_{\text{Bayes}}(\mathbf{x}) = \mathbb{E}_{Y|X=\mathbf{x}}[Y] = 2\mathbb{P}_{Y|X=\mathbf{x}}(Y=1) - 1. \tag{72}$$

The Bayes estimator for the posterior probability is thus the minimal mean square error (MMSE) estimator. We see here that the estimation of the posterior mean of Y is equivalent to the estimation of the probability it takes value one; both quantities are thus trivially related.

We can now use proposition 2, page 12 in [11], which shows that indeed GAMP efficiently achieves Bayes-optimality for the MMSE on Y:

Theorem B.1 (GAMP generalization error, [11]). *Consider a sequence of problems indexed by* d*, with data* $\mathcal{D}_d$ *in dimension* d*, then we have that GAMP estimator asymptotically achieves the Minimal Mean Square Error in estimating the error on new label* Y*. That is, with high probability:*

$$\lim_{d\to\infty} \mathbb{E}_{Y,\boldsymbol{X}|\mathcal{D}_d}\left[(Y-\hat{Y}_{\text{GAMP}}(\boldsymbol{X},\mathcal{D}_d)^2\right] = \text{MMSE}(Y) \tag{73}$$

where $\hat{Y}_{GAMP}(\boldsymbol{x},\mathcal{D}) = 2p-1$*, and* $p = \hat{f}^{AMP}(\boldsymbol{x})$ *(equation (10)), with* $\hat{\boldsymbol{c}}_{\text{amp}}^{\top}(\boldsymbol{x}\odot\boldsymbol{x}) = 1-q$*, with* q *a fixed point of (16).*

The fact that GAMP asymptotically achieves the MMSE, coupled with the uniqueness of the Bayes estimator, implies the GAMP estimator for p is arbitrarily close to the Bayes estimated for p, with high probability over new Gaussian samples, as $d \to \infty$. More precisely, we can use the following lemma:

Lemma B.2 (Bounds on differences of estimators for Y). *Consider a sequence of estimation problems indexed by* d *with data* $\mathcal{D}_d$. *If a (sequence of) estimators* $\hat{f}_d(\boldsymbol{x})$ *achieves the MMSE performance of* $\hat{g}_d^{Bayes}(\boldsymbol{x})$ *as* $d \to \infty$ *for Gaussian distributed* $\boldsymbol{x}$, *then*

$$\lim_{d\to\infty} \mathbb{E}_{\boldsymbol{X}} |f_d(\boldsymbol{X}) - g_d^{\text{Bayes}}(\boldsymbol{X})|^2 \to 0. \tag{74}$$

Proof. The Bayes estimator $g_d^{\text{Bayes}}()$ is the minimum of the MMSE, therefore for any other estimator $f_d(\mathbf{X})$ we have

$$\mathbb{E}\left[(Y - f_d(\mathbf{X})^2\right] \geqslant \mathbb{E}\left[(Y - g_d^{\text{Bayes}}(\mathbf{X})^2\right]. \tag{75}$$

We have, denoting $\delta_d(X) = f_d(\mathbf{X}) - g_d^{\text{Bayes}}(\mathbf{X})$

$$\mathbb{E}\left[(Y - f_d(\mathbf{X})^2\right] = \mathbb{E}\left[(Y - g_d^{\text{Bayes}}(\mathbf{X}) + \delta_d(X))^2\right] \tag{76}$$

$$= \text{MMSE} + \mathbb{E}\left[\delta_d(X)^2 + 2\delta_n(X)(Y - g_d^{\text{Bayes}}(X))\right] \tag{77}$$

$$= \text{MMSE} + \mathbb{E}\left[\delta_d(X)^2\right] + \mathbb{E}_{X,\mathcal{D}}\mathbb{E}_{Y|X,\mathcal{D}}\left[2\delta_d(X)(Y - g_d^{\text{Bayes}}(X))\right] \tag{78}$$

$$= \text{MMSE} + \mathbb{E}\left[\delta_d(X)^2\right] + \mathbb{E}_{X,\mathcal{D}}\left[2\delta_n(X)\mathbb{E}_{Y|X,\mathcal{D}}[Y - g_d^{\text{Bayes}}(X])\right] \tag{79}$$

$$= \text{MMSE} + \mathbb{E}\left[\delta_d(X)^2\right] \tag{80}$$

where we have used $g_d^{\text{Bayes}}(X) = \mathbb{E}_{Y|X,\mathcal{D}}[Y]$. Using the fact that the f_d asymptotically achieves MMSE optimality, we thus obtain:

$$\lim_{d\to\infty} \mathbb{E}_{Y,X,\mathcal{D}}\left[|f_d(X) - g_d^{\text{Bayes}}(X)|^2\right] \to 0. \tag{81}$$

□

Applying this lemma to the GAMP estimator leads to lemma 3.1: with high probability over new sample $\mathbf{x}$ and learning data $\mathcal{D}$, the GAMP estimate is asymptotically equivalent to the Bayes one.

B.2. Joint density of estimators

While a possible strategy to prove the second theorem would be to use state evolution to follow our joint GAMP algorithm (thus monitoring the Bayes *and* the ERM performance), we shall instead again leverage on recent progress on generic proofs of replica equations, in particular the Bayes one (in [11] and the ERM ones (that were the subject of many works recently [5, 16, 38, 52, 61, 63]). Again, all our statements involving the asymptotic limit $d \to \infty$ are implicitly defined for sequences of problems, and the convergence is assumed to be in terms of probability. We start with the following lemma:

Lemma (Joint distribution of pre-activation). *For a fixed set of data* $\mathcal{D}$, *consider the joint random variables (over* $\boldsymbol{X}$*)* $\nu = \boldsymbol{X}\cdot\boldsymbol{w}_\star, \lambda_{\text{erm}} = \boldsymbol{X}\cdot\hat{\boldsymbol{w}}_{\text{erm}}, \lambda_{\text{amp}} = \boldsymbol{X}\cdot\hat{\boldsymbol{w}}_{\text{amp}}$. *Then we have*

$$\mathbb{P}(\nu, \lambda_{\text{amp}}, \lambda_{\text{erm}}) = \mathcal{N}\left(0, \begin{pmatrix} \frac{\boldsymbol{w}_\star\cdot\boldsymbol{w}_\star}{d} & \frac{\boldsymbol{w}_\star\cdot\hat{\boldsymbol{w}}_{\text{amp}}}{d} & \frac{\boldsymbol{w}_\star\cdot\hat{\boldsymbol{w}}_{\text{erm}}}{d} \\ \frac{\hat{\boldsymbol{w}}_{\text{amp}}\cdot\boldsymbol{w}_\star}{d} & \frac{\hat{\boldsymbol{w}}_{\text{amp}}\cdot\hat{\boldsymbol{w}}_{\text{amp}}}{d} & \frac{\hat{\boldsymbol{w}}_{\text{amp}}\cdot\hat{\boldsymbol{w}}_{\text{erm}}}{d} \\ \frac{\hat{\boldsymbol{w}}_{\text{erm}}\cdot\boldsymbol{w}_\star}{d} & \frac{\hat{\boldsymbol{w}}_{\text{erm}}\cdot\hat{\boldsymbol{w}}_{\text{amp}}}{d} & \frac{\hat{\boldsymbol{w}}_{\text{erm}}\cdot\hat{\boldsymbol{w}}_{\text{erm}}}{d} \end{pmatrix}\right). \tag{82}$$

Proof. This is an immediate consequence of the Gaussianity of the new data $\mathbf{x}$, with covariance $\mathbb{I}/d$. □

We now would like to know the asymptotic limit of the parameters of this distribution, for large d. While we have $\frac{w_\star\cdot w_\star}{d} \to \rho$, the other overlap has a deterministic limit given by the replica equations. For empirical risk minimization, this has been proven in the aforementioned series of works, but we shall here use the notation of [38] and utilize use the following results:

Theorem B.4 (ERM overlaps [5, 38, 63]). *Consider a sequence of inference problems indexed by the dimension* d*, then with high probability:*

$$\lim_{d\to\infty} \frac{\hat{\boldsymbol{w}}_{\text{erm}} \cdot \boldsymbol{w}_\star}{d} \to m, \qquad \lim_{d\to\infty} \frac{\hat{\boldsymbol{w}}_{\text{erm}} \cdot \hat{\boldsymbol{w}}_{\text{erm}}}{d} \to q_{\text{erm}}. \tag{83}$$

With q_{erm} *and* m *solutions of the self-consistent equations (18) in the main text.*

GAMP is tracked by its state evolution [31], and is known to achieve the Bayes overlap:

Theorem B.5 (Bayes overlaps [11]). *Consider a sequence of inference problems indexed by the dimension* d*, then with high probability:*

$$\lim_{d\to\infty} \frac{\hat{\boldsymbol{w}}_{\text{amp}} \cdot \boldsymbol{w}_\star}{d} \to q_{\text{bo}}, \qquad \lim_{d\to\infty} \frac{\hat{\boldsymbol{w}}_{\text{amp}} \cdot \hat{\boldsymbol{w}_{\text{amp}}}}{d} \to q_{\text{bo}}. \tag{84}$$

With q_{bo} *given by the self-consistent equation (16).*

The only overlap left to control is thus $Q = \hat{\boldsymbol{w}}_{\text{amp}} \cdot \hat{\boldsymbol{w}}_{\text{erm}}/d$. We shall now prove that it is also concentrating, with high probability, to m. To do this, we first prove the following lemma for the overlap between the Bayes estimate $\boldsymbol{w}_{\text{bo}} = \mathbb{E}_{W|\mathcal{D}}[\boldsymbol{W}]$ and any other vector $\mathbf{V}$, possibly dependent on the data:

Lemma B.6 (Nishimori relation for Bayes overlaps).

$$\mathbb{E}_{\mathcal{D}}\left[\boldsymbol{w}_{\text{bo}} \cdot \boldsymbol{V}(\mathcal{D})\right] = \mathbb{E}_{\mathcal{D},W^*}\left[\boldsymbol{w}^* \cdot \boldsymbol{V}(\mathcal{D})\right]. \tag{85}$$

Proof. The proof is an application of Bayes' formula, and an example of what is often called a Nishimori equality in statistical physics:

$$\mathbb{E}_{\mathcal{D},W^*}\left[\mathbf{w}^* \cdot \mathbf{V}(\mathcal{D})\right] = \mathbb{E}_{\mathcal{D}} E_{W^*|\mathcal{D}}\left[\mathbf{w}^* \cdot \mathbf{V}(\mathcal{D})\right] \tag{86}$$

$$= \mathbb{E}_{\mathcal{D}}\left[(E_{W^*|\mathcal{D}}\mathbf{w}^*) \cdot \mathbf{V}(\mathcal{D})\right] = \mathbb{E}_{\mathcal{D}}\left[\boldsymbol{w}_{\text{bo}} \cdot \mathbf{V}(\mathcal{D})\right]. \tag{87}$$

□

From this lemma, we see immediately that, in expectation

$$\lim_{d\to\infty} \mathbb{E}\left[\frac{\boldsymbol{w}_{\text{erm}} \cdot \boldsymbol{w}_\star}{d}\right] = \lim_{d\to\infty} \mathbb{E}\left[\frac{\boldsymbol{w}_{\text{erm}} \cdot \boldsymbol{w}_{\text{bo}}}{d}\right] = m. \tag{88}$$

Additionally, we already know that the left-hand side concentrates. It is easy to see that the right-hand side does as well:

Lemma B.5 (Concentration of the overlap Q).

$$\lim_{d\to\infty} \mathbb{E}\left[\left(\frac{\boldsymbol{w}_{\text{bo}} \cdot \boldsymbol{w}_{\text{erm}}}{d}\right)^2\right] = \lim_{d\to\infty} \mathbb{E}\left[\frac{\boldsymbol{w}_{\text{bo}} \cdot \boldsymbol{w}_{\text{erm}}}{d}\right]^2. \tag{89}$$

Proof. The proof again uses the Nishimori identity.

$$\mathbb{E}\left[\left(\frac{\boldsymbol{w}_{\text{bo}} \cdot \boldsymbol{w}_{\text{erm}}}{d}\right)^2\right] = \mathbb{E}\left[\left(\frac{\boldsymbol{w}_{\text{bo}} \cdot \boldsymbol{w}_{\text{erm}}}{d}\right)\left(\frac{\boldsymbol{w}_{\text{bo}} \cdot \boldsymbol{w}_{\text{erm}}}{d}\right)\right] \tag{90}$$

$$= \mathbb{E}_{\mathcal{D}}\left[\left(\frac{\mathbb{E}_{W|\mathcal{D}} W \cdot \boldsymbol{w}_{\text{erm}}}{d}\right)\left(\frac{\mathbb{E}_{W|\mathcal{D}} W \cdot \boldsymbol{w}_{\text{erm}}}{d}\right)\right] \tag{91}$$

$$= \mathbb{E}_{\mathcal{D}}\mathbb{E}_{W_1,W_2|\mathcal{D}}\left[\left(\frac{W_1 \cdot \boldsymbol{w}_{\text{erm}}}{d}\right)\left(\frac{W_2 \cdot \boldsymbol{w}_{\text{erm}}}{d}\right)\right] \tag{92}$$

$$= \mathbb{E}_{\mathcal{D},\boldsymbol{W}^*}\left[\left(\frac{\boldsymbol{W}^* \cdot \boldsymbol{w}_{\text{erm}}}{d}\right)\left(\frac{\mathbb{E}_{W|\mathcal{D}} W \cdot \boldsymbol{w}_{\text{erm}}}{d}\right)\right] \tag{93}$$

$$= \mathbb{E}_{\mathcal{D},\boldsymbol{W}^*}\left[\left(\frac{\boldsymbol{W}^* \cdot \boldsymbol{w}_{\text{erm}}}{d}\right)\left(\frac{\boldsymbol{W}_{\text{bo}} \cdot \boldsymbol{w}_{\text{erm}}}{d}\right)\right]. \tag{94}$$

Then, from Cauchy–Schwartz we have

$$\mathbb{E}\left[\left(\frac{\boldsymbol{w}_{\rm bo}\cdot\boldsymbol{w}_{\rm erm}}{d}\right)^2\right]^2 \leqslant \mathbb{E}\left[\left(\frac{\boldsymbol{w}_{\rm bo}\cdot\boldsymbol{w}_{\rm erm}}{d}\right)^2\right]\mathbb{E}\left[\left(\frac{\boldsymbol{w}^*\cdot\boldsymbol{w}_{\rm erm}}{d}\right)^2\right] \tag{95}$$

$$\mathbb{E}\left[\left(\frac{\boldsymbol{w}_{\rm bo}\cdot\boldsymbol{w}_{\rm erm}}{d}\right)^2\right] \leqslant \mathbb{E}\left[\left(\frac{\boldsymbol{w}^*\cdot\boldsymbol{w}_{\rm erm}}{d}\right)^2\right] \tag{96}$$

and as $d\to\infty$, we can use the concentration of the right-hand side to m to obtain

$$\lim_{d\to\infty}\mathbb{E}\left[\left(\frac{\boldsymbol{w}_{\rm bo}\cdot\boldsymbol{w}_{\rm erm}}{d}\right)^2\right] \leqslant m^2 \tag{97}$$

so that, given the second moment has to be larger or equal to its (squared) mean:

$$\lim_{d\to\infty}\mathbb{E}\left[\left(\frac{\boldsymbol{w}_{\rm bo}\cdot\boldsymbol{w}_{\rm erm}}{d}\right)^2\right] = m^2. \tag{98}$$

□

We have thus proven that the overlap Q concentrates in quadratic mean to m as $d\to\infty$: with high probability, it is thus asymptotically equal to m. We shall now prove that $\boldsymbol{w}_{\rm bo}$ can be approximated by $\boldsymbol{w}_{\rm amp}$. In fact, given the concentration of overlap, it will be enough to prove that:

$$\lim_{d\to\infty}\mathbb{E}_{\mathcal{D}_d}\frac{\hat{\boldsymbol{W}}_{\rm amp}(\mathcal{D}_d)\cdot\boldsymbol{W}_{\rm erm}(\mathcal{D})}{d} = \lim_{d\to\infty}\mathbb{E}_{\mathcal{D}_d}\frac{\hat{\boldsymbol{W}}_{\rm bo}(\mathcal{D}_d)\cdot\boldsymbol{W}_{\rm erm}(\mathcal{D})}{d}. \tag{99}$$

This can be done in two steps. First, similarly as in section B.1, we use the fact that GAMP achieves Bayes optimality for the estimation of W^* [11]. This leads to the following lemma

Lemma B.8 (Bounds on differences of estimators for w).

$$\lim_{d\to\infty}\mathbb{E}_{\mathcal{D}}\frac{\|\boldsymbol{w}_{\rm amp}-\boldsymbol{w}_{\rm bo}\|^2}{d}\to 0. \tag{100}$$

Proof. The proof proceeds similarly as in lemma B.2. Denoting $\delta\boldsymbol{W}(\mathcal{D}) = \boldsymbol{w}_{\rm amp}(\mathcal{D}) - \boldsymbol{w}_{\rm bo}(\mathcal{D})$ we write

$$\mathbb{E}_{\mathcal{D},\boldsymbol{W}^*}\frac{\|\boldsymbol{W}_{\rm amp}(\mathcal{D})-\boldsymbol{W}^*\|_2^2}{d} = \mathbb{E}_{\mathcal{D},\boldsymbol{W}^*}\frac{\|\boldsymbol{W}_{\rm bo}(\mathcal{D})+\delta\boldsymbol{W}(\mathcal{D})-\boldsymbol{W}^*\|_2^2}{d} \tag{101}$$

$$= \mathbb{E}_{\mathcal{D},\boldsymbol{W}^*}\frac{\|\boldsymbol{W}_{\rm bo}(\mathcal{D})-\boldsymbol{W}^*\|_2^2}{d} + \mathbb{E}_{\mathcal{D}}\frac{\|\delta\boldsymbol{W}(\mathcal{D})\|_2^2}{d} + \frac{1}{d}2\mathbb{E}_{\mathcal{D}}\mathbb{E}_{\boldsymbol{W}^*|\mathcal{D}}\left[\delta w(\mathcal{D})\left(\boldsymbol{W}^*-\boldsymbol{W}_{\rm bo}\right)\right] \tag{102}$$

$$= \mathbb{E}_{\mathcal{D}}\frac{\|\delta\boldsymbol{W}(\mathcal{D})\|_2^2}{d}. \tag{103}$$

Using the optimality of GAMP for the MMSE yields the lemma. □

We can now prove the equality of overlaps

Lemma B.9.

$$\lim_{d\to\infty}\mathbb{E}_{\mathcal{D}_d}\frac{\hat{\boldsymbol{W}}_{\rm amp}(\mathcal{D}_d)\cdot\boldsymbol{V}(\mathcal{D})}{d} = \lim_{d\to\infty}\mathbb{E}_{\mathcal{D}_d}\frac{\hat{\boldsymbol{W}}_{\rm bo}(\mathcal{D}_d)\cdot\boldsymbol{V}(\mathcal{D})}{d}. \tag{104}$$

Proof. The proof is an application of Cauchy–Schwartz inequality:

$$\left|\mathbb{E}_{\mathcal{D}_d}\left[\frac{(\hat{\boldsymbol{W}}_{\rm amp}-\hat{\boldsymbol{W}}_{\rm bo})(\mathcal{D}_d)\cdot\mathbf{V}(\mathcal{D})}{d}\right]\right| \leqslant \sqrt{\mathbb{E}\frac{\|\mathbf{V}\|_2^2}{d}\mathbb{E}\frac{\|\boldsymbol{W}_{\rm bo}-\boldsymbol{W}_{\rm amp}\|_2^2}{d}} \tag{105}$$

taking the limit $d\to\infty$ yields the lemma. □

Applying the lemma to the ERM estimator, and using the concentration of overlaps, finally leads to

Lemma B.10 (Asymptotic Joint distribution of pre-activation). *Asymptotically, and with high probability over data, the joint distribution of pre-activation is asymptotically given by*

$$\lim_{d\to\infty} \mathbb{P}(\nu, \lambda_{\text{amp}}, \lambda_{\text{erm}}) = \mathcal{N}\left(0, \begin{pmatrix} \rho & q_{\text{bo}} & m \\ q_{\text{bo}} & q_{\text{bo}} & m \\ m & m & q_{\text{erm}} \end{pmatrix}\right). \tag{106}$$

To obtain theorem 3.2, one simply applies the change of variable

$$(\nu, \lambda_{\text{amp}}, \lambda_{\text{erm}}) \to (f_\star(\nu), \hat{f}_{\text{amp}}(\lambda_{\text{amp}}), \hat{f}_{\text{erm}}(\lambda_{\text{erm}})) \tag{107}$$

$$= \left(\sigma_\star(\nu/\tau), \sigma_\star(\lambda_{\text{amp}}/\tau'), \sigma(\lambda_{\text{erm}})\right). \tag{108}$$

B.3. Proof of theorem 3.3

B.3.1. Proof of equation (20)

Consider the local fields $(\nu, \lambda_{\text{erm}}, \lambda_{\text{amp}})$ defined in equation (82). As shown above, these local fields follow a Gaussian distribution with covariance matrix Σ given in equation (13). Then, $(\nu, \lambda_{\text{erm}})$ follows a bivariate Gaussian and the density of ν conditioned on $\hat{f}_{\text{erm}}(\boldsymbol{x}) = \sigma(\lambda_{\text{erm}}) = p$ follows the Gaussian distribution with mean $\mu = \frac{m}{q_{\text{erm}}}\sigma^{-1}(p)$ and variance $\nu^2 = \rho - \frac{m^2}{q_{\text{erm}}}$. Then,

$$\mathbb{E}_x\left[f_\star(\boldsymbol{x})|\hat{f}_{\text{erm}}(\boldsymbol{x}) = p\right] = \int \mathrm{d}\nu \frac{1}{2}\text{erfc}\left(-\frac{\nu}{\sqrt{2\tau^2}}\right)\mathcal{N}(\nu|\mu, \nu^2) \tag{109}$$

$$= \frac{1}{2}\text{erfc}\left(-\frac{\mu}{\sqrt{2(\tau^2+\nu^2)}}\right) = \frac{1}{2}\text{erfc}\left(-\frac{\frac{m}{q_{\text{erm}}}\sigma^{-1}(p)}{\sqrt{2(1-\frac{m^2}{q_{\text{erm}}}+\tau^2)}}\right) \tag{110}$$

$$= \sigma_\star\left(\frac{\frac{m}{q_{\text{erm}}}\sigma^{-1}(p)}{\sqrt{1-\frac{m^2}{q_{\text{erm}}}+\tau^2}}\right) \tag{111}$$

which yields equation (20). We used the property that, for any a, b,

$$\int \text{erf}(ax+b)\mathcal{N}(x|\mu,\sigma^2)\mathrm{d}x = \text{erf}\left(\frac{a\mu+b}{\sqrt{1+2a^2\sigma^2}}\right). \tag{112}$$

B.3.2. Proof of equation (21)

We use the same computation as in the previous paragraph: since the conditioned on the Bayes local field $\hat{f}_{\text{bo}}(\boldsymbol{x}) = \sigma_\star(\frac{\lambda_{\text{amp}}}{\sqrt{\tau^2+1-q_{\text{bo}}}}) = p$, the teacher local field is Gaussian with mean $\mu = \sqrt{\tau^2+1-q_{\text{bo}}}\sigma_\star^{-1}(p)$ and variance $\nu^2 = 1 - q_{\text{bo}}$. As before, we have

$$\mathbb{E}_{\boldsymbol{x}}\left[f_\star(\boldsymbol{x})|\hat{f}_{\text{bo}}(\boldsymbol{x}) = p\right] = \sigma_\star\left(\frac{\mu}{\sqrt{\tau^2+\nu^2}}\right) \tag{113}$$

$$= \sigma_\star\left(\frac{\sqrt{\tau^2+1-q_{\text{bo}}}\sigma_\star^{-1}(p)}{\sqrt{\tau^2+1-q_{\text{bo}}}}\right) = p. \tag{114}$$

Hence the result of equation (21).

B.3.3. Proof of equation (22)

The proof follows the same structure as the previous paragraphs: conditioned on $\sigma(\lambda_{\text{erm}}) = p$, the law of λ_{amp} is $\mathcal{N}(\frac{m}{q_{\text{erm}}}\sigma^{-1}(p), q_{\text{bo}} - \frac{m^2}{q_{\text{erm}}})$ and

$$\mathbb{E}_{\boldsymbol{x}}\left[\hat{f}_{\text{bo}}(\boldsymbol{x})|\hat{f}_{\text{erm}}(\boldsymbol{x}) = p\right] = \int \sigma_\star\left(\frac{-x}{\sqrt{\tau^2+1-q}}\right)\mathcal{N}\left(x|\frac{m}{q_{\text{erm}}}\sigma^{-1}(p), q_{\text{bo}} - \frac{m^2}{q_{\text{erm}}}\right) \tag{115}$$

$$= \sigma_\star\left(\frac{\frac{m}{q_{\text{erm}}}\sigma^{-1}(p)}{\sqrt{\tau^2+1-q_{\text{bo}}+(q_{\text{bo}}-\frac{m}{q_{\text{erm}}})}}\right) \tag{116}$$

$$= \sigma_\star\left(\frac{\frac{m}{q_{\text{erm}}}\sigma^{-1}(p)}{\sqrt{1-\frac{m^2}{q_{\text{erm}}}+\tau^2}}\right) = \mathbb{E}_x\left[f_\star(\boldsymbol{x})|\hat{f}_{\text{erm}}(\boldsymbol{x}) = p\right]. \tag{117}$$

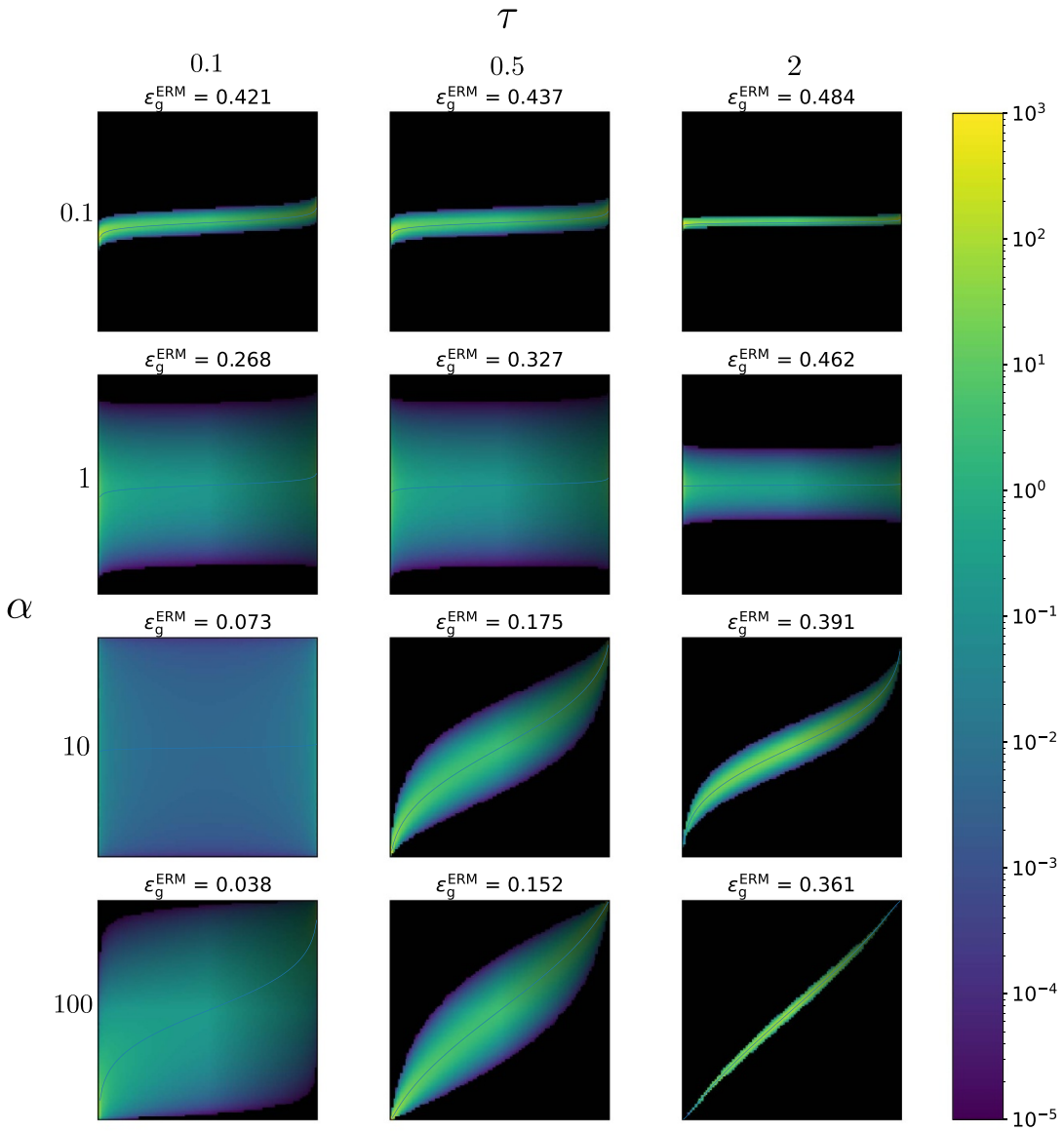


Figure 7. Joint density of $\hat{f}_{\text{erm}}$ (x-axis) and $\hat{f}_{\text{bo}}$ (y-axis), at $\lambda = 0^+$. Blue curve is the mean of $\hat{f}_{\text{bo}}$ at fixed $\hat{f}_{\text{erm}}$. The test error of ERM is indicated above the corresponding plot. The test errors of Bayes for the same parameters are indicated in figure 2.

Appendix C. Additional figures

C.1. Logistic regression uncertainty supplement

Figure 7 complements figure 3 from the main text by showing the same plot as the right panel in figure 3 for other values of sample complexity α and noise τ. We observe that at zero regularization the logistic regression is overconfident in all the depicted cases, in particular so at small α and small noise.

C.2. Choosing optimal regularization supplement

Here we give additional illustration related to the section 4.3 in the Main text.

In figure 8, the calibration Δ_p is shown as a function of λ at different levels p and different noise σ. First, observe that as λ grows the logistic regression is going from overconfident $\Delta_p > 0$ to underconfident $\Delta_p < 0$. For $\lambda \to \infty$, we have $\Delta_p \to p - 1$. Further, we observe that the value of λ at which the calibration is zero (the best calibration) has only mild dependence on the value of p. Finally, the vertical lines mark the values of regularization that minimize the validation error λ_{error}, and loss λ_{loss}. We see that λ_{loss} is closer to the well-calibrated region and that at small α this difference is more pronounced.

The left panel of figure 9 compares λ_{error} and λ_{loss} when $\tau = 0.5$.

The right panel of figure 9 then shows that the test error at λ_{loss} and λ_{error} are extremely close, with the difference being plot in the insert.

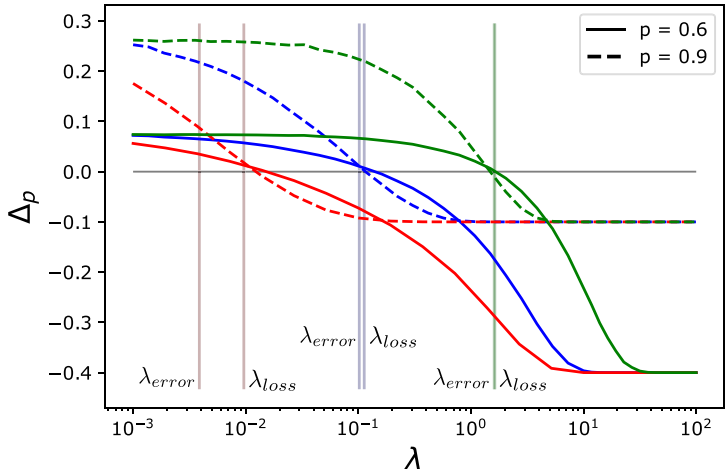


Figure 8. Calibration Δ_p for $p = 0.9$ and $p = 0.6$ as a function of λ, for $\tau = 0$ (red curve) , $\tau = 0.5$ (blue curve), and $\tau = 2$ (green curve), at $\alpha = 5$. Vertical lines correspond to λ_{error} and λ_{loss} defined in 4.3. For $\tau = 2$, λ_{error} and λ_{loss} differ by only 10^{-2} and look indistinguishable on the plot.

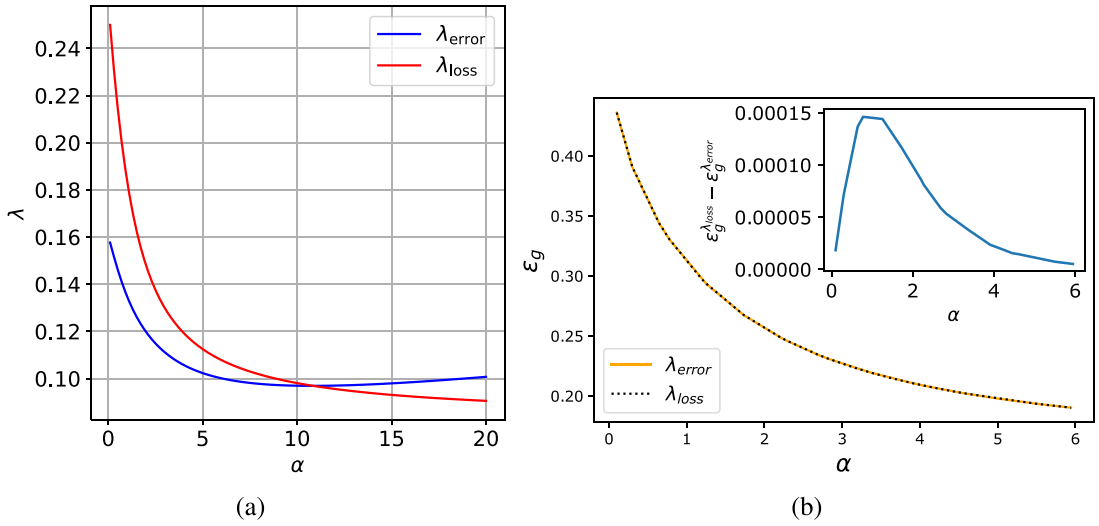


Figure 9. Left: optimal penalization for logistic regression as a function of the sample complexity α, for $\tau = 0.5$. Right: Test error at optimal λ for $\sigma = 0.5$, as a function of α. Orange line (respectively black dotted line) corresponds to λ-error (respectively λ-loss). The two curves are indistinguishable on the plot. The blue curve in the inset shows $\varepsilon_g^{\lambda_{\text{loss}}} - \varepsilon_g^{\lambda_{\text{error}}}$ as a function of α: it appears that the difference is around $\sim 10^{-4}$.

Figure 10 depicts the joint density of $\hat{f}_{\text{erm}}$ (x-axis) and $\hat{f}_{\text{bo}}$ (y-axis) for several values of the regularization λ and the noise τ. As λ increases, we observe that the logistic regression changes from overconfident to underconfident, as we could also observe in figure 8.

Next in figure 11 we depict the densities for λ_{error} and λ_{loss} for different values of α and τ. We observe an overall improvement in the calibration for these optimal regularizations.

Appendix D. Comparison to the data generated by logit model

As mentioned before, our state evolution equations can be adapted to data generated by the logit model, as studied in [9]. The likelihood is defined in equation (120). Since this change only concerns the data distribution, Algorithm 1 is unchanged. However, state evolution is changed in the update of $\hat{m}, \hat{q}, \hat{V}$: the partition function $\mathcal{Z}_0$ is now

$$\mathcal{Z}_0(y, \omega, V) = \int \mathrm{d}z \, \sigma(y \times z) \mathcal{N}(z|\omega, V). \tag{118}$$

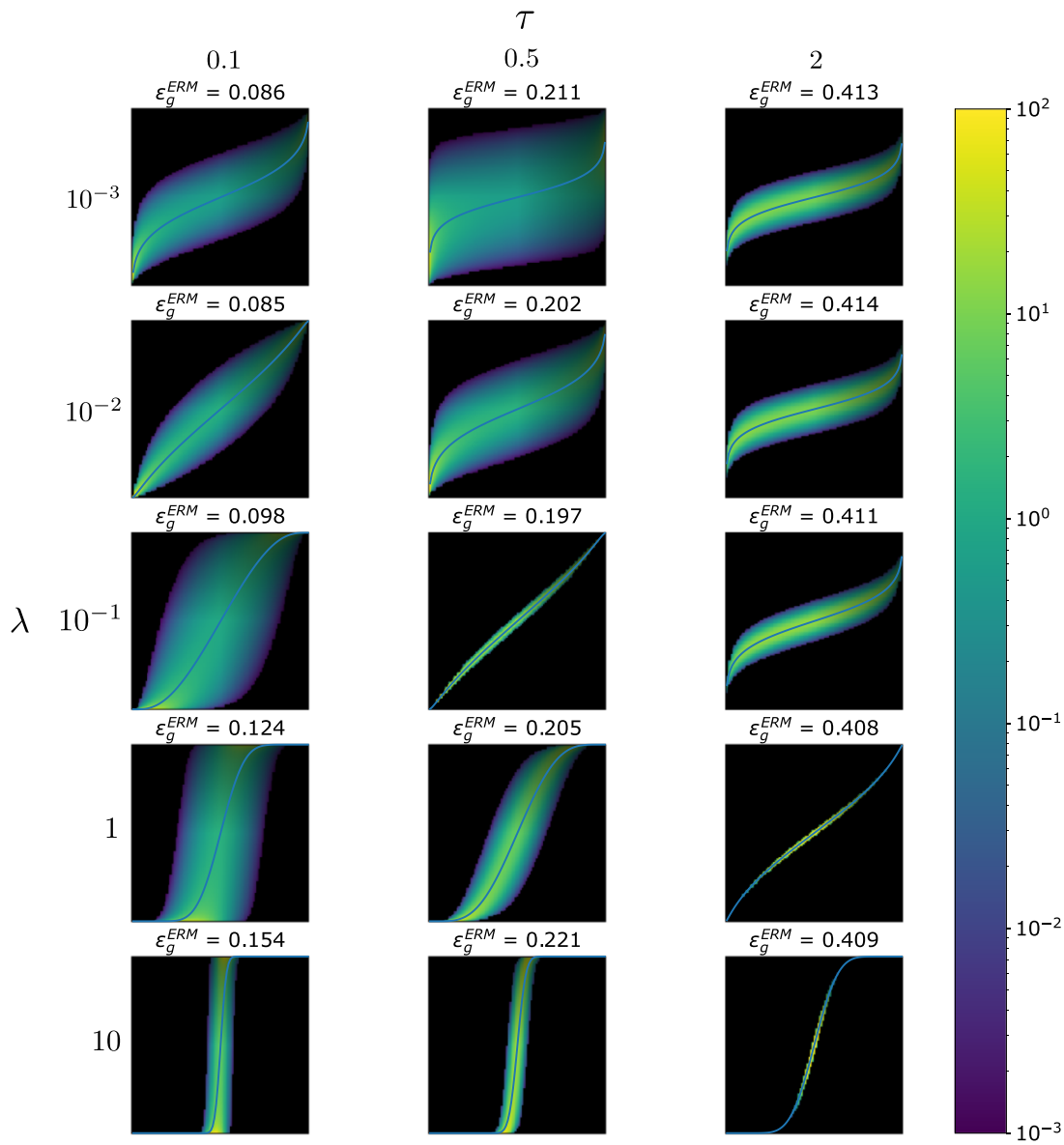


Figure 10. Joint density of $\hat{f}_{\text{erm}}$ (x-axis) and $\hat{f}_{\text{bo}}$ (y-axis) at $\alpha = 5$. The best possible test errors are respectively $\varepsilon_g^\star = 0, 0.148, 0.352$ for $\tau = 0, 0.5, 2$. For the Bayes estimator with $\alpha = 5$, the test errors are $\varepsilon_g^{\text{bo}} = 0.083, 0.198, 0.402$.

Note also that the expression of the calibration is now

$$\Delta_p = p - \int \mathrm{d}x \sigma(x) \mathcal{N}(x | m/q \times \sigma^{-1}(p), \rho - m^2/q). \tag{119}$$

D.1. Behavior at $\lambda = 0^+$

In [9], it has been shown that as the sampling ratio α goes to ∞, the unpenalized logistic classifier is calibrated when the data is generated by the *logit* model

$$\mathbb{P}(y_\star = 1) = \sigma(\mathbf{w}_\star \cdot \mathbf{x}). \tag{120}$$

In this section, we numerically recover the results from [9] i.e the unpenalized logistic estimator is calibrated asymptotically and the calibration decreases as $1/\alpha$. Figure 12 plots the calibration at $p = 0.75, 0.9$ and 0.99 for $\alpha \in [10, 10^4]$. One can observe a decay of Δ_p with a power law, which confirms that with logistic data, the unpenalized logistic classifier is asymptotically calibrated at all levels. Fitting a linear model on these curves gives slopes equal to $-0.99, -1.00, -1.04$ for $p = 0.75, 0.9, 0.99$ respectively, which numerically validates the $1/\alpha$ rate derived in [9].

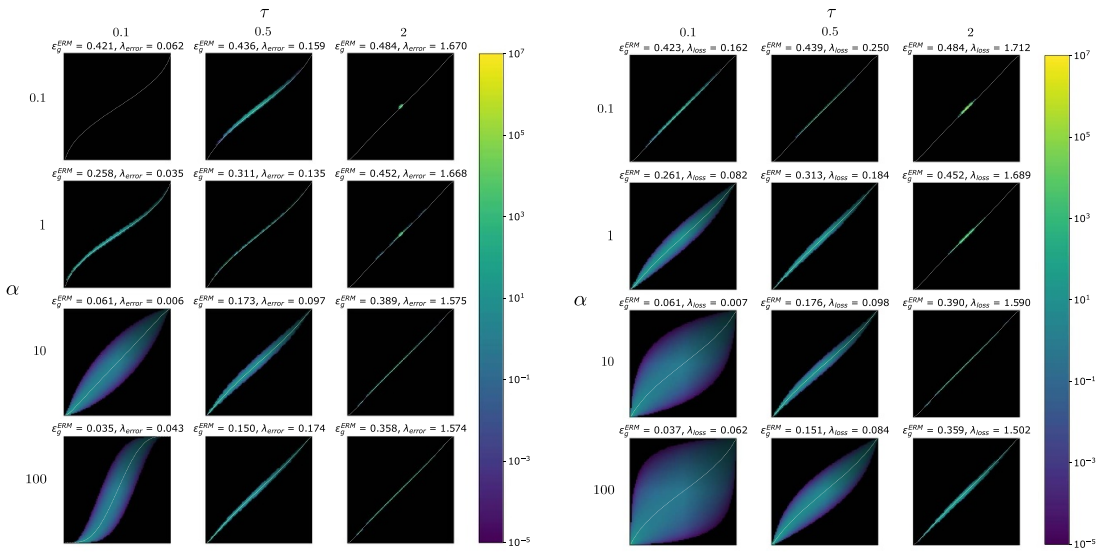


Figure 11. Left: Joint density $\rho_{\mathrm{erm,bo}}$, at $\lambda = \lambda_{\mathrm{error}}$. λ_{error} and the test error of ERM are indicated above the corresponding plot. Right:Joint density $\rho_{\mathrm{erm,bo}}$, at $\lambda = \lambda_{\mathrm{loss}}$. λ_{loss} and the test error of ERM are indicated above the corresponding plot.

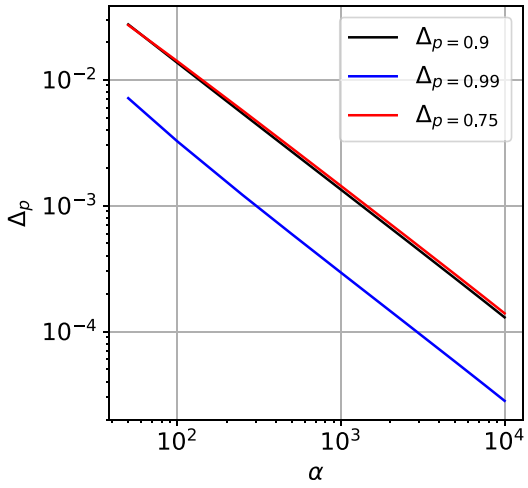


Figure 12. Calibration Δ_p at $p = 0.75, 0.9$ and 0.99 of logistic regression with the logit model as a function of α. The plots are given in log-log scale. On this scale, the curves have respective slopes $-0.99, -1.01, -1.04$.

We compare here to the calibration with probit data, at $\tau = 0.5$. In particular, we exhibit that the logistic classifier cannot be calibrated at all levels p. Indeed, as $\alpha \to \infty$, it can be noted that $\cos(\hat{\mathbf{w}}_{\mathrm{erm}}, \boldsymbol{w}_\star) = m^2/q \to_\infty 1$. Moreover, we observe that $m/q = m^2/q \times 1/m \to_\infty m_\infty := \lim m$. Using the expression for calibration from theorem 3.3, we get that for $p > 1/2$,

$$\Delta_p \to_\infty p - \sigma_\star\left(\frac{\sigma^{-1}(p)}{\tau \times m_\infty}\right). \tag{121}$$

And deduce that

$$\Delta_p = 0 \Leftrightarrow \frac{\sigma_\star^{-1}(p)}{\sigma^{-1}(p)} = \frac{1}{\tau \times m_\infty}. \tag{122}$$

Noting $r(p) := \frac{\sigma_\star^{-1}(p)}{\sigma^{-1}(p)}$, we get the condition

$$p = r^{-1}\left(\frac{1}{\tau \times m_\infty}\right). \tag{123}$$

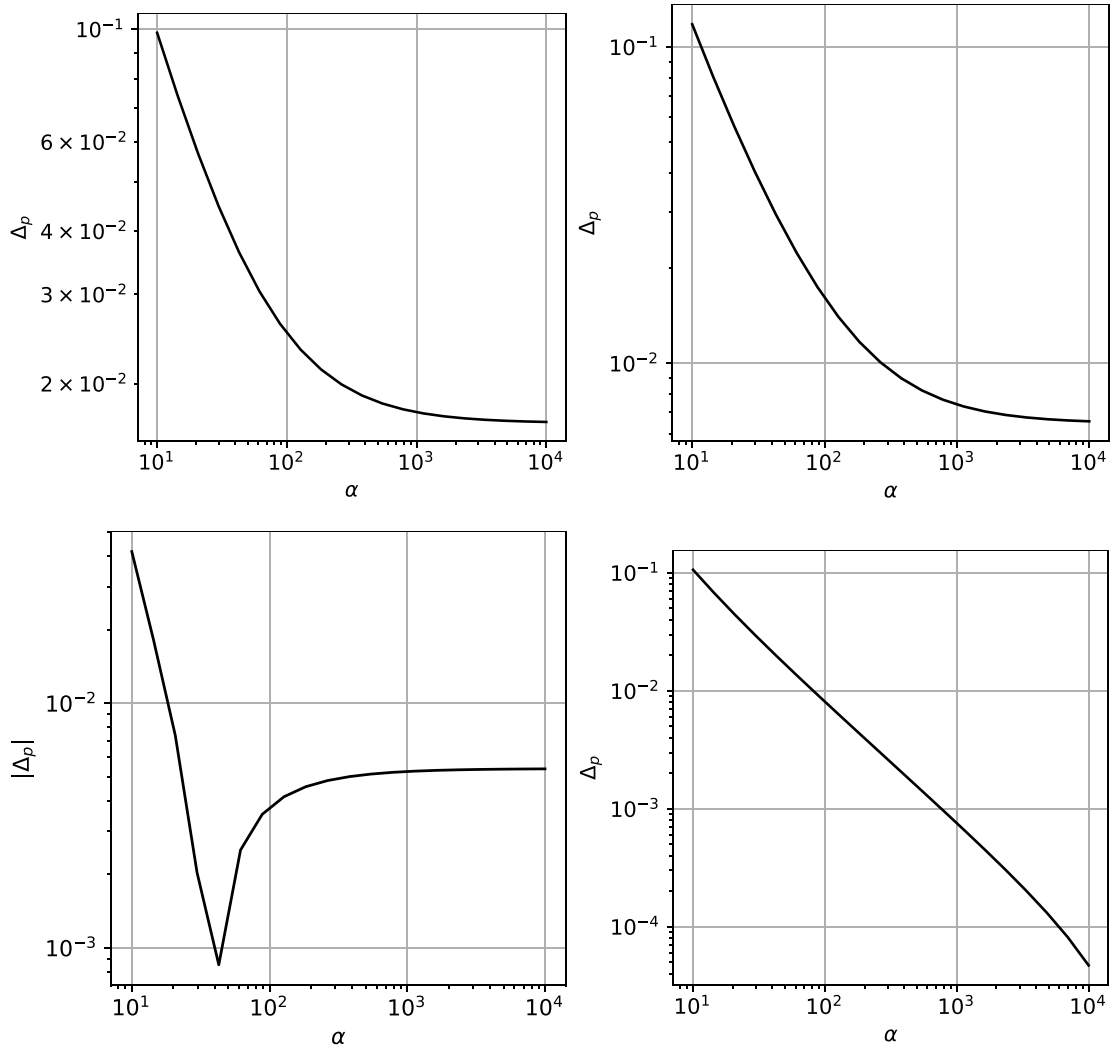


Figure 13. Calibration for logistic regression with $\lambda = 0^+, \tau = 0.5$, at four different levels: $p = 0.75$ (Top left), $p = 0.9$ (Top right), $p = 0.99$ (Bottom left) and $p = p_0 \simeq 0.937$ (Bottom right). For $p = 0.99$, Δ_p becomes negative around $\alpha \simeq 50$ so the absolute value of Δ_p has been plotted instead. C seems to converge to zero for $p = p_0$ only.

With $\tau = 0.5$, we numerically get $m_\infty \simeq 3.53 \Rightarrow \tau \times m_\infty \simeq 1.76$ The level p_0, defined as the only $p > 1/2$ such that $\Delta_p = 0$, is thus

$$p_0 = r^{-1}\left(\frac{1}{\tau \times m_\infty}\right) \simeq r^{-1}(0.57) \simeq 0.937. \tag{124}$$

For $1/2 < p < p_0$ (respectively $1 > p > p_0$), $\Delta_p > 0$ (respectively $\Delta_p < 0$). This can be observed in figure 13 where we have plotted Δ_p for several levels. For $p \neq p_0$, the calibration seems to converge to a finite value. On the other hand, at $p = p_0$, Δ_p converges to 0 as a power-law.

D.2. Behavior a $\lambda = 1$, λ_{error} and λ_{loss}

In this section, we adapt the theoretical results of figure 5 to the logit data model: we compute λ_{error} and λ_{loss} and plot their respective test errors and calibration. Note the definition of the test error and loss in this setting:

$$\begin{cases} \varepsilon_g = \sum_y \mathbb{E}_{\xi\sim\mathcal{N}(0,1)}\left[\mathcal{Z}_0(y, m/\sqrt{q}\xi, 1-m^2/q)\delta(\text{sign}(\xi) = y)\right] \\ \mathcal{L}_g = -\sum_y \mathbb{E}_{\xi\sim\mathcal{N}(0,1)}\left[\mathcal{Z}_0(y, m/\sqrt{q}\xi, 1-m^2/q)\log\sigma(y \times \sqrt{q}\xi)\right] \end{cases}. \tag{125}$$

Moreover, with the logit data model, the empirical risk at $\lambda = 1$ now has a Bayesian interpretation. The risk corresponds to the logarithm of the posterior distribution on $\boldsymbol{w}$, up to a normalization constant, because $\boldsymbol{w}_\star$ is sampled from a Gaussian with identity covariance. At $\lambda = 1$, the empirical risk minimizer $\hat{\boldsymbol{w}}_{\text{erm}}$ is the

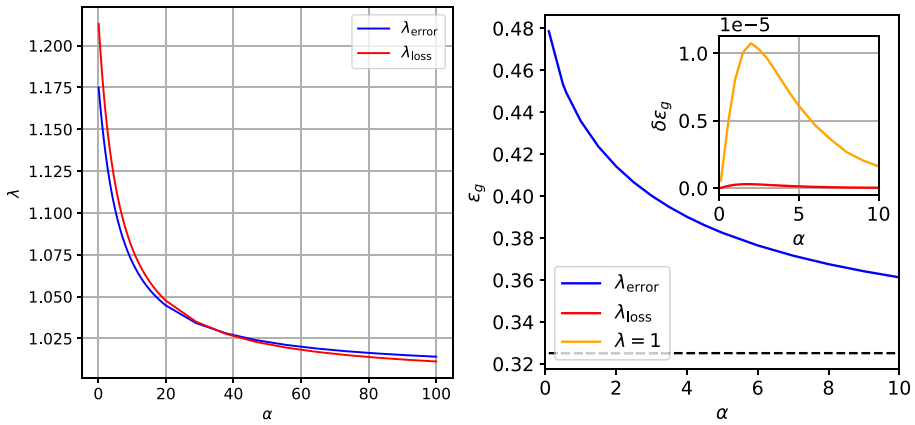


Figure 14. Left: Values of λ_{error} and λ_{loss} as a function of α for the logistic data model. Center: Values of the test error ε_g for λ_{error} (blue curve) . The inset plots the difference of test error $\delta\varepsilon_{g,\text{loss}} := \varepsilon_g(\lambda_{\text{loss}}) - \varepsilon_g(\lambda_{\text{error}})$ (red curve) and $\delta\varepsilon_{g,1} := \varepsilon_g(\lambda = 1) - \varepsilon_g(\lambda_{\text{error}})$ (orange curve). Right: Calibration at $p = 0.75$ of logistic regression on logistic data, for $\lambda = 1$, λ_{error} and λ_{loss}. The curves are given by running state evolution.

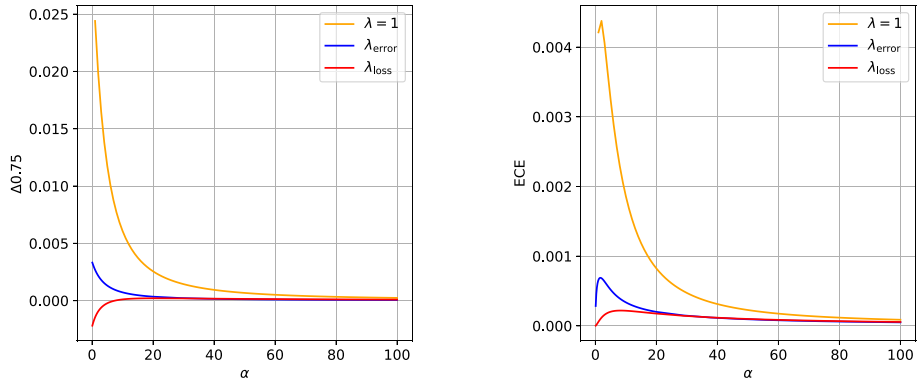


Figure 15. Left: Calibration at $p = 0.75$ of logistic regression on logistic data, for $\lambda = 1$, λ_{error} and λ_{loss}. The curves are given by running state evolution. Right: Expected Calibration Error (ECE) for $\lambda = 1$, λ_{error}, λ_{loss}. The lower ECE, the better.

maximum *a posteriori* (MAP). In this section, we compare the performance of logistic regression with the two different optimal regularizations and with $\lambda = 1$.

The left panel of figure 14 shows the value of λ_{error} and λ_{loss}. As with the probit model, $\lambda_{\text{loss}} > \lambda_{\text{error}}$. Note also that both optimal values are bigger than 1 for this range of α. The right panel shows their respective test error ε_g. As with the probit model, λ_{error} has a lower error than λ_{loss}. Not surprisingly, $\lambda = 1$ has worse test error than both optimal λ. The left panel of figure 15 shows the calibration with the three different regularizations at $p = 0.75$. We observe that $\lambda = 1$ yields an overconfident estimator (consistent with the fact that λ_{error} and λ_{loss} are both bigger than 1), and as before, λ_{loss} is less confident than λ_{error}. Remark that an underconfident estimator is not necessarily better than an overconfident one, and the calibration Δ_p is only a measure on one level p. To compare the different estimators more fairly, we can thus use a metric called Expected Calibration Error (ECE) defined as

$$\text{ECE} := \mathbb{E}_{\hat{f}(\mathbf{x})}\left(|\Delta_{\hat{f}(\mathbf{x})}|\right) = \int \mathrm{d}p |\Delta_p| \frac{\mathcal{N}(\sigma^{-1}(p)|0, q_{\text{erm}})}{p(1-p)}. \tag{126}$$

The ECE measures the average of $|\Delta_p|$ at all levels p weighted by the probability that $\hat{f}(\mathbf{x}) = p$. In other words, at a given level p, if $\mathbb{P}(\hat{f}(\mathbf{x}) = p) = 0$, the ECE of the estimator will not be affected by the calibration of the estimator at p. The right panel of figure 15 plots the ECE as a function of α for $\lambda = 1$, λ_{error} and λ_{loss}. We again observe that λ_{loss} has a lower ECE than λ_{error}, which confirms that optimizing λ for the test loss yields a more calibrated estimator. Moreover, $\lambda = 1$ yields an estimator with the worst ECE, which is coherent with the left panel: at $p = 0.75$, the absolute value of its calibration is higher than λ_{error} and λ_{loss}. Our numerical results show that even if we know the prior distribution on the posterior and the likelihood, using only a point estimate for the parameter (here the MAP) yields an overconfident estimator.

ORCID iD

Lucas Clarté https://orcid.org/0000-0003-0603-508X

References

[1] Abdar M *et al* 2021 A review of uncertainty quantification in deep learning: techniques, applications and challenges *Inf. Fusion* **76** 243–97
[2] Adlam B, Snoek J and Smith S L 2020 Cold posteriors and aleatoric uncertainty (arXiv:2008.00029)
[3] Aitchison L 2021 A statistical theory of cold posteriors in deep neural networks *Int. Conf. on Learning Representations* (available at: https://arxiv.org/abs/2008.05912)
[4] Angelopoulos A N, Bates S, Candès E J, Jordan M I and Lei L 2021 Learn then test: calibrating predictive algorithms to achieve risk control *CoRR* (arXiv:2110.01052)
[5] Aubin B, Krzakala F, Yue L and Zdeborová L 2020 Generalization error in high-dimensional perceptrons: approaching bayes error with convex optimization *Advances in Neural Information Processing Systems* pp 12199–210 (available at: https://proceedings.neurips.cc/paper/2020/file/8f4576ad85410442a74ee3a7683757b3-Paper.pdf)
[6] Aubin B, Loureiro B, Baker A, Krzakala F and Zdeborová L 2020 Exact asymptotics for phase retrieval and compressed sensing with random generative priors *Proc. 1st Mathematical and Scientific Machine Learning Conf.* (*Proc. Machine Learning Research* vol 107), ed J Lu and R Ward (PMLR) pp 55–73
[7] Aubin B, Maillard A, Barbier J, Krzakala F, Macris N and Zdeborová L 2019 The committee machine: computational to statistical gaps in learning a two-layers neural network *J. Stat. Mech.* **2019**
[8] Aubin B, Loureiro B, Maillard A, Krzakala F and Zdeborová L 2021 The spiked matrix model with generative priors *IEEE Trans. Inf. Theory* **2**
[9] Bai Y, Mei S and Xiong C 2021 Don't just blame over-parametrization for over-confidence: theoretical analysis of calibration in binary classification *ICML 2021* (available at: https://proceedings.mlr.press/v139/bai21c.html)
[10] Bai Y, Mei S, Wang H and Xiong C (2021) Understanding the under-coverage bias in uncertainty estimation NeurIPS *2021* (available at: https://proceedings.neurips.cc/paper_files/paper/2021/file/9854d7afce413aa13cd0a1d39d0bcec5-Paper.pdf)
[11] Barbier J, Krzakala F, Macris N, Miolane Leo and Zdeborová L 2019 Optimal errors and phase transitions in high-dimensional generalized linear models *Proc. Natl Acad. Sci.* **116** 5451–60
[12] Barbier J and Macris N 2019 The adaptive interpolation method: a simple scheme to prove replica formulas in bayesian inference *Probab. Theory Relat. Fields* **174** 1133–85
[13] Bayati M and Montanari A 2010 The lasso risk for gaussian matrices *IEEE Trans. Inf. Theory* **58** 1997–2017
[14] Bellec P and Kuchibhotla A 2019 First order expansion of convex regularized estimators *NeurIPS 2019* vol 32, ed H Wallach, H Larochelle, A Beygelzimer, F d' Alché-Buc, E Fox and R Garnett (Curran Associates, Inc.) (available at: https://proceedings.neurips.cc/paper_files/paper/2019/hash/0609154fa35b3194026346c9cac2a248-Abstract.html)
[15] Bruce A D and Saad D 1994 Statistical mechanics of hypothesis evaluation *J. Phys. A: Math. Gen.* **27** 3355–63
[16] Candes E J and Sur P 2020 The phase transition for the existence of the maximum likelihood estimate in high-dimensional logistic regression *Ann. Stat.* **48** 27–42
[17] Cover Thomas M and Thomas Joy A 1991 *Elements of Information Theory* vol 3 (New York: Wiley)
[18] Daxberger E, Kristiadi* A, Immer* A, Eschenhagen* R, Bauer M and Hennig P 2021 Laplace redux — effortless bayesian deep learning *Advances in Neural Information Processing Systems 34 (NeurIPS 2021)* (Curran Associates, Inc.) *equal contribution, pp 20089–103
[19] Deng Z, Kammoun A and Thrampoulidis C 2020 A model of double descent for high-dimensional logistic regression *ICASSP 2020 – 2020 IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*
[20] Dhifallah O and Lu Y M 2020 A precise performance analysis of learning with random features (arXiv:2008.11904 [cs.IT])
[21] Gal Y and Ghahramani Z 2016 Dropout as a bayesian approximation: representing model uncertainty in deep learning *ICML 2016* ed M F Balcan and K Q Weinberger (available at: http://proceedings.mlr.press/v48/gal16.pdf)
[22] Gerace F, Loureiro B, Krzakala F, Mézard M and Zdeborová L 2020 Generalisation error in learning with random features and the hidden manifold model *Int. Conf. on Machine Learning* (PMLR)
[23] Gerbelot C, Abbara A and Krzakala F 2020 Asymptotic errors for high-dimensional convex penalized linear regression beyond gaussian matrices *Proc. 33rd Conf. on Learning Theory* (*Proc. Machine Learning Research*) (PMLR)
[24] Gerbelot C and Berthier Rel 2021 Graph-based approximate message passing iterations (arXiv:2109.11905 [cs.IT])
[25] Goldt S, Mézard M, Krzakala F and Zdeborová L 2020 Modeling the influence of data structure on learning in neural networks: the hidden manifold model *Phys. Rev.* X **10** 041044
[26] Goldt S, Loureiro B, Reeves G, Krzakala F, Mézard M and Zdeborová L 2021 The gaussian equivalence of generative models for learning with shallow neural networks *Conf. on Mathematical and Scientific Machine Learning 2021* (arXiv:https://msml21.github.io/papers/id50.pdf) [stat.ML])
[27] Guo C, Pleiss G, Sun Y and Weinberger K Q 2017 On calibration of modern neural networks ICML *2017 Proc. 34th Int. Conf. on Machine Learning* (available at: https://proceedings.mlr.press/v70/guo17a)
[28] Gupta C, Podkopaev A and Ramdas A 2020 Distribution-free binary classification: prediction sets, confidence intervals and calibration *NeurIPS 2020 Proc. 34th Int. Conf. on Neural Information Processing Systems* (Curran Associates Inc.) (available at: https://proceedings.neurips.cc/paper/2020/hash/26d88423fc6da243ffddf161ca712757-Abstract.html)
[29] Hensman J, Fusi Nò and Lawrence N D 2013 Gaussian processes for big data UAI *Proc. 29th Conf. on Uncertainty in Artificial Intelligence* (*2013*) (AUAI Press) (available at: http://www.auai.org/uai2013/prints/papers/244.pdf)
[30] Iba Y 1999 The nishimori line and bayesian statistics *J. Phys. A: Math. Gen.* **32** 3875–88
[31] Javanmard A and Montanari A 2013 State evolution for general approximate message passing algorithms, with applications to spatial coupling *Inf. Inference: J. IMA* **2** 115–44
[32] Kapoor S, Maddox W J, Izmailov P and Gordon Wilson A 2022 On uncertainty, tempering and data augmentation in bayesian classification (available at: https://arxiv.org/abs/2203.16481)
[33] Kendall A and Gal Y 2017 What uncertainties do we need in bayesian deep learning for computer vision? *NeurIPS 2017 Proc. 31st Int. Conf. on Neural Information Processing Systems* (Curran Associates Inc) p NIS'17 (available at: https://papers.nips.cc/paper_files/paper/2017/hash/2650d6089a6d640c5e85b2b88265dc2b-Abstract.html)

[34] Kristiadi A, Hein M and Hennig P 2020 Being bayesian, even just a bit, fixes overconfidence in relu networks *ICML 2020* (available at: http://proceedings.mlr.press/v119/kristiadi20a/kristiadi20a.pdf)
[35] Lakshminarayanan B, Pritzel A and Blundell C 2017 Simple and scalable predictive uncertainty estimation using deep ensembles NeurIPS 2017 (Curran Associates, Inc.) (available at: https://proceedings.neurips.cc/paper/2017/file/9ef2ed4b7fd2c810847ffa5fa85bce38-Paper.pdf)
[36] Liang T and Sur P 2020 A precise high-dimensional asymptotic theory for boosting and minimum-ℓ_1-norm interpolated classifiers (https://doi.org/10.2139/ssrn.3714013)
[37] Liu J Z, Lin Z, Padhy S, Tran D, Bedrax-Weiss T and Lakshminarayanan B 2020 Simple and principled uncertainty estimation with deterministic deep learning via distance awareness *NeurIPS 2020* (available at: https://proceedings.neurips.cc//paper_files/paper/2020/hash/543e83748234f7cbab21aa0ade66565f-Abstract.html)
[38] Loureiro B, Gerbelot Cedric, Cui H, Goldt S, Krzakala F, Mézard M and Zdeborová L 2022 Learning curves of generic features maps for realistic datasets with a teacher-student model *J. Stat. Mech.* 114001
[39] Loureiro B, Sicuro G, Gerbelot Cedric, Pacco A, Krzakala F and Zdeborová L 2021 Learning gaussian mixtures with generalised linear models: precise asymptotics in high-dimensions *NeurIPS 2021* (available at: https://proceedings.neurips.cc/paper_files/paper/2021/file/543e83748234f7cbab21aa0ade66565f-Paper.pdf)
[40] Mackay D J C 1995 Probable networks and plausible predictions – a review of practical bayesian methods for supervised neural networks *Netw., Comput. Neural Syst.* **6** 469–505
[41] MacKay D J C 1992 Bayesian Interpolation *Neural Comput.* **4** 415–47
[42] Maddox W J, Izmailov P, Garipov T, Vetrov D P and Gordon Wilson A 2019 A simple baseline for bayesian uncertainty in deep learning *NeurIPS 2019 Advances in Neural Information Processing Systems* vol 32, ed H Wallach, H Larochelle, A Beygelzimer, F d' Alché-Buc, E Fox and R Garnett (Curran Associates, Inc.) (available at: https://proceedings.neurips.cc/paper/2019/hash/118921efba23fc329e6560b27861f0c2-Abstract.html)
[43] Mai X, Liao Z and Couillet R 2019 A large scale analysis of logistic regression: asymptotic performance and new insights *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)* (https://doi.org/10.1109/ICASSP.2019.8683376)
[44] Malinin A, Mlodozeniec B and Gales M 2020 Ensemble distribution distillation *ICLR 2020 Int. Conf. on Learning Representations* (available at: https://iclr.cc/virtual_2020/poster_BygSP6Vtvr.html)
[45] Marion G and Saad D 1995 A statistical mechanical analysis of a bayesian inference scheme for an unrealizable rule *J. Phys. A: Math. Gen.* **28** 2159–71
[46] Marion G and Saad D 1994 Hyperparameters evidence and generalisation for an unrealisable rule NIPS *1994 Advances in Neural Information Processing Systems* vol 7, ed G Tesauro, D Touretzky and T Leen (Cambridge, MA: MIT Press) (available at: https://proceedings.neurips.cc/paper_files/paper/1994/hash/e6cb2a3c14431b55aa50c06529eaa21b-Abstract.html)
[47] Mattei P-A 2019 A parsimonious tour of bayesian model uncertainty (available at: https://arxiv.org/abs/1902.05539)
[48] Mei S and Montanari A 2021 The generalization error of random features regression: precise asymptotics and the double descent curve *Commun. Pure Appl. Math.* **75** 667–766
[49] Mezard M and Montanari A 2009 *Information, Physics and Computation* (Oxford: Oxford University Press)
[50] Mézard M, Parisi G and Angel Virasoro M 1987 *Spin Glass Theory and Beyond: An Introduction to the Replica Method and Its Applications* vol 9 (Singapore: World Scientific Publishing Company)
[51] Mignacco F, Krzakala F, Yue L, Urbani P and Zdeborova L 2020 The role of regularization in classification of high-dimensional noisy Gaussian mixture *Proc. 37th Int. Conf. on Machine Learning* (*Proc. Machine Learning Research*) (PMLR)
[52] Montanari A, Ruan F, Sohn Y, and Yan J 2020 The generalization error of max-margin linear classifiers: high-dimensional asymptotics in the overparametrized regime (arXiv:1911.01544 [math.ST])
[53] Mukhoti J, Kulharia V, Sanyal A, Golodetz S, Torr P H S and Dokania P K 2020 Calibrating deep neural networks using focal loss *Advances in Neural Information Processing Systems*
[54] Platt J 2000 Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods *Adv. Large Margin Classif.* **10** 61–74
[55] Posch K, Steinbrener J and Pilz Jurgen 2019 Variational inference to measure model uncertainty in deep neural networks (arXiv:1902.10189 [stat.ML])
[56] Rangan S 2011 Generalized approximate message passing for estimation with random linear mixing *2011 IEEE Int. Symp. on Information Theory Proc.* (IEEE)
[57] Ritter H, Botev A and Barber D 2018 A scalable laplace approximation for neural networks *Int. Conf. on Learning Representations*
[58] Seddik E A M, Louart C, Tamaazousti M and Couillet R 2020 Random matrix theory proves that deep learning representations of gan-data behave as gaussian mixtures *Int. Conf. on Machine Learning* (PMLR) pp 8573–82
[59] Seeger M 2004 Gaussian processes for machine learning *Int. J. Neural Syst.* **14** 69–106
[60] Shafer G and Vovk V 2008 A tutorial on conformal prediction *J. Mach. Learn. Res.* **9** 371–421
[61] Sur P and Emmanuel J Cès 2019 A modern maximum-likelihood theory for high-dimensional logistic regression *Proc. Natl Acad. Sci.* **116** 14516–25
[62] Taheri H, Pedarsani R and Thrampoulidis C 2020 Sharp asymptotics and optimal performance for inference in binary models *Proc. 33rd Int. Conf. on Artificial Intelligence and Statistics* (*Proc. Machine Learning Research*) (PMLR)
[63] Thrampoulidis C, Abbasi E and Hassibi B 2018 Precise error analysis of regularized m-estimators in high dimensions *IEEE Trans. Inf. Theory* **64** 5592–628
[64] Thulasidasan S, Chennupati G, Bilmes J, Bhattacharya T and Michalak S 2019 On mixup training: improved calibration and predictive uncertainty for deep neural networks *Advances in Neural Information Processing Systems*
[65] Wilson A G 2020 The case for bayesian deep learning (arXiv:2001.10995)
[66] Zadrozny B and Elkan C 2001 Obtaining calibrated probability estimates from decision trees and naive bayesian classifiers *ICML* **1** 606–16
[67] Zadrozny B and Elkan C 2002 Transforming classifier scores into accurate multiclass probability estimates *Proc. ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining* (https://doi.org/10.1145/775047.775151)
[68] Zdeborová L and Krzakala F 2016 Statistical physics of inference: thresholds and algorithms *Adv. Phys.* **65** 453–552
[69] Clarté L 2022 SPOC-group/high-dimensional-uncertainty repository (available at: https://github.com/SPOC-group/high-dimensional-uncertainty)