A study of uncertainty quantification in overparametrized high-dimensional models

Lucas Clarté, Bruno Loureiro, Florent Krzakala, Lenka Zdeborová

A study of uncertainty quantification in overparametrized high-dimensional models

Lucas Clarté[1], Bruno Loureiro[2,3],
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

Abstract

Uncertainty quantification is a central challenge in reliable and trustworthy machine learning. Naive measures such as last-layer scores are well-known to yield overconfident estimates in the context of overparametrized neural networks. Several methods, ranging from temperature scaling to different Bayesian treatments of neural networks, have been proposed to mitigate overconfidence, most often supported by the numerical observation that they yield better calibrated uncertainty measures. In this work, we provide a sharp comparison between popular uncertainty measures for binary classification in a mathematically tractable model for overparametrized neural networks: the random features model. We discuss a trade-off between classification accuracy and calibration, unveiling a double descent like behavior in the calibration curve of optimally regularized estimators as a function of overparametrization. This is in contrast with the empirical Bayes method, which we show to be well calibrated in our setting despite the higher generalization error and overparametrization.

1 Introduction

Uncertainty estimation is the cornerstone of reliable data processing. A large body of literature in classical statistical theory is dedicated to providing solid mathematical guarantees on a model's uncertainty, such as confidence scores for classification and confidence intervals for regression [Wasserman, 2013]. Yet, when it comes to modern machine learning methods such as deep neural networks our mathematical understanding of the uncertainty associated with prediction falls short. A key aspect in current machine learning practice is that, in contrast to classical wisdom, models often operate in a regime where the complexity of the hypothesis class (e.g. as measured by the number of parameters in the model) is comparable or larger than the quantity of data available for training. This modern, overparametrized regime defies the common intuition rooted on classical statistics, therefore posing interesting challenges to their mathematical treatments. For example, deep neural networks are able to achieve optimal generalization performance even when the training data are perfectly interpolated [Geman et al., 1992, Geiger et al., 2019, Nakkiran et al., 2020], a behaviour at odds with the bias-variance intuition. This *benign overfitting* property was recently shown to be common among overparametrized convex methods, such as linear regression [Bartlett et al., 2020, Hastie et al., 2022], random features regression [Mei and Montanari, 2022] and classification [Gerace et al., 2020].

While much of the theoretical effort has focused on the generalization properties of point estimates from overparametrized models, less is understood about their confidence. Indeed, a popular method to estimate uncertainty in neural networks consists of interpreting the last layer pre-activations as class probabilities. Numerical experiments suggest that deep neural networks tend to suffer from

overconfidence with respect this notion [Guo et al., 2017], a problem which has motivated many empirical calibration methods in the literature [Hein et al., 2019, Kristiadi et al., 2020, Mukhoti et al., 2020, Liu et al., 2020]. Recently, it has been shown that actually overconfidence is a common problem in high-dimensional classification [Bai et al., 2021], although it can be considerably mitigated by properly regularising the risk [Clarté et al., 2022]. An alternative to the pre-activation scores consists in applying a Bayesian treatment to neural networks, for instance by averaging the last layer weights over the measure induced by the empirical risk. In some contexts, these techniques were shown to provide better calibrated uncertainty measures than pre-activation score. A priori, Bayesian techniques require sampling from a high-dimensional measure, and therefore can be computationally demanding [Alexos et al., 2022]. Despite the success and widespread use of these uncertainty measures, mathematical guarantees relating these notions to intrinsic uncertainty measures such as the true class probabilities or the best uncertainty estimation given the available data (i.e. the true posterior uncertainty given the features) are scarce. In this work, we provide a sharp mathematical comparison between these different uncertainty notions in the context of a simple, solvable model for binary classification on structured features - such as the ones given by the first layers of neural networks. To the best of our knowledge, our work is the first to provide a sharp asymptotic analysis of uncertainty in overparametrized high-dimensional models.

Our **main contributions** are:

• We provide an exact asymptotic description of the generalization error & the joint probability density for different classifiers of interest for the random features model, beyond the empirical risk characterization of [Gerace et al., 2020, Dhifallah and Lu, 2020]. It allows us to evaluate uncertainty measures, such as the calibration and the conditional variance of prediction, and to study the interplay between uncertainty & overparametrization.

• We identify a fundamental trade-off between performance and prediction confidence as a function of the number of parameters in the model. We compare cross-validation schemes with Bayesian model selection, and show that targeting high accuracy can be at odds with achieving a calibrated classifier. In particular, we show that the calibration of the optimally regularized empirical risk classifier peaks around the interpolation threshold, despite the monotonic generalization error. We show that a popular calibration technique known as temperature scaling [Guo et al., 2017] mitigates this peak, and achieves the best combination of a good test error and a good calibration.

• Finally, we discuss a popular uncertainty measure motivated by Bayesian methods: the Laplace approximation [MacKay, 1992, Ritter et al., 2018]. The Laplace approximation is known to produce a consistently less confident uncertainty measure, and in our work we quantify how much. In particular, we show it often provides an underconfident estimate. Our analysis requires an asymptotic characterization of the logistic loss Hessian at the minimum - a result we believe to be of independent interest.

Related work – Uncertainty quantification in deep learning is an active and rapidly evolving field, with many coexisting metrics and methods in the literature, see e.g. [Abdar et al., 2021, Gawlikowski et al., 2022] for two recent reviews. [Nguyen et al., 2015, Guo et al., 2017] empirically observed that different from "small" networks [Niculescu-Mizil and Caruana, 2005], modern deep neural networks tend to give overconfident predictions. [Guo et al., 2017] proposed *temperature scaling*, a simple post-processing variant of Platt scaling [Platt, 2000] consisting of rescaling & cross-validating the norm of the last-layer weights, and showed it can effectively calibrate them. Alternatively, [Kristiadi et al., 2020] has argued that a Bayesian treatment of the last layer of deep networks fixes overconfidence. Bayesian methods typically involve sampling from a high-dimensional posterior [Mattei, 2019], and different methods have been proposed to compute them efficiently [Graves, 2011, Gal and Ghahramani, 2016, Lakshminarayanan et al., 2017, Maddox et al., 2019]. Of particular interest to our work is the Laplace approximation introduced in [MacKay, 1992] for Gaussian process classification and adapted to Bayesian deep learning in [Ritter et al., 2018, Kristiadi et al., 2020, Daxberger et al., 2021]. An asymptotic discussion of evidence maximization in Bayesian ridge regression appeared in [Marion and Saad, 1994, Bruce and Saad, 1994, Marion and Saad, 1995]. [Bai et al., 2021] has shown that the logit model is overconfident in high-dimensions, and [Clarté et al., 2022] discussed how to mitigate it by properly

regularizing. An exact asymptotic characterization of the empirical risk minimizer for random features model has been derived and discussed in [Mei and Montanari, 2022, Gerace et al., 2020, Goldt et al., 2022, Hu and Lu, 2020, Dhifallah and Lu, 2020, Loureiro et al., 2021a]. Particularly relevant to our technical results is the recent progress in approximate message-passing schemes for structured matrices [Gerbelot and Berthier, 2021, Loureiro et al., 2022]. Finally, exact asymptotics for Bayes-optimal estimation has been discussed in the context of generalized linear models in [Barbier et al., 2019, Gabrié et al., 2018].

Notation – We denote vectors with bold letters, and matrices with capital letters. For $n \in \mathbb{N}$, we let $[n] := \{1, \cdots, n\}$. $\mathcal{N}(\boldsymbol{\mu}, \Sigma)$ denotes the Gaussian density, $\sigma(t) := (1 + e^{-t})^{-1}$ denotes the sigmoid function. We define

$$\sigma_v(x) := \int \sigma(z)\mathcal{N}(z|x, v)\mathrm{d}z \tag{1}$$

the averaged sigmoid with a Gaussian noise of variance v.

2 Setting

2.1 Probabilistic classifiers and uncertainty

Consider a supervised binary classification task given by n independent samples $\mathcal{D} = (\boldsymbol{x}^\mu, y^\mu)_{\mu\in[n]} \in \mathcal{X} \times \{-1, +1\}$ from a joint distribution ν, and denote by $f_\star(\boldsymbol{x}) = \nu(y = 1|\boldsymbol{x})$ the oracle class probability obtained by conditioning ν over an input. In this work we are interested in studying the uncertainty associated to probabilistic classifiers $\hat{f}(\boldsymbol{x}) = \mathbb{P}(y = 1|\boldsymbol{x}) \in [0, 1]$ obtained by fitting the data[1], and how they compare with the true class probability $f_\star$. A key motivation is the recent stream of works on uncertainty quantification for neural networks, and in particular the line of works proposing uncertainty measures based on classifiers defined by sampling over the last layer of neural networks [Brosse et al., 2020, Kristiadi et al., 2020]. To set notation, let $\boldsymbol{\varphi} : \mathcal{X} \to \mathbb{R}^p$ denote a *feature map*, for instance the features learned by the first layers of a trained neural network. We shall be interested in the following classifiers:

Empirical risk classifier – The empirical risk classifier is the one obtained by naively interpreting the scores in the last year as probability distributions. Mathematically, it is defined as $\hat{f}_{\mathrm{erm}}(\boldsymbol{x}) = \sigma(\hat{\boldsymbol{\theta}}_{\mathrm{erm}}^\top \varphi(\boldsymbol{x}))$, where $\sigma : \mathbb{R} \to (0, 1)$ is a non-linearity. For concreteness, we will focus on the popular case where $\sigma(z) = (1+e^{-z})^{-1}$ is the sigmoid function, and $\hat{\boldsymbol{\theta}}_{\mathrm{erm}} \in \mathbb{R}^p$ is the minimizer of the associated (regularized) logistic or cross-entropy risk:

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

This is also commonly referred to as the *logit classifier*.

Bayes-optimal classifier – Denoting the training features $\mathcal{D}_\varphi := \{(\boldsymbol{\varphi}(\boldsymbol{x}^\mu), y^\mu)\}_{\mu\in[n]}$, the optimal Bayesian classifier for the last layer is given by:

$$\hat{f}_{\mathrm{bo}}(\boldsymbol{x}) = \int \mathrm{d}\boldsymbol{\theta}\; p\left(y = 1|\boldsymbol{\theta}, \{\boldsymbol{\varphi}\left(\boldsymbol{x}^\mu\right)\}_{\mu\in[n]}\right) p\left(\boldsymbol{\theta}|\mathcal{D}_\varphi\right) \tag{3}$$

where $p\left(y = 1|\boldsymbol{\theta}, \{\boldsymbol{\varphi}(\boldsymbol{x}^\mu)\}_{\mu\in[n]}\right)$ is the likelihood over the labels and $p(\boldsymbol{\theta}|\mathcal{D}_\varphi)$ is the posterior distribution over the weights given the training features and labels. In practice, the Bayes-optimal classifier f_{bo} is not accessible to the statistician, since she doesn't have access to the distribution ν that has generated the data - and even if she had, sampling from the high-dimensional posterior distribution would be computationally cumbersome. However, as we will discuss in Sec. 3.1, for the data generative model considered here, the Bayes-optimal classifier can be asymptotically characterized, and its marginals can be computed by a polynomial-time message passing algorithm.

[1] In the following, we consistently denote with a hat classifiers which are a function of the training data.

Bayesian classifiers – Since the optimal Bayesian classifier is not accessible in practice, different classifiers inspired by Bayesian methods have been proposed in the literature. In this manuscript, we will consider two popular choices.

The first is the *empirical Bayes* classifier $\hat{f}_{\text{eb}}$ [Marion and Saad, 1994, Jospin et al., 2022]. In full generality, the empirical Bayes method consists of postulating a class of plausible likelihoods and priors and doing model selection from the training data via evidence maximization. In the context of Bayesian neural networks, the likelihood and priors are defined by the network architecture and regularization, which are normalized to define proper probability distributions. In our setting, the empirical Bayes classifier is explicitly given by:

$$\hat{f}_{\text{eb}}(\boldsymbol{x}) = \int_{\mathbb{R}^p} \mathrm{d}\boldsymbol{\theta}\ \sigma(\beta \boldsymbol{\theta}^\top \boldsymbol{\varphi}(\boldsymbol{x})) p_{\text{eb}}(\boldsymbol{\theta}|\mathcal{D}, \beta, \lambda), \tag{4}$$

$$p_{\text{eb}}(\boldsymbol{\theta}|\mathcal{D}, \beta, \lambda) = \frac{\prod_\mu \sigma(\beta y^\mu \boldsymbol{\theta}^\top \boldsymbol{\varphi}(\boldsymbol{x}^\mu)) \mathcal{N}(\boldsymbol{\theta}|\mathrm{I}_p/\beta\lambda)}{p(\mathcal{D}|\beta, \lambda)}$$

The normalisation constant $p(\mathcal{D}|\beta, \lambda)$ is known as the *marginal likelihood* or the *evidence*. In the empirical Bayes method the evidence is maximized in order to select the most likely hyperparameters (β, λ) explaining the training data [MacKay, 1996]. In our specific model, we note that the evidence is actually only a function of the ratio λ/β (this can be seen from the change of variables $\boldsymbol{\theta} \leftarrow \beta\boldsymbol{\theta}$). Therefore, without loss of generality we take $\beta = 1$ and optimize only over λ. It is important to stress that the postulated prior and likelihood in $\hat{f}_{\text{eb}}$ may not correspond to the ones that generated the data in general.

Note that, differently from the Bayes-optimal estimator, the empirical Bayes classifier can be a priori computed using only the training data. However, it can be computationally demanding to sample from the posterior distribution above, specially in large dimensions $p, n \gg 1$. To avoid this computational bottleneck, a common approximation consists of expanding the posterior around the $\hat{\boldsymbol{\theta}}_{\text{erm}}$ to second order, known as the *Laplace approximation* [Kristiadi et al., 2020, Ritter et al., 2018, Daxberger et al., 2021]:

$$\hat{f}_{\text{Lap}}(\boldsymbol{x}) = \int \mathrm{d}\boldsymbol{\theta}\ \sigma\left(\hat{\boldsymbol{\theta}}_{\text{erm}}^\top \boldsymbol{\varphi}(\boldsymbol{x})\right) \mathcal{N}(\boldsymbol{\theta}|\hat{\boldsymbol{\theta}}_{\text{erm}}, \mathcal{H}^{-1}) \tag{5}$$

where $\mathcal{H} := \nabla^2_{\boldsymbol{\theta}} \hat{\mathcal{R}}_n(\hat{\boldsymbol{\theta}}_{\text{erm}})$ is the Hessian of the empirical risk evaluated at the minimum. Therefore, in the Laplace approximation the posterior is effectively approximated by a Gaussian distribution centred at $\hat{\boldsymbol{\theta}}_{\text{erm}}$ and with covariance given by the inverse curvature around the minimum. The "sharper" the minimum, the lower the variance and the more confident the Laplace classifier is. Note that the generalization errors associated to the Laplace classifier coincide exactly with the empirical risk classifier. Finally, in the model considered here, the Laplace approximation $\hat{f}_{\text{Lap}}$ will always be less confident than the ERM estimator using $\hat{\boldsymbol{\theta}}_{\text{erm}}$. This is due to the concavity of the logit function σ on $[0, \infty)$.

Performance and uncertainty – Given a probabilistic classifier $\hat{f}$, the most common measure for the generalization performance is the *misclassification test error* (also known as *0/1 error*) :

$$\mathcal{E}_{\text{gen.}}(\hat{f}) = \mathbb{E}_{(\boldsymbol{x},y)\sim\nu} \mathbb{P}\left(\text{sign}(\hat{f}(\boldsymbol{x})) \neq y\right). \tag{6}$$

For $\hat{f}_{\text{erm}}$, another commonly used metric is the *test loss* :

$$\mathcal{L}_{\text{gen.}}(\hat{f}) = -\mathbb{E}_{(\boldsymbol{x},y)\sim\nu} \log(\sigma(y\hat{f}(\boldsymbol{x}))). \tag{7}$$

However, our key goal in this manuscript is to mathematically characterize the uncertainty associated to the prediction of the different classifiers above, and in particular how they correlate with the true class uncertainty as measured by $f_\star$. Mathematically, this can be measured by the following joint density:

$$\rho_{\star,t}(a, b) := \mathbb{E}_{\mathcal{D}} \mathbb{P}_{\boldsymbol{x}}\left(f_\star(\boldsymbol{x}) = a, \hat{f}_t(\boldsymbol{x}) = b\right) \tag{8}$$

where $(a, b) \in [0, 1]^2$ and $\hat{f}_t$, $t \in \{\text{bo}, \text{erm}, \text{Lap}, \text{eb}\}$ can be any of the classifiers defined above, and the expectation is taken both over the training data $\mathcal{D} = \{(\boldsymbol{x}^\mu, y^\mu)\}_{\mu \in [n]}$. In particular, this joint density gives access to different notions used in the literature to quantify uncertainty. For instance, a widely studied notion is the *calibration at level* $\ell \in [0, 1]$ of a classifier $\hat{f}$:

$$\Delta_\ell(\hat{f}) := \ell - \mathbb{E}_{\boldsymbol{x},\mathcal{D}} \left[f_\star(\boldsymbol{x}) | \hat{f}(\boldsymbol{x}) = \ell \right] . \tag{9}$$

A related metric is the *Expected Calibration Error* (ECE):

$$\text{ECE}(\hat{f}) := \mathbb{E}_{\mathbf{x}} \left[|\Delta_{\hat{f}(\mathbf{x})}| \right] . \tag{10}$$

2.2 The random features model

Following our aim to investigate the interplay between overparametrization and uncertainty, we will focus on one of the simplest settings of feature maps defined by two-layer neural networks $\boldsymbol{\varphi} : \boldsymbol{x} \in \mathcal{X} \subset \mathbb{R}^p \mapsto \phi(F\boldsymbol{x})/\sqrt{p}$ with weights $F \in \mathbb{R}^{p \times d}$ and component-wise activation ϕ. We will consider *random features* [Rahimi and Recht, 2007], where the first layer weights $F \in \mathbb{R}^{p \times d}$ are fixed at initialization, typically taken to be i.i.d. standard Gaussian. Random features have been widely studied as a convex proxy for investigating the impact of overparametrization in generalization, since they were shown to display the characteristic non-monotonic *double descent* behaviour of the generalization error [Belkin et al., 2019, Spigler et al., 2019], with optimal generalization achieved beyond interpolation of the data [Mei and Montanari, 2022, Gerace et al., 2020, D'Ascoli et al., 2020], also known as *benign overfitting* [Bartlett et al., 2020].

We will assume Gaussian input data $\boldsymbol{x}^\mu \sim \mathcal{N}(\mathbf{0}, {}^1\!/_d \mathrm{I}_d)$ with labels drawn from a logit model:

$$f_\star(\boldsymbol{x}) = \int_{\mathbb{R}} \sigma \left(\boldsymbol{\theta}_\star^\top \boldsymbol{x} + \tau_0 z \right) \mathcal{N}(z|0, 1) \mathrm{d}z \tag{11}$$

with random weights $\boldsymbol{\theta}_\star \sim \mathcal{N}(\mathbf{0}, \mathrm{I}_d)$ and $\tau_0 \geq 0$ defines a tunable label noise level. This completely specifies the data distribution ν. In the following, we will be interested in the *proportional high-dimensional limit* defined by $n, p, d \to \infty$ with fixed ratios $\alpha := {}^n\!/_p$ and $\gamma := {}^p\!/_d$.

An asymptotic characterization of the generalization and training errors of empirical risk minimization for the random features model in the proportional limit was derived for ridge regression in [Mei and Montanari, 2022] and generalized to convex losses in [Gerace et al., 2020]. A key ingredient in this analysis is a *Gaussian equivalence principle* [Goldt et al., 2020, Goldt et al., 2022] proven in [Hu and Lu, 2020, Montanari and Saeed, 2022] stating that the statistics of the empirical risk minimizer is asymptotically equal to the one of an equivalent Gaussian problem with matching moments. More recently, Gaussian equivalence has been proven for two-layer neural tangent features in [Montanari and Saeed, 2022] and features coming from mixture models in [Gerace et al., 2022], and was conjectured to hold for a broader class including features coming from trained neural networks [Loureiro et al., 2021a]. Although the discussion in this manuscript focus in the random features case, our analysis can be readily extended to all cases in which Gaussian equivalence holds. We provide in Appendix B an extension of our main theoretical result to a general Gaussian covariate model with convex loss encompassing all these cases.

3 Results

3.1 Technical results

Let $\hat{\mu}_p$ denote the empirical spectral distribution of the matrix $\mathrm{F}\mathrm{F}^\top \in \mathbb{R}^{p \times p}$. In the following, we assume that in the proportional high-dimensional limit defined above, $\hat{\mu}_p$ weakly converges to an asymptotic spectral distribution μ on $\mathbb{R}_+$ with normalized second moment $\int \mu(\mathrm{d}x) x^2 = 1$. Further, assume $\kappa_0 = \mathbb{E}[\phi(z)]$, $\kappa_1 = \mathbb{E}[z\phi(z)]$ and $\kappa_\star^2 = \mathbb{E}[\phi(z)^2] - \kappa_1^2 - \kappa_0^2$ are all finite for $z \sim \mathcal{N}(0, 1)$. For simplicity of exposition, in the following we assume $\kappa_0 = 0$, which can always be obtained by letting $\phi \to \phi - \kappa_0$. We also define the effective noise $\tau_{\text{add}}^2 = 1 - \mathbb{E}_{x \sim \mu} \left[\frac{\kappa_1^2 x}{\kappa_1^2 x + \kappa_\star^2} \right]$.

The first step is to characterize the density $\rho_{\star,t}$ with $t \in \{\text{bo}, \text{erm}, \text{Lap}, \text{eb}\}$ defined in eq. (8). All relevant quantities depend on this density. In the asymptotic regime, the estimator $\hat{f}(\boldsymbol{x})$ is characterized by six quantities $(m, q, v, \hat{m}, \hat{q}, \hat{v})$ that are solutions of self-consistent equations.

Classifier	$g_t(y, \omega, v)$	$\hat{\pi}_t(x)$	$\hat{\tau}_t$
$\hat{f}_{\text{erm}}$	$\text{prox}_{\log \sigma(y\times\cdot)}(\omega)$	λ	0
$\hat{f}_{\text{Lap}}$	$\text{prox}_{\log \sigma(y\times\cdot)}(\omega)$	λ	$\mathbb{E}_{x\sim\mu}\left[\frac{\kappa_1^2 x+\kappa_\star}{\lambda+\hat{v}_\star(\kappa_1^2 x+\kappa_\star)}\right]$
$\hat{f}_{\text{eb}}$	$\partial\omega \log \int \sigma(\beta y \times z)\mathcal{N}(z\|\omega, v)\text{d}z$	λ	$v^\star$
$\hat{f}_{\text{bo}}$	$\partial\omega \log \int \sigma_{\tau_0^2+\tau_{\text{add}}^2}(y \times z)\mathcal{N}(z\|\omega, v)\text{d}z$	$\frac{\kappa_1^2 x}{(\kappa_1^2 x+\kappa_\star^2)^2}$	$v^\star + \tau_0^2 + \tau_{\text{add}}^2$

Table 1: Auxiliary functions and value of $\hat{\tau}_t$ for the different classifiers defined in Sec. 2.1.

Theorem 3.1 (Joint density)**.** *Let $\mathcal{D} = \{(\boldsymbol{x}^\mu, y^\mu)\}_{\mu=1}^n$ denote data independently drawn from the model defined in Equation (11). Consider $\hat{f}_t$, $t \in \{\text{bo}, \text{erm}, \text{Lap}, \text{eb}\}$ one of the classifiers defined in Sec. 2.1. Then, in the proportional high-dimensional limit where $n, d, p \to \infty$ with fixed $\alpha = {}^n\!/_p, \gamma = {}^p\!/_d$, the asymptotic joint density $\rho_{\star,t}$ defined in Equation (8) is given by $\rho_{\star,t}^{\lim}(a, b) = \lim_{p\to\infty} \rho_{\star,t}(a, b)$:*

$$\rho_{\star,t}^{\lim}(a,b) = \frac{\mathcal{N}\left(\begin{bmatrix}\sigma^{-1}_{\tau_0^2+\tau_{\text{add}}^2}(a)\\ \sigma^{-1}_{\hat{\tau}_t^2}(b)\end{bmatrix}\middle|\, \mathbf{0}_2, \Sigma_t\right)}{|\sigma'_{\tau_0^2+\tau_{\text{add}}^2}(\sigma^{-1}_{\tau_0^2+\tau_{\text{add}}^2}(a))||\sigma'_{\hat{\tau}_t^2}(\sigma^{-1}_{\hat{\tau}_t^2}(b))|} \tag{12}$$

where

$$\Sigma_t = \begin{bmatrix}1-\tau_{\text{add}}^2 & m_t^\star\\ m_t^\star & q_t^\star\end{bmatrix} \tag{13}$$

and the sufficient statistics $(m_t^\star, q_t^\star, v_t^\star) \in \mathbb{R}^3$ are the unique fixed points of the following system of equations:

$$\begin{cases} v = 2 \times \partial_{\hat{q}} \Psi_w(\hat{m}, \hat{q}, \hat{v}; \hat{\pi}_t) \\ q = 2 \times (\partial_{\hat{q}} \Psi_w - \partial_{\hat{v}} \Psi_w)(\hat{m}, \hat{q}, \hat{v}; \hat{\pi}_t) \\ m = \sqrt{\gamma}\partial_{\hat{m}} \Psi_w(\hat{m}, \hat{q}, \hat{v}; \hat{\pi}_t) \end{cases} \tag{14}$$

$$\begin{cases} \hat{v} &= -\alpha \mathbb{E}_{\xi\sim\mathcal{N}(0,q)}\left[\sum_y \mathcal{Z}_0\left(y, {}^m\!/_q \xi, v_\star\right) \partial_\omega g_t\left(y, \xi, v\right)\right] \\ \hat{q} &= \alpha \mathbb{E}_{\xi\sim\mathcal{N}(0,q)}\left[\sum_y \mathcal{Z}_0\left(y, {}^m\!/_q \xi, v_\star\right) g_t\left(y, \xi, v\right)^2\right] \\ \hat{m} &= \sqrt{\gamma}\alpha \mathbb{E}_{\xi\sim\mathcal{N}(0,q)}\left[\sum_y \partial_\omega \mathcal{Z}_0\left(y, {}^m\!/_q \xi, v_\star\right) g_t\left(y, \xi, v\right)\right] \end{cases}$$

where $\mathcal{Z}_0(y, \omega, v) = \sigma_{v+\tau_0^2+\tau_{\text{add}}^2}(y\omega)$, $v_\star = 1 - {}^{m^2}\!/_q - \tau_{\text{add}}^2$. The functions g_t and $\hat{\pi}_t$ and the scalar $\hat{\tau}_t$ depend on the estimator and the sufficient statistics, and are given in Table 1. Also :

$$\begin{aligned}\Psi_w(\hat{m}, \hat{q}, \hat{v}; \hat{\pi}) = &\frac{1}{2}\mathbb{E}_{x\sim\mu}\left[\frac{\hat{m}\kappa_1^2 x + \hat{q}(\kappa_1^2 x + \kappa_\star^2)}{\hat{\pi}(x) + \hat{v}(\kappa_1^2 x + \kappa_\star^2)}\right] \\ &- \frac{1}{2}\log(\hat{\pi}(x) + \hat{v}(\kappa_1^2 x + \kappa_\star^2))\,.\end{aligned} \tag{15}$$

Proof idea: Let $\boldsymbol{x} \sim \mathcal{N}(\mathbf{0}, {}^1\!/_d \text{I}_d)$. For any of the classifiers $t \in \{\text{bo}, \text{erm}, \text{Lap}, \text{eb}\}$ from Sec. 2.1, the 2-dimensional vector $(f_\star(\boldsymbol{x}), \hat{f}_t(\boldsymbol{x}))$ is asymptotically distributed as $(\sigma(z), \sigma_{\tilde{v}}(z_t'))$ for some $\tilde{v}$ that depends on the estimator, where $(z, z_t') \sim \mathcal{N}(\mathbf{0}_2, \Sigma_t)$, and

$$\Sigma_t = \frac{1}{d}\begin{pmatrix}||\boldsymbol{\theta}_\star||_2^2 & \hat{\boldsymbol{\theta}}_t^\top \Phi \boldsymbol{\theta}_\star \\ \boldsymbol{\theta}_\star^\top \Phi^\top \hat{\boldsymbol{\theta}}_t & \hat{\boldsymbol{\theta}}_t^\top \Omega \hat{\boldsymbol{\theta}}_t\end{pmatrix}$$

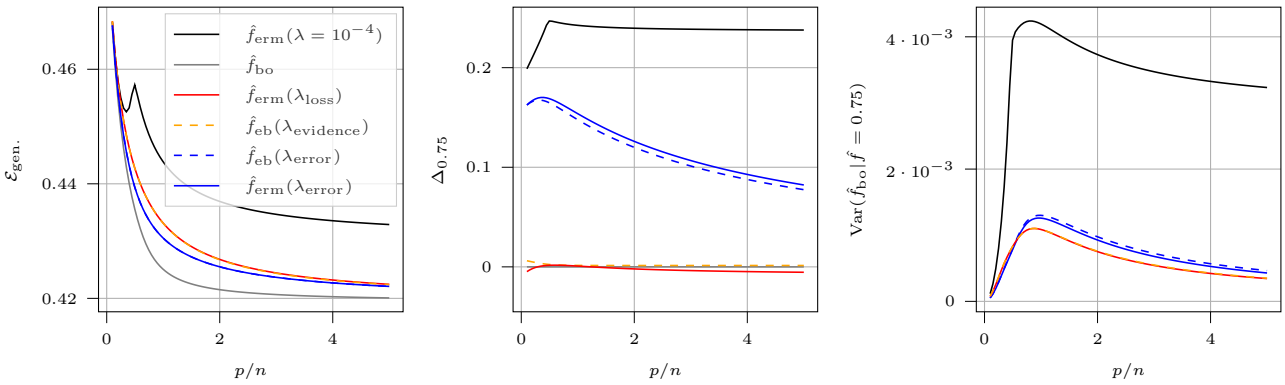

Figure 1: (**Left**) Test errors of the different methods as a function of the number of parameter per sample p/n. ERM, and Empirical Bayes (EB) are used with different penalizations. Here we use a logit teacher with $n/d = 2.0, \tau_0 = 1/2$ and `erf` activation. The curves $\hat{f}_{\rm eb}(\lambda_{\rm error})$ and $\hat{f}_{\rm erm}(\lambda_{\rm error})$ are very close and indistinguishable on the plot, as well as the curves $\hat{f}_{\rm eb}(\lambda_{\rm evidence})$ and $\hat{f}_{\rm erm}(\lambda_{\rm loss})$. Due to the intrinsic noise in the model the oracle error is $\mathcal{E}^{\star}_{gen.} \simeq 0.332$. (**Center**) Calibration at a level $\ell = 0.75$. (**Right**) Variance of $\hat{f}_{\rm bo}$ conditioned on the different other estimators.

where we defined the shorthand $\Phi = \kappa_1 \mathrm{F} \in \mathbb{R}^{p\times d}$ and $\Omega = \kappa_1^2 \mathrm{F}\mathrm{F}^\top + \kappa_\star^2 \mathrm{I}_p$ and $\hat{\boldsymbol{\theta}}_t$ is either the unique minimizer the empirical risk in eq. (2) for $t \in \{\mathrm{erm}, \mathrm{Lap}\}$ or the mean over the respective posterior distribution for $t \in \{\mathrm{bo}, \mathrm{eb}\}$. The computation of $\rho_{\star,t}$ thus boils down to computing the sufficient statistics $(m^\star, q^\star) := (\hat{\boldsymbol{\theta}}_t^\top \Phi \boldsymbol{\theta}_\star, \hat{\boldsymbol{\theta}}_t^\top \Omega \hat{\boldsymbol{\theta}}_t)$. For $\hat{f}_{\rm erm}$ on the random features model, the theorem follows then directly from [Dhifallah and Lu, 2020, Loureiro et al., 2021a], where $(m^\star, q^\star)$ is indeed proven to asymptotically obey a set of self-consistent "state-evolution" equations [Gerbelot and Berthier, 2021, Bayati and Montanari, 2011, Donoho and Montanari, 2016], mathematically equivalent to eqs. (14). We show in Appendix B.3 how to derive analogous results for $t \in \{\mathrm{bo}, \mathrm{Lap}, \mathrm{eb}\}$ with the heuristic replica and cavity method. The Laplace estimator can also be proven by the ERM theorem and the BO one with results from [Barbier et al., 2018, Gabrié et al., 2018]. Finally, the EB is proven under the technical condition that message-passing algorithm sample the posterior efficiently (which is expected to be the case in linear time for such measures [Zdeborová and Krzakala, 2016, Celentano et al., 2020, Barbier et al., 2021b]). Note that a similar theorem was used in [Clarté et al., 2022] for the simpler vanilla logistic model.

Corollary 3.2 (Test error and calibration)**.** *Under the conditions of Theorem 3.1, the asymptotic generalization error and calibration are given by:*

$$\mathcal{E}^{\rm lim}_{\rm gen.} = \iint_{b<0.5,a} a \times \rho^{\rm lim}_{\star,t}(a,b)\mathrm{d}a\mathrm{d}b + \iint_{b>0.5,a} (1-a) \times \rho^{\rm lim}_{\star,t}(a,b)\mathrm{d}a\mathrm{d}b \tag{16}$$

$$\Delta^{\rm lim}_{p} = p - \frac{\int a \times \rho^{\rm lim}_{\star,t}(a,p)\mathrm{d}a}{\int \rho^{\rm lim}_{\star,t}(a,p)\mathrm{d}a}\,. \tag{17}$$

3.2 Trade-off between performance and uncertainty

In sensitive applications of machine learning having a reliable estimation of the model's uncertainty can be as important as having accurate predictions. Therefore, a key question is *"can my model achieve good generalization while being calibrated?"*.

Comparing the performances: In Figure 1 (left) we compare the misclassification test error eq. (6) of the different classifiers defined in Sec. 2.1[2] as a function of the overparametrization ratio p/n at fixed sample complexity $n/d = 2$ for different choices of the hyperparameters (β, λ). First, note

[2]Note that by construction $\mathcal{E}_{\rm gen.}(\hat{f}_{\rm Lap}) = \mathcal{E}_{\rm gen.}(\hat{f}_{\rm erm})$.

the characteristic double descent behaviour of the empirical risk minimizer with $\lambda \to 0^+$, with the peak at the interpolating threshold corresponding in our setting to the existence of linear separator [Rosset et al., 2003]. As discussed in e.g. [Nakkiran et al., 2021] for neural networks and shown in e.g. [Gerace et al., 2020] for random features classification, this peak is mitigated by cross-validation on the ℓ_2 regularization $\lambda > 0$, which is shown in Fig. 1 with the blue and red full lines, corresponding to optimally tuning λ to minimize the misclassification error eq. (6) and the test loss respectively eq. (7).

It is interesting to contrast these ERM estimators to the empirical Bayes classifier, which averages over different classifiers. We see that, evaluating the empirical Bayes with a Gaussian prior of variance given by the cross-validated λ_{error} achieves almost identical performance to the ERM estimator, with a difference of the order of 10^{-5}.

An often quoted strength of the Bayesian approach is that model selection can be performed directly on the training data by evidence maximization over the model hyperparameters [MacKay, 1996]. Curiously, in our setting this yields a very close performance to ERM cross-validated with respect to the test loss, as shown in Fig. 1 (left) in dashed yellow line. Despite achieving similar performances in our setting, it is important to stress that these two classifiers are computationally radically different, as the empirical Bayes classifier requires sampling from a high-dimensional distribution which can be prohibitive in practice. These should be contrasted with the Bayes-optimal classifier, shown in solid grey, which by definition gives the best achievable performance at fixed data availability.

To summarise, from the point-of-view of the performance we observe no significant difference between Bayesian and ERM estimators, with (not surprisingly) best performance achieved by cross-validating over the misclassification error.

Calibration: Despite the relatively small difference in performance, the discussed classifiers are rather different in terms of calibration. Figure 1 (center) shows the calibration at fixed level $\ell = 0.75$ for the same classifiers. Note that the max-margin interpolator $\lambda \to 0^+$ produces consistently overconfident predictions. Indeed, we observe a maximum in the calibration curve around the interpolation threshold reminiscent of the double descent behaviour, with worst possible calibration $\Delta_\ell = \ell - 1/2$ corresponding to a confidence completely uncorrelated with the true class probabilities achieved at the interpolation transition. As noted in [Bai et al., 2021], overconfidence is inherent for unregularized logistic regression in high-dimensions, as it is present even when data is abundant with respect to the number of parameters. However, in their simpler setting of matched linear classifiers the number of parameters is equal to the input dimension $p = d$, and therefore overparametrization cannot be distinguished from high-dimensionality. Indeed, they observe an asymptotic scaling of the calibration $\Delta_\ell \sim d/n$, which suggests that overconfidence increases with the number of parameters. Our setting allow us to decouple the number of parameters p from the data dimension d, suggesting instead that overparametrization can improve calibration at fixed number of samples.

More strikingly, we observe that optimal regularization does not mitigate this double descent-like behaviour in the calibration, which is in contrast with what happens with the error itself that becomes monotonic when optimally regularized. Indeed, while cross-validating with respect to the misclassification error achieves the best accuracy, it produces consistently overconfident predictions for both the empirical risk minimizer and the empirical Bayes classifiers. On the other hand, cross-validation with respect to the loss produces better calibrated estimates, with an interesting non-monotonic behaviour crossing from over- to underconfidence as a function of overparametrization. In contrast, maximising the evidence yields better calibrated estimation with a monotonic calibration curve very close to zero.

To summarise, we observe a fundamental trade-off between optimising the accuracy of classification and obtaining calibrated classifiers. A similar discussion holds for other calibration levels and for the expected calibration error eq. (10), as shown in Appendix E.

Conditional variance: Theorem 3.1 gives us access to a rich set of uncertainty measures, of which the calibration is a particular example. For instance, we have access to the full distribution of the Bayes-optimal classifier $\hat{f}_{\text{bo}}$ conditioned on the predictors defined in Sec. 2.1. Note that since $\mathbb{E}(\hat{f}_{\text{bo}}|\hat{f} = \ell) = \mathbb{E}(f_\star|\hat{f} = \ell) = \ell - \Delta_\ell(\hat{f})$, the mean of this conditional distribution is equal to the calibration up to a constant. A natural measure of uncertainty beyond the calibration is the variance of this conditional

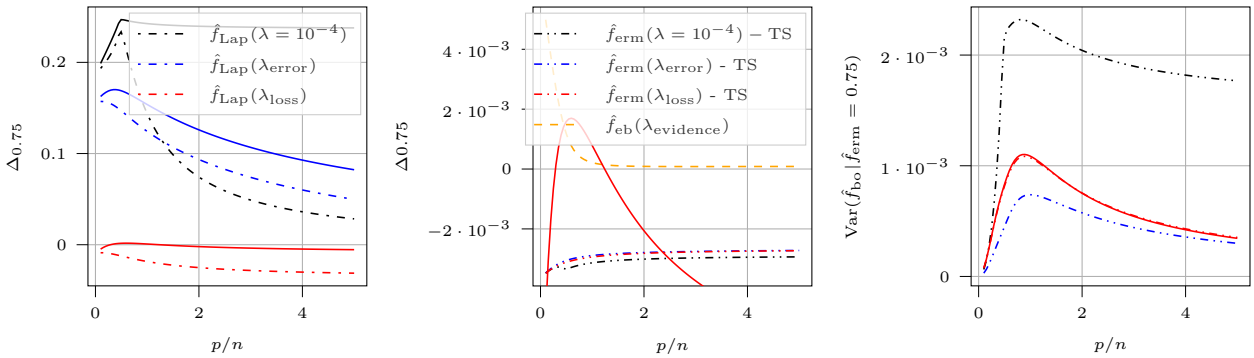

Figure 2: (**Left**) Calibration at level $\ell = 0.75$ of $\hat{f}_{\text{erm}}$ (solid lines, refer to Figure 1 for the legend) and $\hat{f}_{\text{Lap}}$ with the three different regularizations. (**Center**) Calibration at level $\ell = 0.75$ of $\hat{f}_{\text{erm}}$ after temperature scaling (TS), compared to $\hat{f}_{\text{eb}}$ (dashed yellow) and $\hat{f}_{\text{loss}}$ (full red) for reference. (**Right**) Variance of $\hat{f}_{\text{bo}}$ conditioned on $\hat{f}_{\text{erm}} = 0.75$ after temperature scaling, compared to variance at $\hat{f}_{\text{loss}}$ (full red) and $\hat{f}_{\text{eb}}$ (dashed yellow).

distribution $\text{Var}(\hat{f}_{\text{bo}} | \hat{f} = \ell)$, which quantifies how much the prediction $\hat{f}(\boldsymbol{x}) = \ell$ inform us on $\hat{f}_{\text{bo}}(\boldsymbol{x})$, which is by definition the best achievable classifier at finite availability of data.

Theorem 3.3 (Conditional variance)**.** *Under the same setting as Theorem 3.1, an explicit expression for the conditional variance can be derived for any of the classifiers* $t \in \{\text{erm}, \text{bo}, \text{Lap}, \text{eb}\}$*:*

$$\text{Var}(\hat{f}_{\text{bo}}(\boldsymbol{x}) | \hat{f}_t(\boldsymbol{x}) = \ell) = \int \text{d}a \; \sigma_{\hat{v}^\star_{\text{bo}} + \tau_0^2 + \tau^2_{\text{add}}}(a)^2 \times \mathcal{N}\left(a | m_t^\star / q_t^\star \sigma^{-1}_{\hat{\tau}_t}(\ell), q^\star_{\text{bo}} - m_t^{\star 2} / q_t^\star\right) - (\ell - \Delta_\ell)^2 \quad (18)$$

where $(m_t^\star, q_t^\star, q^\star_{\text{bo}})$ *are solutions to the self-consistent equations eq.* (14) *and* Δ_ℓ *is the asymptotic calibration eq.* (17)*.*

The detailed derivation and the proof, that follows from Theorem 3.1, are shown in Appendix D. Figure 1 (right) shows this conditional variance as a function of the overparametrization. We note that here better calibration correlates well with lower variance. We also observe behaviour reminiscent of double descent in the value of the conditional variance that does not go away with optimal regularization and the corresponding peak is even more pronounced than in the calibration.

3.3 Temperature scaling

Temperature scaling is a calibration method introduced in [Guo et al., 2017] to mitigate overconfidence in trained neural networks. It is applied after training, and consists in introducing a "temperature" scaling parameter on the last layer pre-activations $\hat{f}_{\text{erm}}(\boldsymbol{x}) = \sigma(\hat{\boldsymbol{\theta}}_{\text{erm}}^\top \boldsymbol{\varphi}(\boldsymbol{x})/T)$. It is then tuned to minimize the validation loss. In our analysis, this corresponds to simply re-scaling the predictor $\hat{\boldsymbol{\theta}}_{\text{erm}} \to \hat{\boldsymbol{\theta}}_{\text{erm}}/T$, & Thm. 3.1 thus applies mutatis mutandis.

Figure 2 (center) compares the calibration at level $\ell = 0.75$ of the regularized empirical risk minimizers with λ_{loss} and λ_{error} after temperature scaling with the empirical Bayes classifier with $\lambda_{\text{evidence}}$ and ERM classifier at λ_{loss}, the best calibrated in our setting so far. We observe that the temperature scaling yields very similar calibrations for λ_{loss} and λ_{error}. While empirical Bayes remains the best calibrated estimator, temperature scaling has a calibration around 0.1%, which would be satisfying in most practical scenarios. We also observe that the maximum around the interpolation threshold is not present in the calibration curves after temperature scaling.

Looking at the variance of $\hat{f}_{\text{bo}}$ conditioned on $\hat{f}_{\text{erm}}$ after temperature scaling we see that it is lower for λ_{error} than for λ_{loss}, see Fig. 2 (right). We see that again the variance has an increase in the vicinity of the interpolation threshold, reminiscent of the double descent behaviour. As discussed in the previous section, we aim to have the lowest variance possible to ensure that the uncertainty estimation is accurate not only on average but also point-wise. It appears that cross-validating the empirical risk minimizer on the misclassification error and then applying temperature scaling gives an estimator that both has the best test error and is very well calibrated, both on average and point-wise.

3.4 The calibration of the Laplace approximation

Estimating any of the Bayesian classifiers in Sec. 2.1 is computationally demanding, since they involve a sampling over a high-dimensional distribution. This has motivated practitioners to develop different approximations for making Bayesian methods more efficient. These include Bayesian dropout [Gal and Ghahramani, 2016], deep ensembles [Lakshminarayanan et al., 2017], stochastic gradient Langevin dynamics [Welling and Teh, 2011] and the Laplace approximation [Ritter et al., 2018, Daxberger et al., 2021], among others. The Laplace approximation was introduced by [MacKay, 1992] in the context of Gaussian processes, and consists of approximating the posterior distribution by a Gaussian density centred around the empirical risk minimizer - or equivalently to a low-temperature expansion of the posterior - see eq. (5). By construction, the Laplace classifier has the same misclassification error as the empirical risk minimizer, and hence can be effectively seen as endowing this point-estimator with a covariance given by the inverse of the Hessian evaluated at the minimum. Although computing the Hessian of the empirical risk for a deep neural network can be costly, an approximate scheme has been recently proposed [Ritter et al., 2018, Daxberger et al., 2021], making Laplace a viable uncertainty estimation technique for deep learning.

On the theory side, sharp results have been limited to the Gaussian process and ridge regression setting, where the Laplace approximation is exact [Sollich, 1998, Sollich, 2001]. While exact asymptotic results characterizing the statistics of the logit estimator in high-dimensions abound [Sur and Candès, 20 Gerace et al., 2020, Aubin et al., 2020, Deng et al., 2022], to our best knowledge the asymptotic spectral distribution of the Hessian at the minimum is missing. Recently, [Liao and Mahoney, 2021] has computed the asymptotic spectral distribution of the Hessian in a matched logit model under the assumption that the weights are uncorrelated with the input data. Hence, their their results do not apply for the empirical risk minimizer, and cannot be used to characterize the uncertainty of the Laplace classifier. Characterizing the Hessian of the logistic risk eq. (2) at the minimizer is a challenging technical result which we believe is of independent interest to the scope of the discussion in this manuscript.

Theorem 3.4 (Hessian of logit, informal)**.** *Let*

$$\mathcal{H}(\boldsymbol{\theta}) := \sum_{\mu\in[n]} (\sigma'(y^\mu \boldsymbol{\theta}^\top \boldsymbol{\varphi}(\boldsymbol{x}^\mu)) - 1)\boldsymbol{\varphi}(\boldsymbol{x}^\mu)\boldsymbol{\varphi}(\boldsymbol{x}^\mu)^\top + \lambda I_p$$

denote the Hessian of the logistic empirical risk eq. (2), and denote $\hat{\boldsymbol{\theta}}_{\rm erm} = \arg\min_{\boldsymbol{\theta}} \hat{\mathcal{R}}_n(\boldsymbol{\theta})$ *its minimizer. Then, under the same conditions of Thm. 3.1 and additional technical assumptions, the following asymptotic characterization holds:*

$$\mathcal{H}^{-1}(\hat{\boldsymbol{\theta}}_{\rm erm}) \underset{p\to\infty}{\asymp} \left(\hat{v}^\star \left(\kappa_1^2 FF^\top + \kappa_\star^2 I_p\right) + \lambda I_p\right)^{-1} \tag{19}$$

where $\hat{v}^\star \in \mathbb{R}$ *is the solution of the self-consistent eq. (14).*

A derivation of this result is discussed in Appendix C in the more general context of the Gaussian covariate model. With Thm. 3.4 in hands, we can characterize the asymptotic calibration of the Laplace classifier for our model.

Figure 2 (left) shows the calibration curve at level $\ell = 0.75$, at sample complexity $n/d = 2$ and noise variance $\tau_0 = 0.5$ as a function of the number of parameters. As mentioned in the introduction, we observe here that $\hat{f}_{\rm Lap}$ is always less confident than $\hat{f}_{\rm erm}$, due to the concavity of σ. While this might seem desirable in the scenarios where ERM is very overconfident, e.g. for $\lambda \to 0^+$ or $\lambda_{\rm error}$, it hurts calibration when the classifier is well-calibrated as for $\lambda_{\rm loss}$. Moreover, it highly depends on the sample complexity and noise variance, see Appendix E in the supplementary material where we show a setting in which the Laplace approximation yields an underconfident classifier even in the $\lambda \to 0^+$ at mild overparametrization. Then, the Laplace approximation seems to be an unreliable way to control the calibration of the estimators, contrary to temperature scaling.

4 Conclusion

In this paper, we studied the performance of different frequentist and Bayesian classifiers for random features classification. In the high-dimensional limit, the asymptotic behaviour of these algorithms

can be precisely characterized. Our first contribution is the derivation of the Bayes-optimal estimator. By definition it is the estimator with the best possible performance, and although it is inaccessible in practice, it provides a baseline to compare the classifiers. Then, we compared the generalization error of frequentist and Bayesian approaches, showing they yield very similar test error. We then focused on uncertainty quantification, and showed there is a trade-off between generalization and calibration in our model. Moreover, we observed a non-monotonic behaviour of the calibration curve for certain estimators, akin to the famous *double-descent* phenomenon for the test error. Finally, we compared two popular approaches for post-training calibration: temperature scaling and the Laplace approximation, benchmarking them against the baseline classifiers. In our model, we observe that temperature scaling on top of cross-validating the empirical risk classifier on the accuracy achieves the best result : it has both the lowest test error and best calibration. Moreover, despite requiring a validation set, in practice it is a computationally more efficient method than the Bayesian approach, which requires sampling from a high-dimensional distribution. The code used in this project is available at `github.com/SPOC-group/double_descent_uncertainty`.

Limitations: It is worth pointing some limitation of our results. The first resides in the (nevertheless classical) Gaussian assumption for the data. We note, however, that there are good reasons to believe that this can be a very good model in high-dimensions [Hu and Lu, 2020, Montanari and Saeed, 2022]. A second limitation of is the lack of feature learning. While many of the uncertainty quantification methods discussed here apply directly to the last layer of trained neural networks, other methods considered in the literature apply to the full architecture [Abdar et al., 2021]. Since the performance of deep neural networks can be largely attributed to feature learning, it shall be important to take it into account in theoretical studies of uncertainty. We hope that our work can offer a starting point towards this more ambitious goal.

Acknowledgements

We thank Pierre-Alexandre Mattei, Yevgeny Seldin, Anshuk Uppal, Kristoffer Stensbo-Smidt, Simon Bartels and Melih Kandemir for valuable discussions. We acknowledge funding from the ERC under the European Union's Horizon 2020 Research and Innovation Program Grant Agreement 714608-SMiLe, and by the Swiss National Science Foundation grant SNFS OperaGOST, 200021_200390.

References

[Abbasi et al., 2019] Abbasi, E., Salehi, F., and Hassibi, B. (2019). Universality in learning from linear measurements. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

[Abdar et al., 2021] Abdar, M., Pourpanah, F., Hussain, S., Rezazadegan, D., Liu, L., Ghavamzadeh, M., Fieguth, P., Cao, X., Khosravi, A., Acharya, U. R., Makarenkov, V., and Nahavandi, S. (2021). A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Information Fusion*, 76:243–297.

[Aizenman et al., 2006] Aizenman, M., Sims, R., and Starr, S. L. (2006). Mean-field spin glass models from the cavity–rost perspective. *arXiv preprint math-ph/0607060*.

[Alexos et al., 2022] Alexos, A., Boyd, A. J., and Mandt, S. (2022). Structured stochastic gradient MCMC. In Chaudhuri, K., Jegelka, S., Song, L., Szepesvari, C., Niu, G., and Sabato, S., editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 414–434. PMLR.

[Aubin et al., 2020] Aubin, B., Krzakala, F., Lu, Y., and Zdeborová, L. (2020). Generalization error in high-dimensional perceptrons: Approaching bayes error with convex optimization. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H., editors, *Advances in Neural Information Processing Systems*, volume 33, pages 12199–12210. Curran Associates, Inc.

[Aubin et al., 2021] Aubin, B., Loureiro, B., Maillard, A., Krzakala, F., and Zdeborová, L. (2021). The spiked matrix model with generative priors. *IEEE Transactions on Information Theory*, 67(2):1156–1181.

[Aubin et al., 2018] Aubin, B., Maillard, A., barbier, j., Krzakala, F., Macris, N., and Zdeborová, L. (2018). The committee machine: Computational to statistical gaps in learning a two-layers neural network. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

[Bai et al., 2021] Bai, Y., Mei, S., Wang, H., and Xiong, C. (2021). Don't just blame over-parametrization for over-confidence: Theoretical analysis of calibration in binary classification. In Meila, M. and Zhang, T., editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 566–576. PMLR.

[Barbier et al., 2021a] Barbier, J., Chen, W.-K., Panchenko, D., and Sáenz, M. (2021a). Performance of bayesian linear regression in a model with mismatch. *arXiv preprint arXiv:2107.06936*.

[Barbier et al., 2019] Barbier, J., Krzakala, F., Macris, N., Miolane, L., and Zdeborová, L. (2019). Optimal errors and phase transitions in high-dimensional generalized linear models. 116(12):5451–5460.

[Barbier et al., 2018] Barbier, J., Macris, N., Maillard, A., and Krzakala, F. (2018). The mutual information in random linear estimation beyond iid matrices. In *2018 IEEE International Symposium on Information Theory (ISIT)*, pages 1390–1394. IEEE.

[Barbier et al., 2021b] Barbier, J., Panchenko, D., and Sáenz, M. (2021b). Strong replica symmetry for high-dimensional disordered log-concave Gibbs measures. *Information and Inference: A Journal of the IMA*, 11(3):1079–1108.

[Bartlett et al., 2020] Bartlett, P. L., Long, P. M., Lugosi, G., and Tsigler, A. (2020). Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, 117(48):30063–30070.

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

[Bayati and Montanari, 2012] Bayati, M. and Montanari, A. (2012). The lasso risk for gaussian matrices. *IEEE Transactions on Information Theory*, 58(4):1997–2017.

[Belkin et al., 2019] Belkin, M., Hsu, D., Ma, S., and Mandal, S. (2019). Reconciling modern machine-learning practice and the classical bias-variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854.

[Benigni and Péché, 2021] Benigni, L. and Péché, S. (2021). Eigenvalue distribution of some nonlinear models of random matrices. *Electronic Journal of Probability*, 26(none):1 – 37.

[Berthier et al., 2019] Berthier, R., Montanari, A., and Nguyen, P.-M. (2019). State evolution for approximate message passing with non-separable functions. *Information and Inference: A Journal of the IMA*, 9(1):33–79.

[Brosse et al., 2020] Brosse, N., Riquelme, C., Martin, A., Gelly, S., and Moulines, E. (2020). On last-layer algorithms for classification: Decoupling representation from uncertainty estimation.

[Bruce and Saad, 1994] Bruce, A. D. and Saad, D. (1994). Statistical mechanics of hypothesis evaluation. *Journal of Physics A: Mathematical and General*, 27(10):3355–3363.

[Candes and Sur, 2018] Candes, E. J. and Sur, P. (2018). The phase transition for the existence of the maximum likelihood estimate in high-dimensional logistic regression.

[Carmona and Hu, 2004] Carmona, P. and Hu, Y. (2004). Universality in sherrington-kirkpatrick's spin glass model.

[Celentano et al., 2020] Celentano, M., Montanari, A., and Wu, Y. (2020). The estimation error of general first order methods. In *Conference on Learning Theory*, pages 1078–1141. PMLR.

[Chatterjee, 2005] Chatterjee, S. (2005). A simple invariance theorem.

[Clarté et al., 2022] Clarté, L., Loureiro, B., Krzakala, F., and Zdeborová, L. (2022). Theoretical characterization of uncertainty in high-dimensional linear classification.

[Cornacchia et al., 2022] Cornacchia, E., Mignacco, F., Veiga, R., Gerbelot, C., Loureiro, B., and Zdeborová, L. (2022). Learning curves for the multi-class teacher-student perceptron.

[D'Ascoli et al., 2020] D'Ascoli, S., Refinetti, M., Biroli, G., and Krzakala, F. (2020). Double trouble in double descent: Bias and variance(s) in the lazy regime. In III, H. D. and Singh, A., editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 2280–2290. PMLR.

[Daxberger et al., 2021] Daxberger, E., Kristiadi, A., Immer, A., Eschenhagen, R., Bauer, M., and Hennig, P. (2021). Laplace Redux - Effortless Bayesian Deep Learning. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P. S., and Vaughan, J. W., editors, *Advances in Neural Information Processing Systems*, volume 34, pages 20089–20103. Curran Associates, Inc.

[Deng et al., 2022] Deng, Z., Kammoun, A., and Thrampoulidis, C. (2022). A model of double descent for high-dimensional binary linear classification. *Information and Inference: A Journal of the IMA*, 11(2):435–495.

[Dhifallah and Lu, 2020] Dhifallah, O. and Lu, Y. M. (2020). A precise performance analysis of learning with random features.

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

[Donoho and Montanari, 2016] Donoho, D. and Montanari, A. (2016). High dimensional robust m-estimation: Asymptotic variance via approximate message passing. *Probability Theory and Related Fields*, 166(3):935–969.

[Donoho and Tanner, 2009] Donoho, D. and Tanner, J. (2009). Observed universality of phase transitions in high-dimensional geometry, with implications for modern data analysis and signal processing. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 367(1906):4273–4293.

[Erdos et al., 2009] Erdos, L., Schlein, B., and Yau, H.-T. (2009). Universality of random matrices and local relaxation flow.

[Erdos et al., 2010] Erdos, L., Yau, H.-T., and Yin, J. (2010). Bulk universality for generalized wigner matrices.

[Gabrié et al., 2018] Gabrié, M., Manoel, A., Luneau, C., Macris, N., Krzakala, F., Zdeborová, L., et al. (2018). Entropy and mutual information in models of deep neural networks. *Advances in Neural Information Processing Systems*, 31.

[Gal and Ghahramani, 2016] Gal, Y. and Ghahramani, Z. (2016). Dropout as a bayesian approximation: Representing model uncertainty in deep learning.

[Gawlikowski et al., 2022] Gawlikowski, J., Tassi, C. R. N., Ali, M., Lee, J., Humt, M., Feng, J., Kruspe, A., Triebel, R., Jung, P., Roscher, R., Shahzad, M., Yang, W., Bamler, R., and Zhu, X. X. (2022). A survey of uncertainty in deep neural networks.

[Geiger et al., 2019] Geiger, M., Spigler, S., d'Ascoli, S., Sagun, L., Baity-Jesi, M., Biroli, G., and Wyart, M. (2019). Jamming transition as a paradigm to understand the loss landscape of deep neural networks. *Phys. Rev. E*, 100:012115.

[Geman et al., 1992] Geman, S., Bienenstock, E., and Doursat, R. (1992). Neural networks and the bias/variance dilemma. *Neural computation*, 4(1):1–58.

[Gerace et al., 2022] Gerace, F., Krzakala, F., Loureiro, B., Stephan, L., and Zdeborová, L. (2022). Gaussian universality of linear classifiers with random labels in high-dimension.

[Gerace et al., 2020] Gerace, F., Loureiro, B., Krzakala, F., Mezard, M., and Zdeborova, L. (2020). Generalisation error in learning with random features and the hidden manifold model. In III, H. D. and Singh, A., editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 3452–3462. PMLR.

[Gerbelot and Berthier, 2021] Gerbelot, C. and Berthier, R. (2021). Graph-based approximate message passing iterations.

[Goldt et al., 2022] Goldt, S., Loureiro, B., Reeves, G., Krzakala, F., Mezard, M., and Zdeborova, L. (2022). The gaussian equivalence of generative models for learning with shallow neural networks. In Bruna, J., Hesthaven, J., and Zdeborova, L., editors, *Proceedings of the 2nd Mathematical and Scientific Machine Learning Conference*, volume 145 of *Proceedings of Machine Learning Research*, pages 426–471. PMLR.

[Goldt et al., 2020] 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.

[Graves, 2011] Graves, A. (2011). Practical variational inference for neural networks. In Shawe-Taylor, J., Zemel, R., Bartlett, P., Pereira, F., and Weinberger, K., editors, *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc.

[Guo et al., 2017] Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. (2017). On calibration of modern neural networks. In Precup, D. and Teh, Y. W., editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1321–1330. PMLR.

[Hastie et al., 2022] Hastie, T., Montanari, A., Rosset, S., and Tibshirani, R. J. (2022). Surprises in high-dimensional ridgeless least squares interpolation. *The Annals of Statistics*, 50(2):949 – 986.

[Hein et al., 2019] Hein, M., Andriushchenko, M., and Bitterwolf, J. (2019). Why relu networks yield high-confidence predictions far away from the training data and how to mitigate the problem. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[Hu and Lu, 2020] Hu, H. and Lu, Y. M. (2020). Universality laws for high-dimensional learning with random features.

[Jospin et al., 2022] Jospin, L. V., Laga, H., Boussaid, F., Buntine, W., and Bennamoun, M. (2022). Hands-on bayesian neural networks—a tutorial for deep learning users. *IEEE Computational Intelligence Magazine*, 17(2):29–48.

[Karoui, 2009] Karoui, N. E. (2009). Concentration of measure and spectra of random matrices: Applications to correlation matrices, elliptical distributions and beyond. *The Annals of Applied Probability*, 19(6):2362 – 2405.

[Karoui, 2010] Karoui, N. E. (2010). The spectrum of kernel random matrices. *The Annals of Statistics*, 38(1):1 – 50.

[Karoui et al., 2013] Karoui, N. E., Bean, D., Bickel, P. J., Lim, C., and Yu, B. (2013). On robust regression with high-dimensional predictors. *Proceedings of the National Academy of Sciences*, 110(36):14557–14562.

[Korada and Montanari, 2010] Korada, S. B. and Montanari, A. (2010). Applications of lindeberg principle in communications and statistical learning.

[Kristiadi et al., 2020] Kristiadi, A., Hein, M., and Hennig, P. (2020). Being bayesian, even just a bit, fixes overconfidence in ReLU networks. In III, H. D. and Singh, A., editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 5436–5446. PMLR.

[Krzakala et al., 2012] Krzakala, F., Mézard, M., Sausset, F., Sun, Y., and Zdeborová, L. (2012). Probabilistic reconstruction in compressed sensing: algorithms, phase diagrams, and threshold achieving matrices. *Journal of Statistical Mechanics: Theory and Experiment*, 2012(08):P08009.

[Lakshminarayanan et al., 2017] Lakshminarayanan, B., Pritzel, A., and Blundell, C. (2017). Simple and scalable predictive uncertainty estimation using deep ensembles. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

[Liao and Couillet, 2018] Liao, Z. and Couillet, R. (2018). On the spectrum of random features maps of high dimensional data. In Dy, J. and Krause, A., editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 3063–3071. PMLR.

[Liao and Mahoney, 2021] Liao, Z. and Mahoney, M. W. (2021). Hessian eigenspectra of more realistic nonlinear models. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W., editors, *Advances in Neural Information Processing Systems*, volume 34, pages 20104–20117. Curran Associates, Inc.

[Liu et al., 2020] Liu, J., 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. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H., editors, *Advances in Neural Information Processing Systems*, volume 33, pages 7498–7512. Curran Associates, Inc.

[Loureiro et al., 2021a] Loureiro, B., Gerbelot, C., Cui, H., Goldt, S., Krzakala, F., Mezard, M., and Zdeborová, L. (2021a). Learning curves of generic features maps for realistic datasets with a teacher-student model. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W., editors, *Advances in Neural Information Processing Systems*, volume 34, pages 18137–18151. Curran Associates, Inc.

[Loureiro et al., 2022] Loureiro, B., Gerbelot, C., Refinetti, M., Sicuro, G., and Krzakala, F. (2022). Fluctuations, bias, variance & ensemble of learners: Exact asymptotics for convex losses in high-dimension. In Chaudhuri, K., Jegelka, S., Song, L., Szepesvari, C., Niu, G., and Sabato, S., editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 14283–14314. PMLR.

[Loureiro et al., 2021b] Loureiro, B., Sicuro, G., Gerbelot, C., Pacco, A., Krzakala, F., and Zdeborová, L. (2021b). Learning gaussian mixtures with generalized linear models: Precise asymptotics in high-dimensions. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W., editors, *Advances in Neural Information Processing Systems*, volume 34, pages 10144–10157. Curran Associates, Inc.

[MacKay, 1992] MacKay, D. J. C. (1992). Bayesian Interpolation. *Neural Computation*, 4(3):415–447.

[MacKay, 1996] MacKay, D. J. C. (1996). *Hyperparameters: Optimize, or Integrate Out?*, pages 43–59. Springer Netherlands, Dordrecht.

[Maddox et al., 2019] Maddox, W. J., Garipov, T., Izmailov, P., Vetrov, D. P., and Wilson, A. G. (2019). A simple baseline for bayesian uncertainty in deep learning. In *NeurIPS*.

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

[Marion and Saad, 1994] Marion, G. and Saad, D. (1994). Hyperparameters evidence and generalisation for an unrealisable rule. In Tesauro, G., Touretzky, D., and Leen, T., editors, *Advances in Neural Information Processing Systems*, volume 7. MIT Press.

[Marion and Saad, 1995] Marion, G. and Saad, D. (1995). A statistical mechanical analysis of a bayesian inference scheme for an unrealizable rule. *Journal of Physics A: Mathematical and General*, 28(8):2159–2171.

[Mattei, 2019] Mattei, P.-A. (2019). A parsimonious tour of bayesian model uncertainty.

[Mei and Montanari, 2022] Mei, S. and Montanari, A. (2022). The generalization error of random features regression: Precise asymptotics and the double descent curve. *Communications on Pure and Applied Mathematics*, 75(4):667–766.

[Mezard and Montanari, 2009] Mezard, M. and Montanari, A. (2009). *Information, physics, and computation*. Oxford University Press.

[Mézard et al., 1987] Mézard, M., Parisi, G., and Virasoro, M. A. (1987). *Spin glass theory and beyond: An Introduction to the Replica Method and Its Applications*, volume 9. World Scientific Publishing Company.

[Montanari et al., 2020] 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]*.

[Montanari and Saeed, 2022] Montanari, A. and Saeed, B. N. (2022). Universality of empirical risk minimization. In Loh, P.-L. and Raginsky, M., editors, *Proceedings of Thirty Fifth Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pages 4310–4312. PMLR.

[Mukhoti et al., 2020] Mukhoti, J., Kulharia, V., Sanyal, A., Golodetz, S., Torr, P., and Dokania, P. (2020). Calibrating deep neural networks using focal loss. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H., editors, *Advances in Neural Information Processing Systems*, volume 33, pages 15288–15299. Curran Associates, Inc.

[Nakkiran et al., 2020] Nakkiran, P., Kaplun, G., Bansal, Y., Yang, T., Barak, B., and Sutskever, I. (2020). Deep double descent: Where bigger models and more data hurt. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

[Nakkiran et al., 2021] Nakkiran, P., Venkat, P., Kakade, S. M., and Ma, T. (2021). Optimal regularization can mitigate double descent. In *International Conference on Learning Representations*.

[Nguyen et al., 2015] Nguyen, A., Yosinski, J., and Clune, J. (2015). Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 427–436.

[Niculescu-Mizil and Caruana, 2005] Niculescu-Mizil, A. and Caruana, R. (2005). Predicting good probabilities with supervised learning. ICML '05, page 625–632, New York, NY, USA. Association for Computing Machinery.

[Panahi and Hassibi, 2017] Panahi, A. and Hassibi, B. (2017). A universal analysis of large-scale regularized least squares solutions. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

[Pennington and Worah, 2017] Pennington, J. and Worah, P. (2017). Nonlinear random matrix theory for deep learning. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

[Platt, 2000] Platt, J. (2000). Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Adv. Large Margin Classif.*, 10.

[Rahimi and Recht, 2007] Rahimi, A. and Recht, B. (2007). Random features for large-scale kernel machines. In Platt, J., Koller, D., Singer, Y., and Roweis, S., editors, *Advances in Neural Information Processing Systems*, volume 20. Curran Associates, Inc.

[Rangan, 2011] Rangan, S. (2011). Generalized approximate message passing for estimation with random linear mixing. In *2011 IEEE International Symposium on Information Theory Proceedings*. IEEE.

[Ritter et al., 2018] Ritter, H., Botev, A., and Barber, D. (2018). A scalable laplace approximation for neural networks. In *International Conference on Learning Representations*.

[Rosset et al., 2003] Rosset, S., Zhu, J., and Hastie, T. (2003). Margin maximizing loss functions. In Thrun, S., Saul, L., and Schölkopf, B., editors, *Advances in Neural Information Processing Systems*, volume 16. MIT Press.

[Sollich, 1998] Sollich, P. (1998). Learning curves for gaussian processes. In Kearns, M., Solla, S., and Cohn, D., editors, *Advances in Neural Information Processing Systems*, volume 11. MIT Press.

[Sollich, 2001] Sollich, P. (2001). Gaussian process regression with mismatched models. In Dietterich, T., Becker, S., and Ghahramani, Z., editors, *Advances in Neural Information Processing Systems*, volume 14. MIT Press.

[Spigler et al., 2019] Spigler, S., Geiger, M., d'Ascoli, S., Sagun, L., Biroli, G., and Wyart, M. (2019). A jamming transition from under- to over-parametrization affects generalization in deep learning. *Journal of Physics A: Mathematical and Theoretical*, 52(47):474001.

[Stojnic, 2013] Stojnic, M. (2013). A framework to characterize performance of lasso algorithms.

[Sur and Candès, 2019] Sur, P. and Candès, E. J. (2019). A modern maximum-likelihood theory for high-dimensional logistic regression. *Proceedings of the National Academy of Sciences*, 116(29):14516–14525.

[Tao and Vu, 2012] Tao, T. and Vu, V. (2012). Random matrices: universal properties of eigenvectors. *Random Matrices: Theory and Applications*, 01(01):1150001.

[Thrampoulidis et al., 2018] Thrampoulidis, C., Abbasi, E., and Hassibi, B. (2018). Precise error analysis of regularized m -estimators in high dimensions. *IEEE Transactions on Information Theory*, 64(8):5592–5628.

[Thrampoulidis et al., 2015] Thrampoulidis, C., Oymak, S., and Hassibi, B. (2015). Regularized linear regression: A precise analysis of the estimation error. In Grünwald, P., Hazan, E., and Kale, S., editors, *Proceedings of The 28th Conference on Learning Theory*, volume 40 of *Proceedings of Machine Learning Research*, pages 1683–1709, Paris, France. PMLR.

[Wasserman, 2013] Wasserman, L. (2013). *All of Statistics: A Concise Course in Statistical Inference*. Springer Texts in Statistics. Springer New York.

[Welling and Teh, 2011] Welling, M. and Teh, Y. W. (2011). Bayesian learning via stochastic gradient langevin dynamics. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, ICML'11, page 681–688, Madison, WI, USA. Omnipress.

[Zdeborová and Krzakala, 2016] Zdeborová, L. and Krzakala, F. (2016). Statistical physics of inference: Thresholds and algorithms. *Advances in Physics*.

Appendix

A Gaussian equivalence

As discussed in the main, our analysis of the random features model introduced in Sec. 2.2 relies on a recent progress in high-dimensional statistics known as the *Gaussian equivalence theorem* (GET). In this Appendix, we recall the reader of the main results in this line of work.

A.1 Informal discussion and key idea

For convenience let's first recall the model of interest. Consider data $(\boldsymbol{x}^\mu, y^\mu)_{\mu\in[n]} \in \mathbb{R}^d \times \mathcal{Y}$ which we assume was independently drawn from the following model:

$$y^\mu = f_\star(\boldsymbol{\theta}_\star^\top \boldsymbol{x}^\mu/\sqrt{d}), \qquad \boldsymbol{x}^\mu \sim \mathcal{N}(\boldsymbol{0}, \mathrm{I}_d), \qquad \boldsymbol{\theta}_\star \sim \mathcal{N}(\boldsymbol{0}, \mathrm{I}_d) \tag{20}$$

where $f_\star : \mathbb{R} \to \mathcal{Y}$ is an activation function, which we assume can be potentially stochastic (as in the logit model studied in the main, Sec. 2.2). For convenience, define the matrix $\mathrm{X} \in \mathbb{R}^{n\times d}$ and the vector $\boldsymbol{y} \in \mathcal{Y}^n$ obtained by stacking together $\boldsymbol{x}^\mu$ and $\boldsymbol{y}^\mu$ row-wise. We are interested in studying the following generalized linear predictor:

$$\hat{y} = f(\hat{\boldsymbol{\theta}}^\top \boldsymbol{\varphi}(\boldsymbol{x})) \tag{21}$$

where $\boldsymbol{\varphi} : \mathbb{R}^d \to \mathbb{R}^p$ is a feature map, and $\hat{\boldsymbol{\theta}} \in \mathbb{R}^p$ are weights, which generally depend on the training data $\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}(\varphi(\mathrm{X}), \boldsymbol{\theta}_\star)$, where for convenience we defined the feature matrix $\varphi(\mathrm{X}) \in \mathbb{R}^{n\times p}$. The *random features model* correspond to the specific feature map:

$$\boldsymbol{\varphi}(\boldsymbol{x}) = \frac{1}{\sqrt{p}}\phi(\mathrm{F}\boldsymbol{x}) \tag{22}$$

where $\mathrm{F} \in \mathbb{R}^{p\times d}$ is a random matrix and $\phi : \mathbb{R} \to \mathbb{R}$ is a component-wise activation function. Our key goal is to characterize the statistics of the model, i.e. to compute expectations over functions of the test and training predictions:

$$\mathbb{E}_{\mathrm{X},\boldsymbol{x},\boldsymbol{\theta}_\star}\left[\psi\left(f_\star(\boldsymbol{\theta}_\star^\top\boldsymbol{x}), f(\hat{\boldsymbol{\theta}}(\varphi(\mathrm{X}),\boldsymbol{\theta}_\star)^\top\boldsymbol{\varphi}(\boldsymbol{x}))\right)\right], \qquad \mathbb{E}_{\mathrm{X},\boldsymbol{\theta}_\star}\left[\tilde{\psi}\left(f_\star(\mathrm{X}\boldsymbol{\theta}_\star), f(\boldsymbol{\varphi}(\mathrm{X})\hat{\boldsymbol{\theta}}(\varphi(\mathrm{X}),\boldsymbol{\theta}_\star)\right)\right] \tag{23}$$

where $\psi : \mathcal{Y}^2 \to \mathbb{R}$ and $\tilde{\psi} : \mathcal{Y}^{2n} \to \mathbb{R}$ are test functions. Note in particular that the generalization errors eqs. (6), (7) and the density eq. (8) are examples of the above.

Different tools from high-dimensional statistics have been designed to compute such expectations in the limit where $n, p, d \to \infty$ at fixed rates $\alpha = {}^n/_p$ and $\gamma = {}^d/_p$, both rigorously and heuristically, e.g. the replica method [Mézard et al., 1987, Zdeborová and Krzakala, 2016], CGMT [Stojnic, 2013, Thrampoulidis et al., 2015, Thrampoulidis et al., 2018], approximate message passing [Bayati and Montanari, 2011, Krzakala et al., 2012, Gerbelot and Berthier, 2021], cavity / leave-one-out method [Mezard and Montanari, 2009, Karoui et al., 2013, Mai et al., 2019], tools from random matrix theory [Karoui, 2009, Dobriban and Wager, 2018], among others. A shortcoming of all the aforementioned methods is that they typically rely on the Gaussianity of the input data, and therefore are not directly applicable to the random features model (note that even if $\mathrm{F}\boldsymbol{x} \in \mathbb{R}^p$ is a Gaussian vector, the features $\phi(\mathrm{F}\boldsymbol{x})$ are *not* Gaussian).

Gaussian equivalence provides a surprising answer to this hurdle. Assuming for simplicity that the features are centred $\mathbb{E}_{\boldsymbol{x}}[\boldsymbol{\varphi}(\boldsymbol{x})] = 0$ and defining the covariances:

$$\Phi = \mathbb{E}_{\boldsymbol{x}}[\boldsymbol{\varphi}(\boldsymbol{x})\boldsymbol{x}^\top] \in \mathbb{R}^{p\times d}, \qquad \Omega = \mathbb{E}_{\boldsymbol{x}}[\boldsymbol{\varphi}(\boldsymbol{x})\boldsymbol{\varphi}(\boldsymbol{x})^\top] \in \mathbb{R}^{p\times p} \tag{24}$$

Gaussian equivalence states that in the high-dimensional limit, the expectations in eq. (23) can be computed for an *equivalent Gaussian model* with matching second moments:

$$\begin{aligned}
\mathbb{E}_{\mathrm{X},\boldsymbol{x},\boldsymbol{\theta}_\star}\left[\psi\left(f_\star(\boldsymbol{\theta}_\star^\top\boldsymbol{x}), f(\hat{\boldsymbol{\theta}}(\varphi(\mathrm{X}),\boldsymbol{\theta}_\star)^\top\boldsymbol{\varphi}(\boldsymbol{x}))\right)\right] &\xrightarrow[p\to\infty]{} \mathbb{E}_{\mathrm{X},\mathrm{V},\boldsymbol{x},\boldsymbol{v},\boldsymbol{\theta}_\star}\left[\psi\left(f_\star(\boldsymbol{\theta}_\star^\top\boldsymbol{x}), f(\hat{\boldsymbol{\theta}}(\mathrm{V},\boldsymbol{\theta}_\star)^\top\boldsymbol{v})\right)\right] \\
\mathbb{E}_{\mathrm{X},\boldsymbol{\theta}_\star}\left[\tilde{\psi}\left(f_\star(\mathrm{X}\boldsymbol{\theta}_\star), f(\boldsymbol{\varphi}(\mathrm{X})\hat{\boldsymbol{\theta}}(\varphi(\mathrm{X}),\boldsymbol{\theta}_\star)\right)\right] &\xrightarrow[p\to\infty]{} \mathbb{E}_{\mathrm{X},\mathrm{V},\boldsymbol{\theta}_\star}\left[\tilde{\psi}\left(f_\star(\mathrm{X}\boldsymbol{\theta}_\star), f(\mathrm{V}\hat{\boldsymbol{\theta}}(\mathrm{V},\boldsymbol{\theta}_\star)\right)\right]
\end{aligned} \tag{25}$$

where $(\boldsymbol{x}^\mu, \boldsymbol{v}^\mu)_{\mu\in[n]}$ are n independent samples of jointly Gaussian random variables:

$$(\boldsymbol{x}, \boldsymbol{v}) \sim \mathcal{N}\left(\mathbf{0}_{d+p}, \begin{bmatrix} \mathrm{I}_d/d & \Phi/\sqrt{pd} \\ \Phi^\top/\sqrt{pd} & \Omega/p \end{bmatrix}\right) \tag{26}$$

and as before we defined the matrices $\mathrm{X} \in \mathbb{R}^{n\times d}$ and $\mathrm{V} \in \mathbb{R}^{n\times p}$ by stacking the samples row-wise. For the random features model $\boldsymbol{\varphi}(\boldsymbol{x}) = 1/\sqrt{p}\ \phi(\mathrm{F}\boldsymbol{x})$, the asymptotic covariances Φ, Ω can be computed explicitly, and are given by:

$$\Phi \underset{p\to\infty}{\asymp} \frac{\kappa_1}{\sqrt{p}}\mathrm{F}, \qquad\qquad \Omega \underset{p\to\infty}{\asymp} \kappa_0^2 1_p 1_p^\top + \frac{\kappa_1^2}{p}\mathrm{F}\mathrm{F}^\top + \kappa_\star^2 \mathrm{I}_p \tag{27}$$

where the constants $(\kappa_0, \kappa_1, \kappa_\star) \in \mathbb{R}^3$ are the Gaussian moments of the activation function ϕ:

$$\kappa_0 = \mathbb{E}_{z\sim\mathcal{N}(0,1)}[\phi(z)], \qquad \kappa_1 = \mathbb{E}_{z\sim\mathcal{N}(0,1)}[\phi'(z)], \qquad \kappa_\star = \sqrt{\mathbb{E}_{z\sim\mathcal{N}(0,1)}[\phi(z)^2] - \kappa_1^2 - \kappa_0^2}. \tag{28}$$

Therefore, for the random features problem the Gaussian equivalent model can be written explicitly in terms of the input data $\boldsymbol{x} \sim \mathcal{N}(\mathbf{0}, 1/d\mathrm{I}_d)$ and the weights $\mathrm{F} \in \mathbb{R}^{p\times d}$ as:

$$\boldsymbol{v} = \kappa_0 1_p + \frac{\kappa_1}{\sqrt{p}}\mathrm{F}\boldsymbol{x} + \kappa_\star \boldsymbol{z} \tag{29}$$

where $\boldsymbol{z} \sim \mathcal{N}(\mathbf{0}, \mathrm{I}_p)$ is an effective noise vector which is independent from F, $\boldsymbol{x}$ and $\boldsymbol{\theta}_\star$. In summary, in the high-dimensional limit the statistics of the random features model is equivalent to the statistics of a Gaussian equivalent model with noisy features. The later can be directly characterized by the methods mentioned in the last paragraph.

A.2 Gaussian equivalence theorem

Related literature: Gaussian universality has a long history, and appeared in many contexts ranging from random matrix theory [Erdos et al., 2009, Erdos et al., 2010, Tao and Vu, 2012] to signal processing [Donoho and Tanner, 2009], statistical learning [Abbasi et al., 2019, Panahi and Hassibi, 2017, Korada and Montanari, 2010] and physics [Carmona and Hu, 2004, Chatterjee, 2005]. In the context of random features, a precursor of the result discussed here is the observation that for Gaussian data the high-dimensional limit of kernel spectra is linearly related the spectrum of the inputs [Karoui, 2010]. This result was generalized to random features kernels in [Pennington and Worah, 2017, Liao and Couillet, 2018], and leveraged by [Mei and Montanari, 2022] to derive exact asymptotic expressions for the generalization and training error of random features regression. For ridge regression, computing the performance boils down to computing traces of the feature matrix, and therefore Gaussian universality can be seen as an instance of spectral universality of random matrices [Benigni and Péché, 2021]. In non-linear problems where a closed-form solution is not available (as in our classification setting), Gaussian universality for the random features model was shown to hold for the empirical risk minimizer in [Gerace et al., 2020, Goldt et al., 2022], and was later proven in [Hu and Lu, 2020, Dhifallah and Lu, 2020]. More recently, [Montanari and Saeed, 2022] extended this proof to non-convex loss, and showed the universality of the free energy density associated to the empirical risk at finite temperature as well. This version is better suited to our discussion, since it is closer to the Bayesian classifiers studied in the main.

Theorem A.1 (Lemma 1 from [Montanari and Saeed, 2022])**.** *Consider the random features model discussed in Sec. A.1. Assume that the activation ϕ is three times differentiable and has zero Gaussian mean $\kappa_0 = \mathbb{E}_{z\sim\mathcal{N}(0,1)}[\phi(z)] = 0$ and that the weight matrix $F \in \mathbb{R}^{p\times d}$ has rows $\boldsymbol{f}_i \sim \mathcal{N}(\mathbf{0}, I_d)$ for $i \in [p]$. Further, assume that the function $f_\star$ is Lipschitz with i.i.d. bounded sub-Gaussian noise. Define the free energy density at inverse temperature $\beta > 0$:*

$$f_\beta(\varphi(X)) = -\frac{1}{p\beta}\log\int_{\mathbb{R}^p} d\boldsymbol{\theta}\ \exp\left\{-\beta\left[-\sum_{\mu=1}^n \log\sigma\left(f_\star(\boldsymbol{\theta}_\star^\top\boldsymbol{x}^\mu) \times \boldsymbol{\theta}^\top\boldsymbol{\varphi}(\boldsymbol{x}^\mu)\right) + \frac{\lambda}{2}||\boldsymbol{\theta}||_2^2\right]\right\}. \tag{30}$$

Then for any bounded differentiable function ψ with Lipschitz derivative we have:

$$\lim_{p\to\infty} \left|\mathbb{E}\left[\psi\left(f_\beta(\varphi(X))\right)\right] - \mathbb{E}\left[\psi\left(f_\beta(V)\right)\right]\right| = 0 \tag{31}$$

We refer the reader to [Montanari and Saeed, 2022] for the technical details on the proof of this result.

A.3 Beyond random features

The Gaussian equivalence theorem for the random features model motivates the study of generalized linear models with general Gaussian covariates. For instance, consider n independently drawn Gaussian covariates $(\boldsymbol{u}^\mu, \boldsymbol{v}^\mu) \in \mathbb{R}^{d+p}$:

$$\begin{pmatrix}\boldsymbol{u}\\ \boldsymbol{v}\end{pmatrix} \sim \mathcal{N}\left(\mathbf{0}_{d+p}, \begin{bmatrix}\Psi & \Phi\\ \Phi^\top & \Omega\end{bmatrix}\right) \tag{32}$$

for positive definite matrices $\Psi \in \mathbb{R}^{d\times d}$, $\Omega \in \mathbb{R}^{p\times p}$ and $\Phi \in \mathbb{R}^{p\times d}$ such that $\Psi - \Phi\Omega^{-1}\Phi^\top$ is invertible. Labels $y^\mu \in \mathcal{Y}$ are generated from the covariate $\boldsymbol{u} \in \mathbb{R}^p$ from a generalized linear model:

$$y^\mu = f_\star(\boldsymbol{\theta}_\star^\top \boldsymbol{u}^\mu/\sqrt{d}), \qquad\qquad \boldsymbol{\theta}_\star \sim \mathcal{N}(0, \mathrm{I}_d). \tag{33}$$

However, the statistician only observes the pairs $(\boldsymbol{v}^\mu, y^\mu) \in \mathbb{R}^p \times \mathcal{Y}$, from which she tries to learn:

$$\hat{y} = f(\hat{\boldsymbol{\theta}}^\top \boldsymbol{v}/\sqrt{p}). \tag{34}$$

The asymptotic statistics of this Gaussian covariate model has been derived and proven in the recent [Loureiro et al., 2021a] for the particular case in which $\hat{\boldsymbol{\theta}}$ is the empirical risk minimizer. In Appendix B, we recover and generalize this result to the other estimators defined in Sec. 2.1.

Note that thanks to Gaussian equivalence, in the proportional high-dimensional limit, the random features model discussed in Sec. A.1 is a particular case of this Gaussian covariate model where $\boldsymbol{u} = \boldsymbol{x}$ and $\boldsymbol{v} = \boldsymbol{\varphi}(\boldsymbol{x})$. However, the Gaussian covariate model can accommodate a richer class of models. For instance, one could consider the case in which the target covariates themselves come from a feature map: $\boldsymbol{u} = \boldsymbol{\varphi}_\star(\boldsymbol{x})$. Although Gaussian equivalence has only been established for a limited number of feature maps, [Loureiro et al., 2021a] has empirically observed that the asymptotic formulas derived for Gaussian covariates are in good agreement with a rich class of feature maps, including case in which the fixed features are learned from neural networks. While the goal of this work *is not* to investigate Gaussian equivalence, this line of work motivate us to derive our result for general Gaussian covariates, hence making them readily applicable to equivalences proven in the future.

B Derivation of Theorem 3.1

In this Appendix we provide a derivation of the self-consistent equations (14) characterizing the sufficient statistics $(m_t^\star, q_t^\star, v_t^\star, \hat{m}_t^\star, \hat{q}_t^\star, \hat{v}_t^\star)$ for $t \in \{\text{bo}, \text{erm}, \text{eb}, \text{Lap}\}$. As motivated in Appendix A, our discussion will focus on the more general Gaussian covariate model, which contains the random features setting as a particular case. The key idea is to design an approximate message passing and show that the associated state evolution equations coincide exactly the with the self-consistent equations for the sufficient statistics in Theorem 3.1.

This Appendix is organized as follows. We start by a recap of the Gaussian covariate model for our specific setting in Sec. B.1, and introduce a convenient change of variables. Next, in Sec. B.2 we introduced a tailored message passing algorithm, and provide an informal derivation of the associated state evolution equations. In Sec. B.3 we provide a heuristic derivation of the self-consistent equations from the replica method, and show that it agrees with the state evolution equations for our algorithm. In Sec. B.4 we discuss how these equations are made rigorous from recent progress in the literature. Finally, in the three last subsections we discuss variations of this result to the context of the Laplace approximation and temperature scaling, and a simplification for the the random features case.

B.1 Recap of the setting

As motivated in Sec. A.3, our goal is to derive the self-consistent equations in the more general setting of the Gaussian covariate model (GCM), which thanks to Gaussian universality contains the random features model as a special case. For the ease of reading, we first recall the reader of the model of interest, which specializes Sec. A.3 to binary classification.

Data model: Let $(\boldsymbol{u}, \boldsymbol{v})$ denote a pair of Gaussian covariates:

$$\begin{pmatrix} \boldsymbol{u} \\ \boldsymbol{v} \end{pmatrix} \sim \mathcal{N}\left(\mathbf{0}_{d+p}, \begin{bmatrix} \Psi & \Phi \\ \Phi^\top & \Omega \end{bmatrix}\right). \tag{35}$$

and define the oracle classifier as:

$$f_\star(\boldsymbol{u}) = \mathbb{P}(y=1|\boldsymbol{\theta}_\star^\top \boldsymbol{u}) = \sigma_{\tau_0^2}\left(\frac{\boldsymbol{\theta}_\star^\top \boldsymbol{u}}{\sqrt{d}}\right), \qquad \boldsymbol{\theta}_\star \sim \mathcal{N}(\mathbf{0}, \mathrm{I}_d), \tag{36}$$

where we remind the reader of the convenient notation:

$$\sigma_\tau(x) := \int \sigma(z)\mathcal{N}(z|x,\tau)\mathrm{d}z \tag{37}$$

with $\sigma(z) = (1+e^{-z})^{-1}$ the sigmoid function.

Classifiers: Given n independent pairs $(\boldsymbol{v}^\mu, y^\nu)_{\mu\in[n]} \in \mathbb{R}^p \times \{-1,1\}$ from the model above and defining the training data $\mathcal{D} = \{(\boldsymbol{v}^\mu, y^\mu)_{\mu\in[n]}\}$ we are interested in studying the family of probabilistic classifiers of the type:

$$\hat{f}_t(\boldsymbol{v}) = \mathbb{P}(y=1|\tau_t, \mathcal{D}) = \int \mathrm{d}\boldsymbol{\theta}\ \sigma_{\tau_t}\left(\frac{\boldsymbol{\theta}^\top \boldsymbol{v}}{\sqrt{p}}\right) p_t(\boldsymbol{\theta}|\mathcal{D}) \tag{38}$$

where the "posterior" $p_t(\boldsymbol{\theta}|\mathcal{D})$ and the noise level τ_t depend on the specific classifier $t \in \{\text{bo}, \text{erm}, \text{eb}, \text{Lap}\}$ of interest introduced in Sec. 2.1.

A convenient rewriting: Since the covariate $\boldsymbol{u} \in \mathbb{R}^d$ is not observed by the statistician, it is useful to rewrite it explicitly as a function of $\boldsymbol{v}$ and an effective uncorrelated noise. Additionally, it is also convenient to write $\boldsymbol{v}$ in terms of an uncorrelated variable. Mathematically, this is given by a standard Gaussian partition:

$$\boldsymbol{u} = \Phi\Omega^{-1}\boldsymbol{v} + \left(\Psi - \Phi\Omega^{-1}\Phi^\top\right)^{1/2}\boldsymbol{z}, \tag{39}$$

for $\boldsymbol{z} \sim \mathcal{N}(0, I_d)$ uncorrelated with $\boldsymbol{v}$. This motivate us to define the *projected oracle weights*:

$$\boldsymbol{w}_\star = \Omega^{-1}\Phi^\top \boldsymbol{\theta}_\star \tag{40}$$

Then, the oracle classifier can be equivalently written as:

$$\mathbb{P}(y=1|\boldsymbol{w}_\star^\top \boldsymbol{v}) = \int \sigma_{\tau_0^2}\left(\frac{\boldsymbol{w}_\star^\top \boldsymbol{v}}{\sqrt{d}} + \frac{1}{\sqrt{d}}\boldsymbol{\theta}_\star^\top \left(\Psi - \Phi\Omega^{-1}\Phi^\top\right)^{1/2} \boldsymbol{z}\right) \mathcal{N}\left(\boldsymbol{z}|0, \mathrm{I}_d\right) \mathrm{d}\boldsymbol{z} \tag{41}$$

$$= \int \sigma_{\tau_0^2}\left(\frac{\boldsymbol{w}_\star^\top \boldsymbol{v}}{\sqrt{d}} + \xi\right) \mathcal{N}\left(\xi \Big| 0, \frac{\boldsymbol{\theta}_\star^\top \left(\Psi - \Phi\Omega^{-1}\Phi^\top\right)\boldsymbol{\theta}_\star}{d}\right) \mathrm{d}\xi \tag{42}$$

which is a logit model on the observed features with an effective mismatch noise

$$\xi \sim \mathcal{N}(0, \frac{1}{d}\boldsymbol{\theta}_\star^\top \left(\Psi - \Phi\Omega^{-1}\Phi^\top\right)\boldsymbol{\theta}_\star).$$

Recalling that $\boldsymbol{\theta}_\star \sim \mathcal{N}(\mathbf{0}_d, \mathrm{I}_d)$, in the asymptotic limit the noise variance concentrates:

$$\frac{1}{d}\boldsymbol{\theta}_\star^\top \left(\Psi - \Phi\Omega^{-1}\Phi^\top\right)\boldsymbol{\theta}_\star \to \frac{1}{d}\operatorname{Tr}\left(\Psi - \Phi\Omega^{-1}\Phi^\top\right) =: \tau^2_{\text{add}}. \tag{43}$$

Therefore, the oracle classifier is equivalent to:

$$f_\star(\boldsymbol{v}) = \mathbb{P}(y=1|\boldsymbol{w}_\star^\top \boldsymbol{v}) = \sigma_{\tau_0^2+\tau^2_{\text{add}}}\left(\frac{\boldsymbol{w}_\star^\top \boldsymbol{v}}{\sqrt{d}}\right) \tag{44}$$

with

$$\boldsymbol{w}_\star \sim \mathcal{N}(0, \Sigma_\star), \quad \Sigma_\star = \Omega^{-1}\Phi^\top\Phi\Omega^{-1} \tag{45}$$

Finally, to further simplify the algebra it is convenient to consider the following change of variables:

$$\boldsymbol{v} \to \Omega^{-1/2}\boldsymbol{v}, \quad \boldsymbol{w}_\star \to \Omega^{1/2}\boldsymbol{w}_\star \tag{46}$$

Such that $\boldsymbol{v} \sim \mathcal{N}(\mathbf{0}_p, \mathrm{I}_p)$ and $\boldsymbol{w}_\star \sim \mathcal{N}(\mathbf{0}_d, \tilde{\Phi}^\top\tilde{\Phi})$ with $\tilde{\Phi} \equiv \Phi\Omega^{-1/2}$. Note that the labels are invariant under this change, and therefore we can assume input data with identity covariance.

B.2 State evolution for GAMP

Algorithm 1: GAMP for an estimator $t \in \{\text{erm}, \text{bo}, \text{eb}\}$

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

With the model in hands, we now show discuss how the sufficient statistics $(v_t^\star, q_t^\star, m_t^\star)$ needed to characterize the asymptotic density defined in eq. (12) satisfy a set of self-consistent equations, which in the particular case of the random features model are explicitly written given in eq. (14).

Our derivation follows from the analysis of an approximate message passing scheme, which provides a powerful tool to derive exact asymptotic results in an unified way, and has been employed in many works in the high-dimensional statistics literature, for instance

Classifier	$f_{out,t}(y,\omega,v)$	$f_{w,t}(\boldsymbol{b},\mathrm{A})$
$\hat{f}_{\rm erm}$	$\mathrm{prox}_{\log\sigma(y\times\cdot)}(\omega)$	$(\lambda \mathrm{I}_p + \mathrm{A})\boldsymbol{b}$
$\hat{f}_{\rm bo}$	$\partial_\omega \log \int \mathbb{P}(y=1\|z)\mathcal{N}(z\|\omega,v)\mathrm{d}z$	$(\Sigma_\star^{-1} + \mathrm{A})\boldsymbol{b}$
$\hat{f}_{\rm eb}$	$\partial_\omega \log \int \sigma(\beta y \times z)\mathcal{N}(z\|\omega,v)\mathrm{d}z$	$(\lambda \mathrm{I}_p + \mathrm{A})\boldsymbol{b}$

Table 2: GAMP denoising functions for the ERM, Bayes-optimal and empirical Bayes estimators. We recall that the covariance matrix is given by $\Sigma_\star = \Omega^{-1}\Phi^\top\Phi\Omega^{-1}$.

see [Bayati and Montanari, 2012, Bayati and Montanari, 2011, Rangan, 2011, Krzakala et al., 2012, Donoho and Montanari, 2016, Sur and Candès, 2019, Loureiro et al., 2021b, Celentano et al., 2020, Gerbelot and Berthier, 2021, Loureiro et al., 2022, Cornacchia et al., 2022].

Given the training data $\mathcal{D} = (\mathrm{V}, \boldsymbol{y})$, the initial step is to consider the following set of iterates known as Generalized Approximate Message Passing (GAMP) algorithm 1, where the *denoising functions* $(f_{\mathrm{out},t}, f_{w,t})$ depend on the specific classifier of interest $t \in \{\mathrm{bo}, \mathrm{erm}, \mathrm{eb}\}$, and are summarized in table 2. The convenience of the GAMP is precisely to allow us to deal with classifiers of very different nature ($t \in \{\mathrm{bo}, \mathrm{eb}\}$ are defined by sampling, while $t = \mathrm{erm}$ is a point-estimator) in an unified framework. Note that the GAMP algorithm 1 is close to the one in [Rangan, 2011], with the important difference that the denoising functions $f_{w,t}$ is vector valued - a consequence of the fact that implicitly the classifiers of interest have non-separable priors. A second convenient property of GAMP is that in the high-dimensional limit of interest here, the statistics of the sequence of estimators $\hat{\boldsymbol{\theta}}_t^{T,\mathrm{amp}}, \hat{\boldsymbol{c}}_t^{T,\mathrm{amp}}$ can be exactly tracked by a set of equations known as *state evolution*. Therefore, the key idea in the proof strategy is to show that the statistics of the iterates $\hat{\boldsymbol{\theta}}_t^{T,\mathrm{amp}}, \hat{\boldsymbol{c}}_t^{T,\mathrm{amp}}$ (given by the state evolution equations) coincide with the statistics of the classifiers defined in Sec. 2.1. The state evolution for GAMP with non-separable priors was rigorously derived in [Berthier et al., 2019, Gerbelot and Berthier, 2021]. Therefore, in the following we limit ourselves to an informal but intuitive derivation. In Sec. B.3 we show that the state evolution for the GAMP estimators indeed coincides with the fixed-point equations describing the statistics of the classifiers of interest according to the replica method. The fact that GAMP (rigorous) state evolution equations corresponds to the replica saddle-point equations is a very general fact [Zdeborová and Krzakala, 2016], which is at the roots of many rigorous proofs to the replica predictions.

In the limit where $n, p \to \infty$ with fixed $\alpha = {}^n\!/_p$, it can be shown that the GAMP algorithm 1 is asymptotically equivalent to the following rBP equations (this is discussed in for instance [Aubin et al., 2018, Aubin et al., 2021]):

$$\begin{cases} \omega^T_{\mu\to i} = \sum_{j\neq i} v_j^\mu \hat{\theta}^T_{j\to\mu} \\ V^T_{\mu\to i} = \sum_{j\neq i} (v_j^\mu)^2 \hat{c}^T_{j\to\mu} \end{cases}, \qquad \begin{cases} g^T_{\mu\to i} = f_{\mathrm{out},t}(y^\mu, \omega^T_{\mu\to i}, V^T_{\mu\to i}) \\ \partial g^T_{\mu\to i} = \partial_\omega f_{\mathrm{out},t}(y^\mu, \omega^T_{\mu\to i}, V^T_{\mu\to i}) \end{cases} \tag{47}$$

$$\begin{cases} b^T_{\mu\to i} = \sum_{\nu\neq\mu} v_i^\nu g^T_{\nu\to i} \\ A^T_{\mu\to i} = -\sum_{\nu\neq\mu} (v_i^\nu)^2 \partial g^T_{\nu\to i} \end{cases}, \qquad \begin{cases} \hat{\theta}^{T+1}_{i\to\mu} = f_{w,t}(b^T_{i\to\mu}, A^T_{i\to\mu}) \\ \hat{c}^{T+1}_{i\to\mu} = \partial_b f_{w,t}(b^T_{\mu\to i}, A^T_{\mu\to i}) \end{cases} \tag{48}$$

where we recall the reader $i \in [p]$, $\mu \in [n]$, and to lighten notation we have dropped the indexes $t \in \{\mathrm{bo}, \mathrm{erm}, \mathrm{eb}\}$ for the classifier and $^{\mathrm{amp}}$ which stresses that the messages concern GAMP estimators. By construction, the rBP messages are independent, and are only coupled to each other through the data, which we recall is given by:

$$y^\mu \sim P_0(\cdot|\boldsymbol{w}_\star^\top \boldsymbol{v}^\mu), \qquad \boldsymbol{v}^\mu \sim \mathcal{N}(0, \mathrm{I}_p), \qquad \boldsymbol{w}_\star \sim \mathcal{N}(\boldsymbol{0}, \Sigma_\star) \tag{49}$$

For convenience, we define the so-called *teacher local field*:

$$z_\mu = \boldsymbol{w}_\star^\top \boldsymbol{v}^\mu / \sqrt{d} \tag{50}$$

Without loss of generality, we can write $y^\mu = f_0(z_\mu, \eta^\mu)$ for $\eta^\mu \sim \mathcal{N}(0,1)$. We now characterize the joint statistics of the rBP messages.

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

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

$$\mathbb{E}\left[z^2_\mu\right] = \frac{1}{d}\mathbb{E}\left[\sum_{i=1}^p\sum_{j=1}^p v^\mu_i v^\mu_j w_{\star i} w_{\star j}\right] = \frac{1}{d}\sum_{i=1}^p\sum_{j=1}^p \mathbb{E}\left[v^\mu_i v^\mu_j\right] w_{\star i} w_{\star j} = \frac{1}{d}\sum_{i=1}^p\sum_{j=1}^p \delta_{ij} w_{\star i} w_{\star j}$$
$$\underset{p\to\infty}{\to} \rho \tag{51}$$

$$\mathbb{E}\left[\left(\omega^T_{\mu\to i}\right)^2\right] = \frac{1}{p}\mathbb{E}\left[\sum_{j\neq i}^p\sum_{k\neq i}^p v^\mu_j v^\mu_k \hat\theta^T_{j\to\mu}\hat\theta^T_{k\to\mu}\right] = \frac{1}{p}\sum_{j\neq i}^p\sum_{k\neq i}^p \mathbb{E}\left[v^\mu_j v^\mu_k\right]\hat\theta^T_{j\to\mu}\hat\theta^T_{k\to\mu}$$
$$= \frac{1}{p}\sum_{j\neq i}^p\sum_{k\neq i}^p \delta_{jk}\hat\theta^T_{j\to\mu}\hat\theta^T_{k\to\mu} \underset{p\to\infty}{\to} q^T \tag{52}$$

$$\mathbb{E}\left[z_\mu \omega^T_{\mu\to i}\right] = \frac{1}{\sqrt{dp}}\mathbb{E}\left[\sum_{j\neq i}^p\sum_{k=1}^p v^\mu_j v^\mu_k \hat\theta^T_{j\to\mu} w_{\star k}\right] = \frac{1}{\sqrt{dp}}\sum_{j\neq i}^p\sum_{k=1}^p \mathbb{E}\left[v^\mu_j v^\mu_k\right]\hat\theta^T_{j\to\mu} w_{\star k}$$
$$= \frac{1}{\sqrt{dp}}\sum_{j\neq i}^p\sum_{k=1}^p \delta_{jk}\hat\theta^T_{j\to\mu} w_{\star k} \underset{p\to\infty}{\to} m^T \tag{53}$$

$$\tag{54}$$

where we have used that $\hat\theta^T_{i\to\mu} = O(p^{-1/2})$ to simplify the sums at large p. Summarising our findings:

$$(z_\mu, \omega^T_{\mu\to i}) \sim \mathcal{N}\left(\mathbf{0}_3, \begin{bmatrix}\rho & m^T\\ m^T & q^T\end{bmatrix}\right) \tag{55}$$

with:

$$\rho \equiv \frac{1}{d}\boldsymbol{w}_\star^\top \boldsymbol{w}_\star, \qquad q^T \equiv \frac{1}{p}(\hat{\boldsymbol{\theta}}^T_t)^\top \hat{\boldsymbol{\theta}}^T_t, \qquad m^T \equiv \frac{1}{\sqrt{dp}}(\hat{\boldsymbol{\theta}}^T_t)^\top \boldsymbol{w}_\star \tag{56}$$

Step 2: Concentration of variances $V^T_{\mu\to i}$

Since the variance $V^T_{\mu\to i}$ depends on $(\boldsymbol{v}^\mu_i)^2$, in the asymptotic limit $p\to\infty$ it concentrates around its mean :

$$\mathbb{E}\left[V^T_{\mu\to i}\right] = \frac{1}{p}\sum_{j\neq i}\mathbb{E}\left[(v^\mu_i)^2\right]\hat c^T_{j\to\mu} = \frac{1}{p}\sum_{j\neq i}\hat c^T_{j\to\mu} = \frac{1}{p}\sum_{j=1}^p \hat c^T_{j\to\mu} - \frac{1}{p}\hat c^T_{i\to\mu} \underset{p\to\infty}{\to} V^T \equiv \frac{1}{p}\sum_{j=1}^p \hat c^T_j \tag{57}$$

where we have defined the variance overlap V^T. We thus have $V^T_{\mu\to i}\to V^T$. Note that V^T corresponds to the divergence with respect to $\boldsymbol{b}$ of $\log \mathcal{Z}_w(\boldsymbol{b}, \mathrm{A})$.

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

By definition, we have

$$b^T_{\mu\to i} = \frac{1}{\sqrt{p}}\sum_{\nu\neq\mu} v^\nu_i g^T_{\nu\to i} = \frac{1}{\sqrt{p}}\sum_{\nu\neq\mu} v^\nu_i f_{\rm out}(y^\mu, \omega^T_{\nu\to i}, V^T_{\nu\to i}) = \frac{1}{\sqrt{p}}\sum_{\nu\neq\mu} v^\nu_i f_{\rm out}(f_0(z_\nu,\eta^\nu), \omega^T_{\nu\to i}, V^T_{\nu\to i})$$

$$\tag{58}$$

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

$$z_\mu = \frac{1}{\sqrt{d}} \sum_{j=1}^{p} v_j^\mu w_{\star j} = \underbrace{\frac{1}{\sqrt{d}} \sum_{j \neq i} v_j^\mu w_{\star j}}_{z_{\mu \to i}} + \frac{1}{\sqrt{d}} v_i^\mu w_{\star i} \tag{59}$$

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

$$\begin{aligned} f_{\text{out}}(f_0(z_\mu, \eta^\nu), \omega^T_{\nu \to i}, V^T_{\nu \to i}) &= f_{\text{out}}(f_0(z_{\nu \to i}, \eta^\nu), \omega^T_{\nu \to i}, V^T_{\nu \to i}) \\ &\quad + \frac{1}{\sqrt{d}} \partial_z f_{\text{out}}(f_0(z_{\nu \to i}, \eta^\nu), \omega^T_{\nu \to i}, V^T_{\nu \to i}) v_i^\nu w_{\star i} + O(p^{-1}) \end{aligned} \tag{60}$$

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

$$\begin{aligned} b^T_{\mu \to i} &= \frac{1}{\sqrt{p}} \sum_{\nu \neq \mu} v_i^\nu f_{\text{out}}(f_0(z_{\nu \to i}, \eta^\nu), \omega^T_{\nu \to i}, V^T_{\nu \to i}) \\ &\quad + \frac{1}{\sqrt{dp}} \sum_{\nu \neq \mu} (v_i^\nu)^2 \partial_z f_{\text{out}}(f_0(z_{\nu \to i}, \eta^\nu), \omega^T_{\nu \to i}, V^T_{\nu \to i}) w_{\star i} + O(p^{-3/2}) \end{aligned} \tag{61}$$

Therefore:

$$\begin{aligned} \mathbb{E}\left[b^T_{\mu \to i}\right] &= \frac{w_{\star i}}{\sqrt{dp}} \sum_{\nu \neq \mu} \partial_z f_{\text{out}}(f_0(z_{\nu \to i}, \eta^\nu), \omega^T_{\nu \to i}, V^T_{\nu \to i}) + O(p^{-3/2}) \\ &= \frac{w_{\star i}}{\sqrt{dp}} \sum_{\nu = 1}^{n} \partial_z f_{\text{out}}(f_0(z_{\nu \to i}, \eta^\nu), \omega^T_{\nu \to i}, V^T_{\nu \to i}) + O(p^{-3/2}) \end{aligned} \tag{62}$$

Note that as $p \to \infty$, 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 eq. (55). Therefore,

$$\frac{1}{\sqrt{dp}} \sum_{\nu = 1}^{n} \partial_z f_{\text{out}}(f_0(z_{\nu \to i}, \eta^\nu), \omega^T_{\nu \to i}, V^T_{\nu \to i}) \underset{p \to \infty}{\to} \alpha \sqrt{\gamma}\, \mathbb{E}_{(\omega, z), \eta}\left[\partial_z f_{\text{out}}(f_0(z, \eta), \omega, V^T)\right] \equiv \hat{m}^T \tag{63}$$

so:

$$\mathbb{E}\left[b^T_{\mu \to i}\right] \underset{p \to \infty}{\to} w_{\star i} \hat{m}^T. \tag{64}$$

Similarly, the variance is given by:

$$\begin{aligned} &\text{Var}\left[b^T_{\mu \to i}\right] \\ &= \frac{1}{p} \sum_{\nu \neq \mu} \sum_{\kappa \neq \mu} \mathbb{E}\left[v_i^\nu v_i^\kappa\right] f_{\text{out}}(f_0(z_{\nu \to i}, \eta^\nu), \omega^T_{\nu \to i}, V^T_{\nu \to i}) f_{\text{out}}(f_0(z_{\kappa \to i}, \eta^\kappa), \omega^T_{\kappa \to i}, V^T_{\kappa \to i}) + O(d^{-2}) \\ &= \frac{1}{p} \sum_{\nu \neq \mu} f_{\text{out}}(f_0(z_{\nu \to i}, \eta^\nu), \omega^T_{\nu \to i}, V^T_{\nu \to i})^2 + O(p^{-2}) \\ &= \frac{1}{p} \sum_{\nu = 1}^{n} f_{\text{out}}(f_0(z_{\nu \to i}, \eta^\nu), \omega^T_{\nu \to i}, V^T_{\nu \to i})^2 + O(p^{-2}) \\ &\underset{d \to \infty}{\to} \alpha\, \mathbb{E}_{(z, \omega), \xi}\left[f_{\text{out}}(f_0(z, \eta), \omega, V^T)^2\right] \equiv \hat{q}^T \end{aligned} \tag{65, 66}$$

To summarise, we have:

$$b^T_{\mu \to i} \sim \mathcal{N}\left(w_{\star i} \hat{m}^T, \hat{q}^T\right) \tag{67}$$

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:

$$A^T_{\mu\to i} = -\frac{1}{p}\sum_{\nu\neq\mu}(v^\nu_i)^2\partial_\omega f_{\text{out},t}(y^\nu, \omega^T_{\nu\to i}, V^T_{\nu\to i}) \tag{68}$$

$$= -\frac{1}{p}\sum_{\nu\neq\mu}(x^\nu_i)^2\partial_\omega f_{\text{out},t}(f_0(z_{\nu\to i}, \eta^\nu), \omega^T_{\nu\to i}, V^T_{\nu\to i}) + O(p^{-3/2})$$

$$= -\frac{1}{p}\sum_{\nu=1}\partial_\omega f_{\text{out},t}(f_0(z_{\nu\to i}, \eta^\nu), \omega^T_{\nu\to i}, V^T_{\nu\to i}) + O(d^{-3/2})$$

$$\underset{p\to\infty}{\to} -\alpha\,\mathbb{E}_{(z,\omega),\xi}\left[\partial_\omega f_{\text{out},t}(f_0(z,\eta),\omega,V^T)\right] \equiv \hat{v}^T \tag{69}$$

Summary

We now have all the ingredients we need to characterize the asymptotic distribution of the GAMP iterates for any of the classifiers $t \in \{\text{bo}, \text{erm}, \text{eb}\}$:

$$\hat{\boldsymbol{\theta}}^{T,\text{amp}}_t \sim f_{\text{out},t}(\boldsymbol{w}_\star \hat{m}^{T,\text{amp}}_t + \sqrt{\hat{q}^{T,\text{amp}}_t}\boldsymbol{\xi}, \hat{v}^{T,\text{amp}}_t) \tag{70}$$

where $\boldsymbol{\xi} \sim \mathcal{N}(\mathbf{0}, \mathrm{I}_p)$ is an independent Gaussian vector. Therefore, we recover the GAMP state evolution equations [Rangan, 2011, Berthier et al., 2019] for the overlaps:

$$\begin{cases} V^{T+1,\text{amp}} &= \mathbb{E}_{(\boldsymbol{w}_\star,\boldsymbol{\xi})}\left[\partial_{\boldsymbol{b}}\cdot f_{w,t}(\hat{m}^{T,\text{amp}}\boldsymbol{w}_\star + \sqrt{\hat{q}^{T,\text{amp}}}\boldsymbol{\xi}, \hat{v}^{T,\text{amp}}\mathrm{I}_p)\right] \\ q^{T+1,\text{amp}} &= \mathbb{E}_{(\boldsymbol{w}_\star,\boldsymbol{\xi})}\left[f_{w,t}(\hat{m}^{T,\text{amp}}\boldsymbol{w}_\star + \sqrt{\hat{q}^{T,\text{amp}}}\boldsymbol{\xi}, \hat{v}^T\mathrm{I}_p)^2\right] \\ m^{T+1,\text{amp}} &= \sqrt{\gamma}\mathbb{E}_{(\boldsymbol{w}_\star,\boldsymbol{\xi})}\left[f_{w,t}(\hat{m}^{T,\text{amp}}\boldsymbol{w}_\star + \sqrt{\hat{q}^{T,\text{amp}}}\boldsymbol{\xi}, \hat{v}^{T,\text{amp}}\mathrm{I}_p)^\top\boldsymbol{w}_\star\right] \end{cases}, \tag{71}$$

$$\begin{cases} \hat{v}^{T,\text{amp}} &= -\alpha\mathbb{E}_{(z,\omega),\eta}\left[\partial_\omega f_{\text{out},t}(f_0(z,\eta),\omega,V^{T,\text{amp}})\right] \\ \hat{q}^{T,\text{amp}} &= \alpha\mathbb{E}_{(z,\omega),\eta}\left[f_{\text{out},t}(f_0(z,\eta),\omega,V^{T,\text{amp}})^2\right] \\ \hat{m}^{T,\text{amp}} &= \alpha\sqrt{\gamma}\mathbb{E}_{(z,\omega),\eta}\left[\partial_z f_{\text{out},t}(f_0(z,\eta),\omega,V^{T,\text{amp}})\right] \end{cases} \tag{72}$$

where $\boldsymbol{w}_\star \sim \mathcal{N}(\mathbf{0}, \Sigma_\star), \xi \sim \mathcal{N}(\mathbf{0}, I_p)$ and $(z,\omega) \sim \mathcal{N}(\mathbf{0}, \begin{bmatrix}\rho & m^T \\ m^T & q^T\end{bmatrix}), \eta \sim \mathcal{N}(0,1)$

Interestingly, we can show that the equations (72) are strictly equivalent to the self-consistent equations (14) of Theorem 3.1. Consider first the update equations for $V^{T,\text{amp}}_t, q^{T,\text{amp}}_t, m^{T,\text{amp}}_t$. First, note that for all the estimators considered here, the function $f_{w,t}$ has the form $(\Sigma_t^{-1} + \mathrm{A})\boldsymbol{b}$ for some matrix Σ_t. Now, let us introduce

$$\tilde{\Psi}(\boldsymbol{b}, \mathrm{A}, \Sigma) = \frac{1}{2p}\operatorname{Tr}\log(\Sigma^{-1} + \mathrm{A}) + \frac{1}{2p}\boldsymbol{b}^\top(\Sigma^{-1} + \mathrm{A})^{-1}\boldsymbol{b} \tag{73}$$

$$\Psi_w(\hat{m}, \hat{q}, \hat{v}) = \mathbb{E}_{\boldsymbol{w}_\star,\boldsymbol{\xi}}\tilde{\Psi}(\hat{m}\boldsymbol{w}_\star + \hat{q}\boldsymbol{\xi}, \hat{v}\mathrm{I}_p) \tag{74}$$

With some algebra, we can see that for any estimator described in Table 2 we have

$$\partial_{\hat{m}}\Psi_w(\hat{m},\hat{q},\hat{v}) = \mathbb{E}_{(\boldsymbol{w}_\star,\boldsymbol{\xi})}\left[f_{w,t}(\hat{m}^T\boldsymbol{w}_\star + \sqrt{\hat{q}^T}\boldsymbol{\xi}, \hat{v}^T\mathrm{I}_p)^\top\boldsymbol{w}_\star\right] \tag{75}$$

$$\partial_{\hat{q}}\Psi_w - \partial_{\hat{v}}\Psi_w = \frac{1}{2}\mathbb{E}_{(\boldsymbol{w}_\star,\boldsymbol{\xi})}\left[f_{w,t}(\hat{m}^T\boldsymbol{w}_\star + \sqrt{\hat{q}^T}\boldsymbol{\xi}, \hat{v}^T\mathrm{I}_p)^2\right] \tag{76}$$

$$\partial_{\hat{q}}\Psi_w = \frac{1}{2}\mathbb{E}_{(\boldsymbol{w}_\star,\xi)}\left[\partial_{\boldsymbol{b}}\cdot f_{w,t}(\hat{m}^T\boldsymbol{w}_\star + \sqrt{\hat{q}^T}\boldsymbol{\xi}, \hat{v}^T\mathrm{I}_p)\right] \tag{77}$$

Thus the update equations (72) for m, q, v are equivalent to Equations (14). It is the same for $\hat{m}, \hat{q}, \hat{v}$: consider the update equation for $\hat{q}^{T,\text{amp}}$. We can rewrite it with a Dirac delta

$$\hat{q}^{T,\text{amp}} = \alpha \sum_y \mathbb{E}_{(z,\omega),\eta} \left[f_{\text{out},t}(y, \omega, V^{T,\text{amp}})^2 \delta(y - f_0(z, \eta)) \right] \tag{78}$$

$$= \alpha \sum_y \mathbb{E}_{\omega} \left[f_{\text{out},t}(y, \omega, V^{T,\text{amp}})^2 \mathbb{E}_{z|\omega,\eta}(\delta(y - f_0(z, \eta))) \right] \tag{79}$$

$$\tag{80}$$

The distribution of z conditioned on ω is a Gaussian with mean $m^{T,\text{amp}}/q^{T,\text{amp}} \times \omega$ and variance $\rho - m^{T,\text{amp}} \times m^{T,\text{amp}}/q^{T,\text{amp}}$. Then, $\mathbb{E}_{z|\omega,\eta}(\delta(y - f_0(z, \eta))$ can be written

$$\begin{aligned}
\mathbb{E}_{z|\omega,\eta}(\delta(y - f_0(z, \eta)) &= \int \mathrm{d}z \mathbb{E}_{\eta} \left(\delta(y - f_0(z, \eta)) \right) \mathcal{N}(z | m^{T,\text{amp}}/q^{T,\text{amp}} \times \omega, \rho - m^{T,\text{amp}} \times m^{T,\text{amp}}/q^{T,\text{amp}}) \\
&= \int \mathrm{d}z \mathbb{P}(y = 1|z) \mathcal{N}(z | m^{T,\text{amp}}/q^{T,\text{amp}} \times \omega, \rho - m^{T,\text{amp}} \times m^{T,\text{amp}}/q^{T,\text{amp}}) \\
&= \mathcal{Z}_0(y, m^{T,\text{amp}}/q^{T,\text{amp}} \times \omega, \rho - m^{T,\text{amp}} \times m^{T,\text{amp}}/q^{T,\text{amp}})
\end{aligned}$$

we thus recover the equation for $\hat{q}$ in (14) :

$$\hat{q}^{T+1} = \alpha \sum_y \mathbb{E}_{\omega} \left[f_{\text{out},t}(y, \omega, V^{T,\text{amp}})^2 \mathcal{Z}_0(y, m^T/q^T \omega, \rho - {}^{m^2}/_{q}) \right] \tag{81}$$

Similar computations can be done for $\hat{m}$ and $\hat{v}$.

We have thus eqs. (72) with the self-consistent equations (14) of Theorem 3.1. It remains to show two points. First, that in the particular case of the random feature model the expression of Ψ_w simplifies to Equation (15) - this is discussed in Appendix B.6. Second, to show that the fixed points of the state evolution equations $(m_t^{\text{amp}}, q_t^{\text{amp}}, v_t^{\text{amp}})$ indeed corresponds to the sufficient statistics $(m_t^\star, q_t^\star, v_t^\star)$ for the classifiers of interest. First, we provide a heuristic derivation of this fact, based on the replica method from statistical physics [Mézard et al., 1987]. We defer the discussion of the formal aspects to Sec. B.4.

B.3 Self-consistent equation from the replica method

As discussed above, the goal of this section is to provide a derivation of the self-consistent equations in eq. (14) from the replica method. For the particular case of $\hat{f}_{\text{erm}}$, this derivation appeared [Loureiro et al., 2021a], where it was also rigorously proven using CGMT. Here, we extend this analysis to $t \in \{\text{bo}, \text{eb}\}$.

We can treat the different classifiers of interest in the replica analysis by defining the following Gibbs distribution:

$$\mu_t(\boldsymbol{\theta}|\mathcal{D}) = \frac{1}{\mathcal{Z}_t} \prod_{\mu \in [n]} P_\sigma^t(y^\mu | \boldsymbol{\theta}^\top \boldsymbol{v}^\mu) \times P_{\boldsymbol{\theta}}^t(\boldsymbol{\theta}) \tag{82}$$

where $(P_\sigma^t, P_{\boldsymbol{\theta}}^t)$ are a likelihood and priors (not necessarily normalized) depending on the particular classifier, and are explicitly given in Table 3, and the normalization constant $\mathcal{Z}_t$ is the partition function.

The aim of the replica method is to compute the free energy density defined as:

$$\beta f_\beta = - \lim_{p \to \infty} \frac{1}{p} \mathbb{E}_{\mathcal{D}} \log \mathcal{Z}_t \tag{83}$$

The free energy is the cumulant generating function of the Gibbs measure, and therefore computing it give us access to the statistics of the measure, which in particular allow us to compute the test error and calibration (among others quantities) of the classifiers defined in Sec. 2.1. Since taking the expectation over the log is intractable, we resort to the *replica method* [Mézard et al., 1987], which consists of the following trick:

$$\log \mathcal{Z}_t = \lim_{r \to 0^+} \frac{1}{r} \mathcal{Z}_t^r \tag{84}$$

Classifier	$P_\sigma^t(y\|z)$	$P_\theta^t(\boldsymbol{\theta})$
$\hat{f}_{\text{erm}}$	$\sigma(y \times z)^\beta$	$e^{-\beta\lambda/2\|\boldsymbol{\theta}\|^2}$
$\hat{f}_{\text{eb}}$	$\sigma(\beta y \times z)$	$\mathcal{N}(\boldsymbol{\theta}\|\mathbf{0}, {}^1/_\lambda I_p)$
$\hat{f}_{\text{bo}}$	$\sigma_{\tau_0^2+\tau_{\text{add}}^2}(y \times z)$	$\mathcal{N}(\boldsymbol{\theta}\|\mathbf{0}, \Sigma_\star)$

Table 3: Prior and likelihood for the different estimators. For $\hat{f}_{\text{erm}}$, the temperature β must be taken in the limit $\beta \to \infty$, and the Gibbs measure $\mu_{\text{erm}}(\boldsymbol{\theta}|\mathcal{D})$ is peaked around the minimizer of the empirical risk $\hat{\boldsymbol{\theta}}_{\text{erm}}$.

Swapping the limit and the expectation, what we need to compute is:

$$\begin{aligned}\mathbb{E}_{\mathcal{D}}\mathcal{Z}_t^r &= \prod_{\mu=1}^n \mathbb{E}_{\boldsymbol{v}^\mu, y^\mu} \prod_{a=1}^r \int_{\mathbb{R}^d} P_{\boldsymbol{\theta}}^t(\boldsymbol{\theta}^a)\, P_\sigma^t\left(y^\mu|\frac{\boldsymbol{v}^{\mu\top}\boldsymbol{\theta}^a}{\sqrt{d}}\right) \\ &= \prod_{\mu=1}^n \sum_y \int P(\boldsymbol{w}_\star) \int \left(\prod_a P_{\boldsymbol{\theta}}^t(\boldsymbol{\theta}^a)\right) \mathbb{E}_{\boldsymbol{v}^\mu}\left[P_0\left(y^\mu|\frac{\boldsymbol{v}^{\mu\top}\boldsymbol{w}_\star}{\sqrt{d}}\right)\prod_a P_\sigma^t\left(y^\mu|\frac{\boldsymbol{v}^{\mu\top}\boldsymbol{\theta}^a}{\sqrt{p}}\right)\right]\end{aligned}$$

Next, we introduce the *local fields* $\nu_\star^\mu = \frac{1}{\sqrt{d}}\boldsymbol{v}^{\mu\top}\boldsymbol{w}_\star$ and $\nu_a^\mu = \frac{1}{\sqrt{d}}\boldsymbol{v}^{\mu\top}\boldsymbol{\theta}^a$. Then, the term between brackets in the above equation is equal to

$$\int \mathrm{d}\nu_\star^\mu P_0(y^\mu|\nu_\star^\mu) \int \prod_a \mathrm{d}\nu_a^\mu P_\sigma^t(y^\mu|\nu_a^\mu)\mathbb{E}_{\boldsymbol{v}^\mu}\left[\delta(\nu_\star^\mu - \frac{\boldsymbol{v}^{\mu\top}\boldsymbol{w}_\star}{\sqrt{d}})\prod_a \delta(\nu_a^\mu - \frac{\boldsymbol{v}^{\mu\top}\boldsymbol{\theta}^a}{\sqrt{p}}))\right] \tag{85}$$

Note that $\mathbb{E}_{\boldsymbol{v}^\mu}\left[\delta(\nu_\star^\mu - \boldsymbol{v}^{\mu\top}\boldsymbol{w}_\star)\prod_a \delta(\nu_a^\mu - \boldsymbol{v}^{\mu\top}\boldsymbol{\theta}^a))\right]$ defines the joint distribution of the local fields. It is straightforward to show that this is a Gaussian distribution on zero mean and covariance Σ_ν:

$$\mathbb{E}(\nu_\star, \nu_\star) = \frac{1}{d}\boldsymbol{w}_\star^\top\Omega\boldsymbol{w}_\star = \rho, \quad \mathbb{E}(\nu_\star, \nu_a) = \frac{1}{\sqrt{pd}}\boldsymbol{w}_\star^\top\Omega\boldsymbol{\theta}^a = m^a, \quad \mathbb{E}(\nu_a, \nu_b) = \frac{1}{p}\boldsymbol{\theta}^{a\top}\Omega\boldsymbol{\theta}^b = Q^{ab} \tag{86}$$

Then, we have

$$\mathbb{E}\mathcal{Z}_t^r = \prod_\mu \sum_{y^\mu} \int P_{\boldsymbol{\theta},0}(\boldsymbol{w}_\star) \int \prod_a P_{\boldsymbol{\theta}}^t(\boldsymbol{\theta}^a) \int \mathrm{d}\nu_\star^\mu \prod_a \mathrm{d}\nu_a^\mu P_0(y^\mu|\nu_*^\mu)\prod_a P_\sigma^t(y^\mu|\nu_a^\mu) \times \mathcal{N}(\nu_*^\mu, \nu_a^\mu|\mathbf{0}, \Sigma_\nu) \tag{87}$$

The elements of the covariance matrix Σ_ν are fixed by eq. (86). We can free these overlaps by doing the Fourier transform of the Dirac delta. We get in the end

$$\mathbb{E}_{\mathcal{D}}\mathcal{Z}_t^r \propto \int \mathrm{d}\rho\mathrm{d}\hat{\rho}\prod_a \mathrm{d}m^a\mathrm{d}\hat{m}^a\prod_{a,b}\mathrm{d}Q^{ab}\mathrm{d}\hat{Q}^{ab}e^{p\Phi(r)} \tag{88}$$

Where

$$\Phi(r) = -\frac{1}{\gamma}\rho\hat{\rho} - \frac{1}{\sqrt{\gamma}}\sum_a m^a\hat{m}^a - \sum_{a\leqslant b} Q^{ab}\hat{Q}^{ab} + \alpha \times \Psi_y^{(r)} + \Psi_w^{(r)} \tag{89}$$

$$\Psi_y^{(r)} = \frac{1}{p}\log\int P_{\boldsymbol{\theta},0}(\boldsymbol{w}_\star)\int\prod_a P_{\boldsymbol{\theta}}^t(\boldsymbol{\theta}^a)e^{\hat{\rho}\|\boldsymbol{w}_\star\|^2 + \sum_a \hat{m}^a\boldsymbol{w}_\star\Omega\boldsymbol{\theta}^a + \sum_{a\leqslant b}\hat{Q}^{a,b}\boldsymbol{\theta}^a\Omega\boldsymbol{\theta}^b} \tag{90}$$

$$\Psi_w^{(r)} = \frac{1}{p}\log\sum_y\int \mathrm{d}\nu_\star P_0(y|\nu_\star)\int\prod_a \mathrm{d}\nu_a P_\sigma^t(y|\nu_a)\mathcal{N}(\nu, \nu_a; \Sigma_\nu) \tag{91}$$

B.3.1 Replica symmetric ansatz

In the replica symmetric ansatz, we assume $m^a = m$, $Q^{ab} = q$ for $a \neq b$, $Q^{aa} = v + q$, $\hat{m}^a = \hat{m}$, $\hat{Q}^{ab} = \hat{q}$ for $a \neq b$, $\hat{Q}^{aa} = -\frac{1}{2}(\hat{v} - \hat{q})$ where the quantities $m, q, v, \hat{m}, \hat{q}, \hat{v}$ are to be determined.

We refer to [Gerace et al., 2020, Aubin et al., 2020] for the detailed computation of $\lim_{r\to 0^+} \Psi_y^{(r)}$ and $\lim_{r\to 0^+} \Psi_w^{(r)}$. In the end, we obtain :

$$\boxed{\begin{aligned} f_\beta &= \mathrm{extr}_{m,q,v,\hat{m},\hat{q},\hat{v}} \left\{ -\frac{1}{\sqrt{\gamma}} m\hat{m} + \frac{1}{2}(q\hat{v} - \hat{q}v + \hat{v}v) + \Psi_w + \alpha \times \Psi_y \right\} && (92)\\ \Psi_w &= \lim_{d\to\infty} \frac{1}{p} \mathbb{E}_{\xi, \boldsymbol{w}_\star} \log \int \mathrm{d}\boldsymbol{\theta} P_{\boldsymbol{\theta}}^t(\boldsymbol{\theta}) e^{-\hat{v}^2 \boldsymbol{\theta}^\top \Omega \boldsymbol{\theta} + \boldsymbol{\theta}^\top (\hat{m}\Omega \boldsymbol{w}_\star + \hat{q}\Omega^{-1/2}\boldsymbol{\xi})} && (93)\\ \Psi_y &= \mathbb{E}_{\xi\sim\mathcal{N}(0,q)} \left[\sum_y \mathcal{Z}_0(y, {}^m/_q \xi, \rho - {}^{m^2}/_q) \log \mathcal{Z}_g(y, \xi, v) \right] && (94) \end{aligned}}$$

where

$$\mathcal{Z}_{0/g}(y, \omega, v) = \int \mathrm{d}z P_{0/g}(y|z) \mathcal{N}(z|\omega, v) \tag{95}$$

The self-consistent equations (14) are obtained by cancelling the derivative of the free energy with respect to each of $(m, q, v, \hat{m}, \hat{q}, \hat{v})$.

Ψ_w for Gaussian priors: For all the estimators considered here, the prior distribution $P_{\boldsymbol{\theta}}^t(\boldsymbol{\theta})$ is Gaussian $\mathcal{N}(0, \Sigma)$, where Σ depends on the considered estimator. Then,

$$\begin{aligned} \Psi_w &= \int \mathrm{d}\boldsymbol{\theta} e^{-\frac{1}{2}\boldsymbol{\theta}^\top \Sigma^{-1} \boldsymbol{\theta}} e^{-\frac{v}{2}\boldsymbol{\theta}^\top \Omega \boldsymbol{\theta} + \boldsymbol{\theta}^\top (\hat{m}\Omega \boldsymbol{w}_\star + \sqrt{\hat{q}}\Omega^{-1/2}\boldsymbol{\xi})} && (96)\\ &= \frac{\exp\left(\frac{1}{2}(\hat{m}\boldsymbol{w}_\star + \sqrt{\hat{q}}\Omega^{-1/2}\boldsymbol{\xi})^\top (\Sigma + \hat{v}\Omega)^{-1} (\hat{m}\boldsymbol{w}_\star + \sqrt{\hat{q}}\Omega^{-1/2}\boldsymbol{\xi})\right)}{\sqrt{\det(\Sigma + \hat{v}\Omega)}} && (97)\\ &= \lim -\frac{1}{2p} \operatorname{Tr} \log(\Sigma + \hat{v}\Omega) + \frac{1}{2p} \operatorname{Tr}(\hat{m}^2 \Omega \boldsymbol{w}_\star \boldsymbol{w}_\star^\top \Omega + \hat{q}\Omega)(\Sigma + \hat{v}\Omega) && (98) \end{aligned}$$

We get in the end the following expression for Ψ_w

$$\Psi_w = -\frac{1}{2p} \operatorname{Tr} \log(\hat{v}\Omega + \Sigma) + \frac{1}{2p} \operatorname{Tr}\left((\hat{m}^2 \Omega \boldsymbol{w}_\star \boldsymbol{w}_\star^\top \Omega + \hat{q}\Omega)(\hat{v}\Omega + \Sigma)^{-1} \right) \tag{99}$$

$$\tag{100}$$

Saddle-point equations: To compute the free energy, we cancel its derivative with respect to $m, q, v, \hat{m}, \hat{q}, \hat{v}$. We have :

$$\begin{cases} \partial_{\hat{m}} f_\beta &= -\frac{1}{\sqrt{\gamma}} m + \partial_{\hat{m}} \Psi_w \\ \partial_{\hat{q}} f_\beta &= -\frac{1}{2} v + \partial_{\hat{q}} \Psi_w \\ \partial_{\hat{v}} f_\beta &= \frac{1}{2}(v + q) + \partial_{\hat{v}} \Psi_w \end{cases} \tag{101}$$

Cancelling the derivatives gives the condition:

$$\boxed{\begin{cases} m = \sqrt{\gamma} \partial_{\hat{m}} \Psi_w \\ v = 2 \times \partial_{\hat{q}} \Psi_w \\ q = -v - 2 \times \partial_{\hat{v}} \Psi_w = 2 \times (\partial_{\hat{q}} \Psi_w - \partial_{\hat{v}} \Psi_w) \end{cases}} \tag{102}$$

Which are the first three equations of Theorem 3.1. The derivative of the free energy with respect to (m, q, v) is given by

$$\begin{cases} \partial_m f_\beta &= -\frac{1}{\sqrt{\gamma}} \hat{m} + \alpha \partial_m \Psi_y \\ \partial_q f_\beta &= \frac{1}{2}\hat{v} + \alpha \partial_q \Psi_y \\ \partial_v f_\beta &= \frac{1}{2}(\hat{v} - \hat{q}) + \alpha \partial_v \Psi_y \end{cases} \tag{103}$$

Cancelling the derivatives, and computing the derivatives of Ψ_y gives then

$$\boxed{\begin{cases} \hat{v} &= -\alpha \mathbb{E}_{\xi\sim\mathcal{N}(0,q)} \left[\sum_y \mathcal{Z}_0 \left(y, {}^{m}\!/_{q}\xi, v_\star\right) \partial_\omega g_t \left(y, \xi, v\right) \right] \\ \hat{q} &= \alpha \mathbb{E}_{\xi\sim\mathcal{N}(0,q)} \left[\sum_y \mathcal{Z}_0 \left(y, {}^{m}\!/_{q}\xi, v_\star\right) g_t \left(y, \xi, v\right)^2 \right] \\ \hat{m} &= \alpha\sqrt{\gamma} \mathbb{E}_{\xi\sim\mathcal{N}(0,q)} \left[\sum_y \partial_\omega \mathcal{Z}_0 \left(y, {}^{m}\!/_{q}\xi, v_\star\right) g_t \left(y, \xi, v\right) \right] \end{cases}} \tag{104}$$

which are the last three equations for $\hat{m}, \hat{q}, \hat{v}$ in Theorem 3.1.

Therefore, we have shown that the self-consistent equations characterizing the sufficient statistics in Theorem 3.1 can be obtained from the replica method by computing the asymptotic free energy density. Moreover, in Sec. B.2 we have shown that these equations exactly agree with the state evolution equations for a tailored GAMP algorithm 1.

B.4 Rigorous version of replica and self-consistent equations

As discussed in the introduction of this Appendix, the derivation of Theorem 3.1 consists in two steps. First, one constructs a tailored GAMP algorithm 1 for which the estimates can be exactly tracked by a set of state evolution equations. Second, one shows that these equations actually agree with the self-consistent equations describing the sufficient statistics for the joint density of interest in Theorem 3.1. The first part was discussed in Sec. B.2, and although we provided an informal derivation of the state evolution equations, they rigorously follow from the recent progress on state evolution proofs for structured message passing schemes with non-separable priors [Berthier et al., 2019, Gerbelot and Berthier, 2021]. For the second part, in Sec.B.3 we discussed a heuristic derivation of the self-consistent equations from the replica method, and showed it agrees with the state evolution equations from GAMP. Therefore, it remains to rigorously justify this last step. Thankfully, one can resort to a large number of recent progress on generic proofs of the replica predictions [Bayati and Montanari, 2012, Barbier et al., 2019, Sur and Candès, 2019, Candes and Sur, 2018, Montanari et al., 2020, Dhifallah and Lu, 2020, Loureiro et al., 2021a], which we now discuss in detail.

First, let us recall the statement of the theorem in the more general context of the Gaussian covariate model. Let $(\boldsymbol{u}, \boldsymbol{v})$ denote a pair of Gaussian covariates:

$$\begin{pmatrix} \boldsymbol{u} \\ \boldsymbol{v} \end{pmatrix} \sim \mathcal{N}\left(\mathbf{0}_{d+p}, \begin{bmatrix} \Psi & \Phi \\ \Phi^\top & \Omega \end{bmatrix} \right). \tag{105}$$

For any of the classifiers $t \in \{\text{bo}, \text{erm}, \text{Lap}, \text{eb}\}$ from Sec. 2.1, the 2-dimensional vector $(f_\star(\boldsymbol{u}), \hat{f}_t(\boldsymbol{v}))$ is asymptotically distributed as $(\sigma(z), \sigma_{\tilde{v}}(z_t'))$ for some $\tilde{v}$ that depends on the estimator, where $(z, z_t') \sim \mathcal{N}(\mathbf{0}_2, \Sigma_t)$, and

$$\Sigma_t = \begin{pmatrix} \boldsymbol{\theta}_\star^\top \Psi \boldsymbol{\theta}_\star^\top & \hat{\boldsymbol{\theta}}_t^\top \Phi \boldsymbol{\theta}_\star \\ \boldsymbol{\theta}_\star^\top \Phi^\top \hat{\boldsymbol{\theta}}_t & \hat{\boldsymbol{\theta}}_t^\top \Omega \hat{\boldsymbol{\theta}}_t \end{pmatrix}$$

where $\hat{\boldsymbol{\theta}}_t$ is either the unique minimizer the empirical risk in eq. (2) for $t \in \{\text{erm}, \text{Lap}\}$ or the mean over the respective posterior distribution for $t \in \{\text{bo}, \text{eb}\}$. The computation of $\rho_{\star,t}$ thus boils down to computing the sufficient statistics $(\hat{\boldsymbol{\theta}}_t^\top \Phi \boldsymbol{\theta}_\star, \hat{\boldsymbol{\theta}}_t^\top \Omega \hat{\boldsymbol{\theta}}_t)$. A first important point is that, asymptotically in p, these quantities converge in probability to single, deterministic quantities. This was shown in general for sampling problems with log concave measure (such as the one we use in the Bayes-optimal and empirical Bayes method) in [Barbier et al., 2021b], and for empirical risk minimization with convex risks in [Loureiro et al., 2021a]. We shall thus use the following lemma:

Lemma B.1 (Overlap Concentration, from [Barbier et al., 2021b, Loureiro et al., 2021a])**.** *In the asymptotic limit $p \to \infty$, the random variables $(\hat{\boldsymbol{\theta}}_t^\top \Phi \boldsymbol{\theta}_\star, \hat{\boldsymbol{\theta}}_t^\top \Omega \hat{\boldsymbol{\theta}}_t)$ converge in probability to some value $(m_t^\star, q_t^\star)$ for $t \in \{\text{bo}, \text{erm}, \text{Lap}, \text{eb}\}$.*

The problem is thus reduced to the computation of these statistics, as a function of the parameters of the problems (α, γ, τ_0, etc.) for each of the estimators of interest. In Theorem 3.1, we claim that these are given by the replica equations derived in Appendix B.3. Thankfully, for different estimators these equations were proven in the literature in slightly different contexts, written as formal proofs of replica predictions.

- For $\hat{f}_{\rm erm}$ on the random features model the self-consistent equations for $(m^\star_{\rm erm}, q^\star_{\rm erm})$ were heuristically derived in [Gerace et al., 2020] and rigorously proven in [Dhifallah and Lu, 2020]. In the more general context of the Gaussian covariate model, analogous equations were proven in [Loureiro et al., 2021a]. In both cases, they agree with our equations in eq. (14). While these works use the Gordon minimax approach to prove these equations, we note an independent GAMP-based proof for both the random features and Gaussian covariate models appeared in [Loureiro et al., 2022], leveraging recent progress on structured message passing schemes from [Gerbelot and Berthier, 2021].

- As noted in Sec. 2.1, the average over the Laplace posterior agrees exactly with the empirical risk minimizer:

$$p_{\rm Lap}(\boldsymbol{\theta}|\mathcal{D}) = \mathcal{N}(\boldsymbol{\theta}|\hat{\boldsymbol{\theta}}_{\rm erm}, \mathcal{H}^{-1}) \tag{106}$$

  Therefore, the self-consistent equations for $(m^\star_{\rm Lap}, q^\star_{\rm Lap})$ agree exactly with the ones for $(m^\star_{\rm erm}, q^\star_{\rm erm})$. Therefore, they are also rigorous.

  In both cases, our result follows from:

  Theorem B.2 (ERM statistics, Thms. 4 & 5 from [Loureiro et al., 2021a], Informal)**.** *In the setting of Theorem 3.1, the ERM predictions from the replica are correct:* (m, q, v) *converges in probability to their replica fixed points* $(m^\star_{\rm erm}, q^\star_{\rm erm}, v^\star_{\rm erm})$*, while the minimum training error converges in probability to the replica free energy density.*

- The "finite temperature" sampling problems related to Bayesian estimation pose different challenges. We start by discussing the Bayes-optimal $\hat{f}_{\rm bo}$ classifier. For i.i.d. Gaussian data, the rigour of the replica prediction has been proven for Generalized linear models in [Barbier et al., 2019], together with the GAMP optimality. Thanks to Gaussian equivalence, our problem can be framed as a Bayesian generalized linear reconstruction problem, but with data matrix that are instead correlated. In the random features case, the data matrix is a product of two random matrix (see eq.(29)). Thus, in this the case replica predictions were for the Bayes-optimal problem was rigorously proven in [Gabrié et al., 2018, Barbier et al., 2018]. Note that while these works only prove the correctness of the replica free energy density, the techniques in [Barbier et al., 2019] can be readily applied to generalize the proof to overlaps:

  Theorem B.3 (BO statistics, Th. 1 from [Barbier et al., 2018] and Th. 1 from [Barbier et al., 2018], Informal)**.** *In the setting of Theorem 3.1, the BO prediction from the replica is correct:* (m, q, v) *converges in probability to their replica fixed points* $(m^\star_{\rm bo}, q^\star_{\rm bo}, v^\star_{\rm bo})$*, while the minimum training error converges in probability to the replica free energy density.*

  Additionally, given that the performance of the GAMP algorithms follows the same self-consistent equations as the replica's [Berthier et al., 2019, Gerbelot and Berthier, 2021], it follows that GAMP performs Bayes-optimal estimation for this problem, a classical property in Bayesian estimation [Zdeborová and Krzakala, 2016]. [3]

- The remaining case is the empirical Bayes (EM) classifier $\hat{f}_{\rm eb}$. In this case, where Bayesian estimation is performed *with mismatched noise*, the complete proof of the replica equation is not available in the literature. In principle, this can be done following the steps of [Barbier et al., 2021a] for the square loss (recall we consider the logistic loss in this work). Indeed, [Barbier et al., 2021a] shows how the concentration of $(m^\star, q^\star)$ (referred to as strong replica symmetry in [Barbier et al., 2021b]) can be used together with rigorous control of the cavity method [Aizenman et al., 2006] to prove the cavity equations. While this is, we believe, a worthwhile direction of research, we instead shall redefine the empirical Bayes method performance as the one of the *best empirical Bayesian estimator* in linear time, that is, the estimator achieved by the GAMP algorithm 1 with the corresponding empirical Bayes denoiser. It can indeed be shown that GAMP is the best *first-order algorithm* for this

[3] Note that this crucially relies on the strong replica symmetry [Barbier et al., 2021b] condition, which impose the existence of an unique fixed point in our problem. Without this property, one could generically have more than a fixed point, associated to a so-called "hard phase" where GAMP is not optimal, see [Zdeborová and Krzakala, 2016].

class of Bayesian estimation problems [Celentano et al., 2020], and it is widely expected to perform an exact sampling for these problems [Zdeborová and Krzakala, 2016] (as it was proven in for the Bayes-optimal case). With this definition, the performance of GAMP is by construction given by its rigorous state evolution [Berthier et al., 2019, Gerbelot and Berthier, 2021], which we recall the reader matches the replica prediction.

B.5 Laplace approximation : computing the inverse Hessian

In this section, we show how to compute the prediction for the Laplace approximation

$$\hat{f}_{\text{Lap}}(\boldsymbol{v}) = \int \mathrm{d}z\sigma(z)\mathcal{N}(z|\hat{\boldsymbol{\theta}}_{\text{erm}}^{\top}\boldsymbol{v}, \boldsymbol{v}^{\top}\mathcal{H}^{-1}\boldsymbol{v}) \tag{107}$$

with $\mathcal{H}$ the Hessian of the empirical risk at $\hat{\boldsymbol{\theta}}_{\text{erm}}$. Note that in the high-dimensional limit, $\boldsymbol{v}^{\top}\mathcal{H}^{-1}\boldsymbol{v} \to_{p\to\infty} \text{Tr}(\mathcal{H}^{-1}\Omega)$. As shown in Appendix C, to compute this quantity we can add the term $\boldsymbol{h}^{\top}\boldsymbol{\theta}$ to the loss and compute the second derivative of the free energy density with respect to $\boldsymbol{h}$. The computations are the same as those done in Section B.3, except that the Gibbs distribution $\mu_t(\boldsymbol{\theta})$ is replaced by

$$\mu_t(\boldsymbol{\theta}) = \frac{1}{\mathcal{Z}_t(\boldsymbol{h})}\prod_i P_{\sigma}^t(y_i|\boldsymbol{\theta}^{\top}\boldsymbol{\varphi}(\boldsymbol{x}_i)) \times P_{\boldsymbol{\theta}}^t(\boldsymbol{\theta}) \times e^{\beta\boldsymbol{h}^{\top}\boldsymbol{\theta}} \tag{108}$$

Adapting the derivation from Sec. B.3 for $\hat{f}_{\text{erm}}$ and taking the temperature $\beta \to \infty$ and get as before

$$f_0 := \lim_{\beta\to\infty} f_{\beta} = \underset{m,q,v,\hat{m},\hat{q},\hat{v}}{\textbf{extr}} \left\{ -\frac{1}{\sqrt{\gamma}}m\hat{m} + \frac{1}{2}(q\hat{v} - \hat{q}v) + \Psi_w(\hat{m},\hat{q},\hat{v},\boldsymbol{h}) + \alpha\Psi_y(m,q,v) \right\} \tag{109}$$

$$\Psi_y = \mathbb{E}_{\xi\sim\mathcal{N}(0,q)}\left[\sum_y \mathcal{Z}_0(y, {}^m/_q\xi, \rho - {}^{m^2}/_q)\log\mathcal{Z}_g(y,\xi,v)\right] \tag{110}$$

However, now Ψ_w is

$$\Psi_w = -\frac{1}{2p}\,\text{Tr}\log\left(\hat{v}\Omega + \Sigma\right) + \frac{1}{2p}\,\text{Tr}\left[((\hat{m}\Omega\boldsymbol{w}_{\star} + \boldsymbol{h})(\hat{v}\Omega + \lambda I)^{-1}(\hat{m}\Omega\boldsymbol{w}_{\star} + \boldsymbol{h}) + \hat{q}\Omega(\hat{v}\Omega + \Sigma)^{-1})\right] \tag{111}$$

The second derivative of Ψ_w with respect to $\boldsymbol{h}$ is $(\lambda I_d + \hat{v}\Omega)^{-1}$. As a consequence, the second derivative of the free energy

$$(\nabla_{\mathbf{h}}^2\log\mathcal{Z}_{\text{erm}})_{|\boldsymbol{h}=\boldsymbol{0}} = \nabla_{\mathbf{h}}^2\Psi_w(m_{\text{erm}}^{\star}, q_{\text{erm}}^{\star}, v_{\text{erm}}^{\star}, \boldsymbol{h}) = (\lambda I_d + \hat{v}_{\text{erm}}^{\star}\Omega)^{-1}$$

and $\nabla_{\mathbf{h}}^2 f_0 = -(\lambda I_d + \hat{v}_{\text{erm}}^{\star}\Omega)^{-1}$. We then deduce that the inverse Hessian is equal to

$$\mathcal{H}^{-1} = (\lambda \mathrm{I}_d + \hat{v}_{\text{erm}}^{\star}\Omega)^{-1} \tag{112}$$

B.6 Simplification for random features

As discussed in Appendix B, the random features model $\boldsymbol{\varphi}(\boldsymbol{x}) = \phi(\mathrm{F}\boldsymbol{x})$ is asymptotically equivalent to the Gaussian covariate model up to an identification of the covariances:

$$\Omega = \kappa_1^2 FF^{\top} + \kappa_{\star}^2\mathrm{I}_p, \qquad \Phi = \kappa_1 F, \qquad \Psi = \mathrm{I}_d \tag{113}$$

where:

$$\kappa_1 = \mathbb{E}_{z\sim\mathcal{N}(0,1)}\left[\phi'(z)\right], \qquad \kappa_{\star} = \sqrt{\mathbb{E}_{z\sim\mathcal{N}(0,1)}\left[\phi(z)^2\right] - \kappa_1^2} \tag{114}$$

where for simplicity we assume $\kappa_0 = \mathbb{E}_{z\sim\mathcal{N}(0,1)}[\phi(z)] = 0$. Thus, in this case we can explicitly write:

$$\Sigma_{\star} = \frac{\kappa_1^2\mathrm{F}\mathrm{F}^{\top}}{\kappa_{\star}^2\mathrm{I}_p + \kappa_1^2\mathrm{F}\mathrm{F}^{\top}}. \tag{115}$$

. Note that the matrices $\Omega, \Sigma, \Sigma_\star$ are diagonalizable in the same basis, since Σ is either a multiple of the identity, or a function of $\Omega, \Phi\Phi^\top$. Assuming that $\mathrm{F}\mathrm{F}^\top$ has an asymptotic spectral distribution μ, we can write Ψ_w directly in terms of an average over μ:

$$\Psi_w = \frac{1}{2}\mathbb{E}_{x\sim\mu)}\left[\log\left(\hat{v}(\kappa_1^2 x + \kappa_\star^2) + \pi(x)\right) + \left(\frac{\hat{m}^2\frac{\kappa_1^2 x}{\kappa_1^2 x+\kappa_\star^2} + \hat{q}(\kappa_1^2 x + \kappa_\star^2)}{\hat{v}(\kappa_1^2 x + \kappa_\star^2) + \pi(x)}\right)\right] \tag{116}$$

where the function π represents the eigenvalues of Σ : since we can write $\Sigma = f(\Phi\Phi^\top)$ here, we have, $\pi(x) = f(x)$. For $\hat{f}_{\text{erm}}$ and $\hat{f}_{\text{eb}}$, $\pi(x) = \lambda$. For $\hat{f}_{\text{bo}}$, $\pi(x) = \frac{\kappa_1^2 x}{\kappa_1^2 x+\kappa_\star}$. This gives us the values of $\hat{\pi}_t$ in Table 1.

In particular, when F has Gaussian i.i.d. entries (as in all plots presented here), μ is simply the Marcenko-Pastur distribution with shape parameter γ.

B.7 Temperature scaling

In this section, we show how to compute the optimal temperature T that minimizes the test loss for $\hat{f}_{\text{erm}}$. Once the overlaps $m^\star, q^\star, v^\star, \hat{m}^\star, \hat{q}^\star, \hat{v}^\star$ are computed, we get the test loss with the expression

$$\mathcal{L}_{\text{gen.}}(m, q) = \sum_y \mathbb{E}_{\xi\sim\mathcal{N}(0,1)}\left[\mathcal{Z}_0(y, {}^{m^\star}\!/_{q^\star}\xi, \rho - m^2/q) \times (-\log\sigma(y \times \sqrt{q}\xi))\right] \tag{117}$$

Given a temperature T, temperature scaling will divide the weights such that the prediction is now $\sigma(\boldsymbol{\theta}^\top\boldsymbol{\varphi}(\boldsymbol{x}/T)$. It is easy to see that in this case, the overlaps $m^\star, q^\star$ now become $m^\star/T, q^\star/T^2$. Then, temperature scaling amounts to finding

$$T^\star = \arg\min_T \mathcal{L}_{\text{gen.}}(m^\star/T, q^\star/T^2) \tag{118}$$

C Confidence function and Hessian of Laplace Method

C.1 Computing the Hessian of the training loss

In this section, we show how we can compute the (inverse of) the Hessian thanks to classical properties of Legendre transforms. We consider the ERM estimator $\hat{f}_{\text{erm}}$ trained by minimizing the following loss :

$$\mathcal{L}(\boldsymbol{w}) = -\sum_i \log \sigma(\boldsymbol{\theta}^\top \boldsymbol{\varphi}(\boldsymbol{x})_i \times y_i) + \lambda/2\|\boldsymbol{\theta}\|^2 \tag{119}$$

whose Hessian at the minimum is given by

$$\mathcal{H} := \nabla^2 \mathcal{L} = -\sum_i (1 - \sigma'(\boldsymbol{\theta}^\top \boldsymbol{\varphi}(\boldsymbol{x})_i \times y_i))\boldsymbol{\varphi}(\boldsymbol{x})_i \boldsymbol{\varphi}(\boldsymbol{x})_i^\top + \lambda I_d \big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_{\text{erm}}} \tag{120}$$

Our starting point to compute this Hessian is a very classical lemma in statistical mechanics, that uses the Legendre transform of the loss.

Lemma C.1 (Inverse Hessian from Legendre Transforms)**.** *We define the Legendre transform of the loss by adding a source term to the loss (an external field in the parlance of statistical mechanics)*

$$\mathcal{L}^L(\boldsymbol{h}) = \min_{\boldsymbol{\theta}} \left[-\sum_i \log \sigma(y_i \boldsymbol{\theta}^\top \boldsymbol{\varphi}(\boldsymbol{x})_i) + \lambda/2\|\boldsymbol{\theta}\|^2 + \boldsymbol{h}^\top \boldsymbol{\theta} \right] = \min_{\boldsymbol{\theta}} \left[\mathcal{L}(\boldsymbol{\theta}) + \boldsymbol{h}^\top \boldsymbol{\theta} \right] \tag{121}$$

then the Inverse of the Hessian (120) *is the Hessian of the Legendre transform* $\mathcal{L}(\boldsymbol{h})$

$$\mathcal{H}^{-1}(\hat{\boldsymbol{\theta}}_{\text{erm}}) = -\frac{\partial^2 \mathcal{L}^L(\boldsymbol{h})}{\partial^2 \boldsymbol{h}} \big|_{\boldsymbol{h}=0} \tag{122}$$

Proof This is a classical result from Legendre transform of strongly convex functions, which we informally recall. First notice that at the minimum of $\mathcal{L}(\boldsymbol{\theta}) + \boldsymbol{h}^\top \boldsymbol{\theta}$ over $\boldsymbol{\theta}$ is characterized by

$$\frac{\partial \mathcal{L}(\boldsymbol{\theta})}{\partial \theta_j} + h_j = 0 \ \forall j \tag{123}$$

so that

$$\begin{aligned} \frac{\partial \mathcal{L}^L(\boldsymbol{h})}{\partial h_i} &= \frac{\partial \left[\mathcal{L}(\boldsymbol{\theta}) + \boldsymbol{h}^\top \boldsymbol{\theta}\right]}{\partial h_i} \big|_{\hat{\boldsymbol{\theta}}_{\text{erm}}} = \sum_{j=1}^p \left[\frac{\partial \mathcal{L}(\boldsymbol{\theta})}{\partial \theta_j} \frac{\partial \theta_j}{\partial h_i} + h_j \frac{\partial \theta_j}{\partial h_i} \right] \big|_{\hat{\boldsymbol{\theta}}_{\text{erm}}} + \theta_i \big|_{\hat{\boldsymbol{\theta}}_{\text{erm}}} \\ &= \sum_{j=1}^p \frac{\partial \theta_j}{\partial h_i} \left[\frac{\partial \mathcal{L}(\boldsymbol{\theta})}{\partial \theta_j} + h_j \right] \big|_{\hat{\boldsymbol{\theta}}_{\text{erm}}} + \theta_i \big|_{\hat{\boldsymbol{\theta}}_{\text{erm}}} = \theta_{\text{erm},i} \end{aligned} \tag{124}$$

It thus follows that

$$\frac{\partial^2 \mathcal{L}^L(\boldsymbol{h})}{\partial h_i \partial h_j} = \frac{\partial \theta_i}{\partial h_j}. \tag{125}$$

However, we have from eq.(123)

$$\frac{\partial^2 \mathcal{L}(\boldsymbol{\theta})}{\partial \theta_i \partial \theta_j} = -\frac{\partial h_j}{\partial \theta_i}. \tag{126}$$

Using both eqs (125) and (126) at $h = 0$ concludes the proof. □

Note that this relation is not asymptotic and is valid for a given instance of the problem. This lemma is however particularly practical in the large n limit, since an asymptotic expression for the loss $\mathcal{L}^L$ is known so that we can use it to obtain the asymptotic expression. Using the value of the minimal loss from [Loureiro et al., 2021a, Loureiro et al., 2022], we deduce, taking its second derivative, that for large n we must have (See Section B.5 for the derivation)

$$\mathcal{H}^{-1} \underset{p\to\infty}{\asymp} (\lambda \mathrm{I}_p + \hat{v}^\star_{\text{erm}} \Omega)^{-1} \tag{127}$$

Where $\hat{v}^\star_{\text{erm}}$ is the unique solution the following self-consistent equations:

$$\begin{cases} m &= \frac{\gamma\hat{m}}{p}\operatorname{Tr}\left(\Omega \boldsymbol{w}_\star \boldsymbol{w}_\star^\top \Omega(\lambda \mathrm{I}_p + \hat{v}\Omega)^{-1}\right) \\ q &= \frac{1}{p}\operatorname{Tr}\left((\hat{q}\Omega + \hat{m}^2\Omega \boldsymbol{w}_\star \boldsymbol{w}_\star^\top \Omega)\Omega(\lambda \mathrm{I}_p + \hat{v}\Omega)^{-2}\right) , \\ v &= \frac{1}{p}\operatorname{Tr}(\lambda \mathrm{I}_p + \hat{v}\Omega)^{-1}\Omega) \end{cases}$$

$$\begin{cases} \hat{v} &= -\alpha \mathbb{E}_{\xi\sim\mathcal{N}(0,q)}\left[\sum_y \mathcal{Z}_0\left(y, {}^m\!/_q\xi, v_\star\right)\partial_\omega f_{\text{out,erm}}\left(y,\xi,v\right)\right] \\ \hat{q} &= \alpha \mathbb{E}_{\xi\sim\mathcal{N}(0,q)}\left[\sum_y \mathcal{Z}_0\left(y, {}^m\!/_q\xi, v_\star\right) f_{\text{out,erm}}\left(y,\xi,v\right)^2\right] \\ \hat{m} &= \alpha \mathbb{E}_{\xi\sim\mathcal{N}(0,q)}\left[\sum_y \partial_\omega \mathcal{Z}_0\left(y, {}^m\!/_q\xi, v_\star\right) f_{\text{out,erm}}\left(y,\xi,v\right)\right] \end{cases}$$

This leads us to the following theorem:

Theorem C.2 (Asymptotic form of the inverse Hessian)**.** *Assume that the convergence of the free energy to its asymptotic value in [Loureiro et al., 2021a] is such that the difference has bounded second derivative, then*

$$\mathcal{H}^{-1} = (\lambda I_p + \hat{v}^\star_{\text{erm}}\Omega)^{-1} + o(1) \tag{128}$$

While routinely made in the statistical physics literature, note that the assumption -akin to a uniform convergence of the function and its first and second derivatives- is a strong but unfortunately necessary one, that would warrant a dedicated investigation in itself. Such a delicate control of the correction to the free energy is however needed since, in our identification, we are exchanging the second derivative and the limit $p \to \infty$. Fortunately, it can be checked numerically that the assumption is empirically satisfied, and that the approximation given by the Hessian is extremely accurate, as is shown in the next section. Additionally, we note that the statement is made in term of an expression of the inverse of the Hessian (which, conveniently, is actually what we want to know).

C.2 Comparison with numerics

In this section, we apply the computations of the previous section and show that they gives extremly good prediction even at very moderate sizes. In Figure 3, we compare the theoretical value of $\boldsymbol{\varphi}(\boldsymbol{x})^\top \mathcal{H}^{-1}\boldsymbol{\varphi}(\boldsymbol{x})$ for $\boldsymbol{\varphi}(\boldsymbol{x}) = \operatorname{erf}(\mathrm{Fx})$ from eq. (127) and the one observed experimentally. Experiments are done by training the logistic classifier $\hat{f}_{\text{erm}}$ on training data $(\boldsymbol{x}^\mu, y^\mu)_{\mu\in[n]}$ and computing the Hessian (C.1) at the minimizer $\hat{\boldsymbol{\theta}}_{\text{erm}}$. We observe a good fit between theory and experiment, validation our analysis.

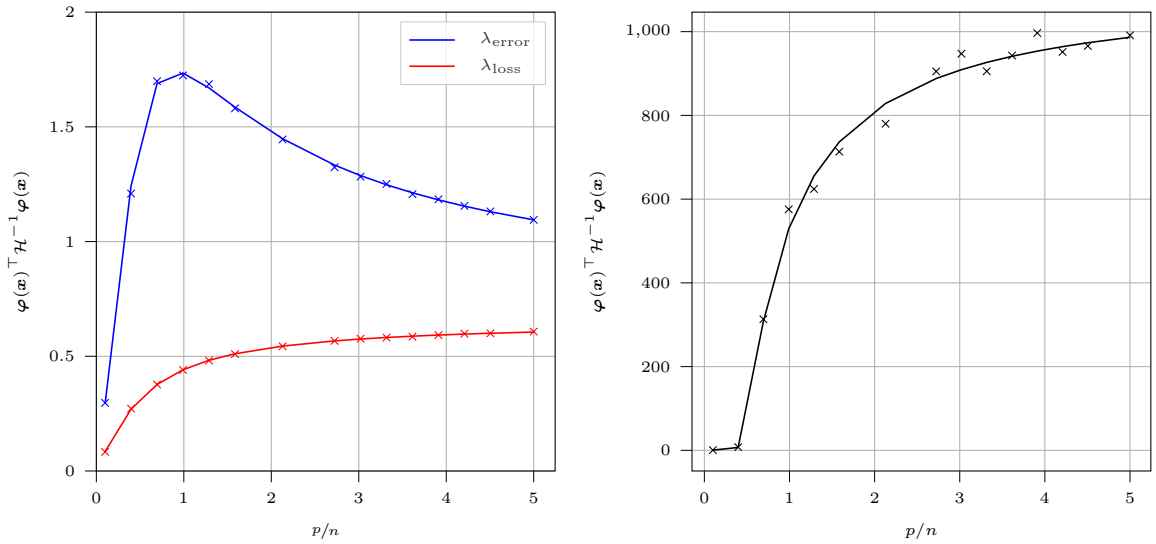

Figure 3: (**Left**) Theoretical predictions (lines) and experimental values (crosses) of $\boldsymbol{\varphi}(\boldsymbol{x})^\top \mathcal{H}^{-1}\boldsymbol{\varphi}(\boldsymbol{x})$ with ${}^n\!/_d = 2, \tau_0^2 = 0.5$, $\boldsymbol{\varphi}(\boldsymbol{x}) = \operatorname{erf}(\mathrm{Fx})$ and F Gaussian, as in Figure 1, for λ_{error} and λ_{loss}. Experimental values are obtained by fixing $d = 256$. (**Right**) Theoretical and experimental values for $\lambda = 10^{-4}$.

D Conditional variance of the Bayes-optimal estimator

In this section, we prove the expression of the variance of $\hat{f}_{\rm bo}$ conditioned on the confidence of other estimators that was given in theorem 3.3:

$$\mathrm{Var}(\hat{f}_{\rm bo}(\boldsymbol{x})|\hat{f}_t(\boldsymbol{x})=\ell)=\int \mathrm{d}a\ \sigma_{\hat{v}^\star_{\rm bo}+\tau_0^2+\tau^2_{\rm add}}(a)^2\times\mathcal{N}\left(a|m^\star_t/q^\star_t\sigma^{-1}_{\hat{\tau}_t}(\ell),q^\star_{\rm bo}-{m^\star_t}^2/q^\star_t\right)-(\ell-\Delta_\ell)^2 \quad (129)$$

The first step is to show that for any estimator $t\in\{\mathrm{erm},\mathrm{eb},\mathrm{Lap}\}$, the joint density of the confidence of $\hat{f}_{\rm bo},\hat{f}$, defined as

$$\rho_{{\rm bo},t}(a,b)=\mathbb{P}_{\boldsymbol{x}}(\hat{f}_{\rm bo}(\boldsymbol{x})=a,\hat{f}_t(\boldsymbol{x})=b) \quad (130)$$

can be computed in the similar way as $\rho_{\star,t}$ in Theorem 3.1. This was shown previously for a simpler model in [Clarté et al., 2022], where the teacher and input data have identity covariance.

Lemma D.1. *In the same setting as Theorem 3.1, in the asymptotic limit, the density $\rho_{{\rm bo},t}(a,b)$ converges to $\rho^{\rm lim}_{{\rm bo},t}(a,b)$*

$$\rho^{\rm lim}_{{\rm bo},t}(a,b)=\frac{\mathcal{N}\left(\begin{bmatrix}\sigma^{-1}_{\tau_0^2+\tau^2_{\rm add}}(a)\\ \sigma^{-1}_{\hat{\tau}_t^2}(b)\end{bmatrix}\middle|\mathbf{0}_2,\Sigma_{bo,t}\right)}{|\sigma'_{\tau_0^2+\tau^2_{\rm add}}(\sigma^{-1}_{\tau_0^2+\tau^2_{\rm add}}(a))||\sigma'_{\hat{\tau}_t^2}(\sigma^{-1}_{\hat{\tau}_t^2}(b))|} \quad (131)$$

where this time

$$\Sigma_{{\rm bo},t}=\begin{bmatrix}q^\star_{\rm bo} & m^\star_t\\ m^\star_t & q^\star_t\end{bmatrix} \quad (132)$$

To prove Lemma D.1, the main idea is to observe that, as with $f_\star$, to compute the density we need the covariance matrix

$$\frac{1}{d}\begin{pmatrix}\hat{\boldsymbol{\theta}}^\top_{\rm bo}\Omega\hat{\boldsymbol{\theta}}_{\rm bo} & \hat{\boldsymbol{\theta}}^\top_{\rm bo}\Omega\boldsymbol{\theta}_t\\ \hat{\boldsymbol{\theta}}^\top_{\rm bo}\Omega\hat{\boldsymbol{\theta}}_t & \hat{\boldsymbol{\theta}}^\top_t\Omega\boldsymbol{\theta}_t\end{pmatrix} \quad (133)$$

The diagonal terms are $q^\star_{\rm bo},q^\star_t$ respectively by definition. We then just need to compute the overlap $m_{{\rm bo},t}=\frac{1}{d}\hat{\boldsymbol{\theta}}^\top_{\rm bo}\Omega\hat{\boldsymbol{\theta}}_t$. Our goal is to prove that $m_{{\rm bo},t}=m^\star_t$,

However, using the Nishimori identity from statistical physics, for any vector $\boldsymbol{z}(\mathcal{D})$ that can depend on the training data, we have

$$\mathbb{E}_{\mathcal{D}}\left(\hat{\theta}^\top_{\rm bo}\boldsymbol{z}(\mathcal{D})\right)=\mathbb{E}_{\boldsymbol{w}_\star,\mathcal{D}}\left(\boldsymbol{w}_\star^\top\boldsymbol{z}(\mathcal{D})\right) \quad (134)$$

Equation 134 is just an application of Bayes formula. In particular, if we take $\boldsymbol{z}(\mathcal{D})=\hat{\boldsymbol{\theta}}_t$, we obtain that

$$\mathbb{E}_{\mathcal{D}}(\hat{\theta}^\top_{\rm bo}\hat{\boldsymbol{\theta}}_t)=\mathbb{E}_{\boldsymbol{w}_\star,\mathcal{D}}(\boldsymbol{w}_\star^\top\hat{\boldsymbol{\theta}}_t) \quad (135)$$

and we see that in expectation, $\mathbb{E}_{\mathcal{D}}(m_{{\rm bo},t})=\mathbb{E}_{\boldsymbol{w}_\star,\mathcal{D}}(m^\star_t)$. We already know that the right-hand side of the equality self-averages, i.e $\lim_{d\to\infty}\mathbb{E}_{\boldsymbol{w}_\star,\mathcal{D}}(m^\star_t)=m^\star_t$. It remains to show that the left-hand side also self-averages.

Lemma D.2 (Concentration of the overlap $m_{{\rm bo},t}$)**.**

$$\lim_{d\to\infty}\mathbb{E}\left[\left(\frac{\hat{\boldsymbol{\theta}}^\top_{\rm bo}\boldsymbol{\theta}_t}{d}\right)^2\right]=\lim_{d\to\infty}\mathbb{E}\left[\frac{\hat{\boldsymbol{\theta}}^\top_{\rm bo}\boldsymbol{\theta}_t}{d}\right]^2 \quad (136)$$

Proof. The proof again uses Nishimori identity.

$$\mathbb{E}\left[\left(\frac{\hat{\boldsymbol{\theta}}_{\rm bo}^\top \boldsymbol{\theta}_t}{d}\right)^2\right] = \mathbb{E}\left[\left(\frac{\hat{\boldsymbol{\theta}}_{\rm bo}^\top \boldsymbol{\theta}_t}{d}\right)\left(\frac{\hat{\boldsymbol{\theta}}_{\rm bo}^\top \boldsymbol{\theta}_t}{d}\right)\right] \tag{137}$$

$$= \mathbb{E}_{\mathcal{D}}\left[\left(\frac{\mathbb{E}_{\hat{\boldsymbol{\theta}}|\mathcal{D}}\hat{\boldsymbol{\theta}}^\top \boldsymbol{\theta}_t}{d}\right)\left(\frac{\mathbb{E}_{\hat{\boldsymbol{\theta}}|\mathcal{D}}\hat{\boldsymbol{\theta}}\cdot \boldsymbol{\theta}_t}{d}\right)\right] \tag{138}$$

$$= \mathbb{E}_{\mathcal{D}}\mathbb{E}_{\hat{\boldsymbol{\theta}}_1,\hat{\boldsymbol{\theta}}_2|\mathcal{D}}\left[\left(\frac{\hat{\boldsymbol{\theta}}_1^\top \boldsymbol{\theta}_t}{d}\right)\left(\frac{\hat{\boldsymbol{\theta}}_2\cdot \boldsymbol{\theta}_t}{d}\right)\right] \tag{139}$$

$$= \mathbb{E}_{\mathcal{D},\boldsymbol{w}_\star}\left[\left(\frac{\boldsymbol{w}_\star^\top \boldsymbol{\theta}_t}{d}\right)\left(\frac{\mathbb{E}_{\hat{\boldsymbol{\theta}}|\mathcal{D}}\hat{\boldsymbol{\theta}}\cdot \boldsymbol{\theta}_t}{d}\right)\right] \tag{140}$$

$$= \mathbb{E}_{\mathcal{D},\boldsymbol{w}_\star}\left[\left(\frac{\boldsymbol{w}_\star^\top \boldsymbol{\theta}_t}{d}\right)\left(\frac{\boldsymbol{w}_{\rm bo}^\top \boldsymbol{\theta}_t}{d}\right)\right] \tag{141}$$

Then, from Cauchy-Schwartz we have

$$\mathbb{E}\left[\left(\frac{\hat{\boldsymbol{\theta}}_{\rm bo}^\top \boldsymbol{\theta}_t}{d}\right)^2\right]^2 \leq \mathbb{E}\left[\left(\frac{\hat{\boldsymbol{\theta}}_{\rm bo}^\top \boldsymbol{\theta}_t}{d}\right)^2\right]\mathbb{E}\left[\left(\frac{\boldsymbol{w}_\star^\top \boldsymbol{\theta}_t}{d}\right)^2\right] \tag{142}$$

$$\mathbb{E}\left[\left(\frac{\hat{\boldsymbol{\theta}}_{\rm bo}^\top \boldsymbol{\theta}_t}{d}\right)^2\right] \leq \mathbb{E}\left[\left(\frac{\boldsymbol{w}_\star^\top \boldsymbol{\theta}_t}{d}\right)^2\right] \tag{143}$$

and as $d \to \infty$, we can use the concentration of the right hand side to $m_t^\star$ to obtain

$$\lim_{d\to\infty}\mathbb{E}\left[\left(\frac{\hat{\boldsymbol{\theta}}_{\rm bo}^\top \boldsymbol{\theta}_t}{d}\right)^2\right] \leq (m_t^\star)^2 \tag{144}$$

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

$$\lim_{d\to\infty}\mathbb{E}\left[\left(\frac{\hat{\boldsymbol{\theta}}_{\rm bo}^\top \boldsymbol{\theta}_t}{d}\right)^2\right] = (m_t^\star)^2 \tag{145}$$

□

We have thus shown that $m_t^\star = m_{{\rm bo},t}$, proving Lemma D.1.

Computing the conditional variance Fix now the confidence $\hat{f}_t = \ell$, the local field of the estimator t is $\nu_t := \sigma_{\hat{\tau}_t}^{-1}(\ell)$. The conditional distribution of the Bayes-optimal local field $\lambda_{\rm bo}$ is a Gaussian $\mathcal{N}(m_t^\star/q_t^\star \nu_t, q_{\rm bo}^\star - m_t^{\star\,2}/q_t^\star)$. Thus,

$$\mathbb{E}\left(\hat{f}_{\rm bo}^2|\hat{f}=\ell\right) = \int \mathrm{d}z \sigma_{v_{\rm bo}^\star+\tau_0^2+\tau_{\rm add}^2}(z)^2 \mathcal{N}(z|m_t^\star/q_t^\star\nu_t, q_{\rm bo}^\star - m_t^{\star\,2}/q_t^\star) \tag{146}$$

The last step to prove eq. (18) is to show that $\mathbb{E}\left(\hat{f}_{\rm bo}|\hat{f}_t=\ell\right) = \ell - \Delta_\ell$:

$$\begin{aligned}\mathbb{E}\left(\hat{f}_{\rm bo}|\hat{f}_t=\ell\right) &= \int \sigma_{\tau_0^2+\tau_{\rm add}^2+\hat{v}_{\rm bo}^*}(z)\mathcal{N}(z|m_t^\star/q_t^\star\nu_t, q_{\rm bo}^\star - m_t^{\star\,2}/q_t^\star)\mathrm{d}z \\ &= \sigma_{\tau_0^2+\tau_{\rm add}^2+\hat{v}_{\rm bo}^*+q_{\rm bo}^\star - m_t^{\star\,2}/q_t^\star}(m_t^\star/q_t^\star\nu_t) \\ &= \sigma_{\tau_0^2+\tau_{\rm add}^2+\rho-m_t^{\star\,2}/q_t^\star}(m_t^\star/q_t^\star\nu_t) = \mathbb{E}\left(f_\star|\hat{f}_t=\ell\right) = \ell - \Delta_\ell\end{aligned}$$

since, due to Bayes optimality, $\hat{v}_{\rm bo}^* = \rho - q_{\rm bo}^\star$.

E Additional numerical evaluations

E.1 Additional setting : $\tau_0^2 = 0, n/d = 10.0$

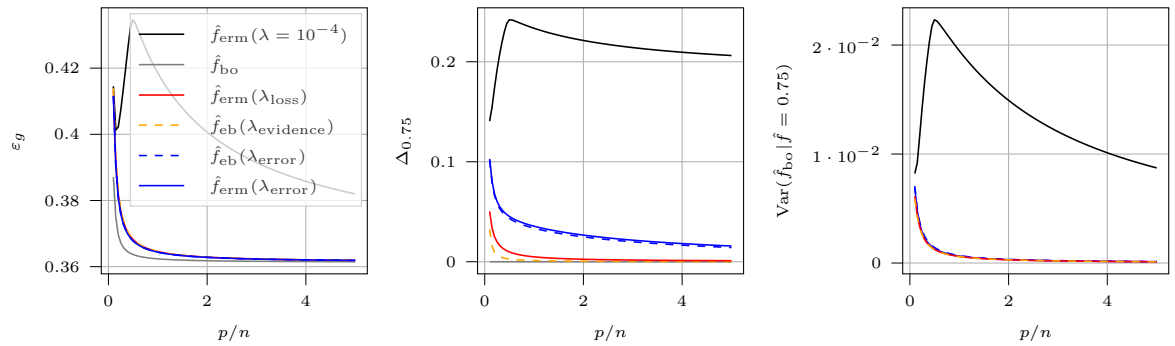

Figure 4: (**Left**) Test error of the estimators as a function of $1/\alpha$ in the setting of Section E.1 : $\|\boldsymbol{\theta}_\star\|^2 = 1, \tau_0^2 = 0, n/d = 10$. (**Middle**) Calibration of the estimators. (**Right**) Variance of $\hat{f}_{\text{bo}}$ conditioned on $\hat{f} = 0.75$ for the different estimators.

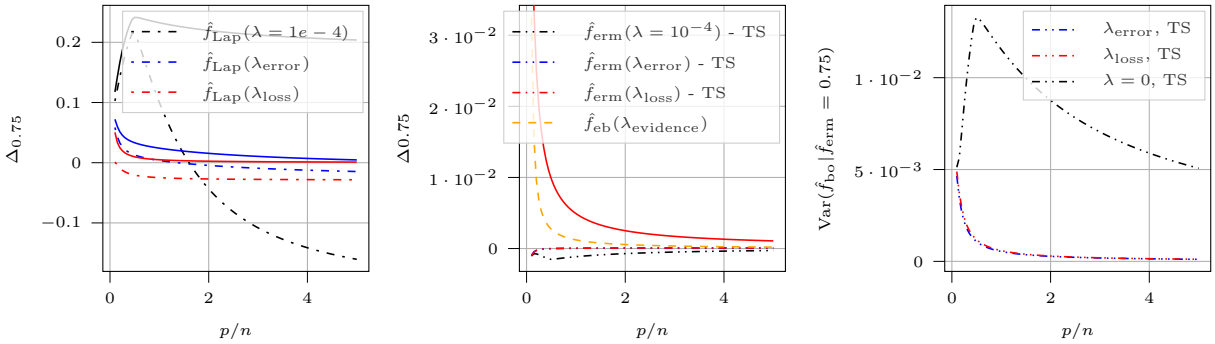

Figure 5: (**Left**) Calibration of $\hat{f}_{\text{Lap}}$ and $\hat{f}_{\text{erm}}$ in the setting of Section E.1. (**Middle**) Calibration of $\hat{f}_{\text{erm}}$ after temperature scaling. Curves for λ_{error} and λ_{loss} are indistinguishable on the plot. (**Right**) Variance of $\hat{f}_{\text{bo}}$ conditioned on the confidence of temperature scaling.

In this section, we consider a setting where $\tau_0^2 = 0.0$. This allows us to consider a setting where the test error of our estimators will be lower, as we reduce the noise in the teacher and increase the amount of training data. This is confirmed by the first panel of Figure 4, where the test error of the estimators is smaller than in Figure 1. Moreover, compared to the setting of Figure 1, the curves for $\hat{f}_{\text{erm}}(\lambda_{\text{error}})$, $\hat{f}_{\text{erm}}(\lambda_{\text{loss}})$ and $\hat{f}_{\text{eb}}(\lambda_{\text{evidence}})$ are much closer. Looking at the second panel, we note that as before, doing ERM with λ_{loss} or empirical-Bayes with $\lambda_{\text{evidence}}$ yields the best calibration. However, the calibration curves $\Delta_{0.75}$ do not exhibit the *double descent*-like behaviour shown in Figure 1. On Figure 5, we see the calibration of $\hat{f}_{\text{Lap}}$ (left plot) and temperature scaling (center). We see that in this setting, $\hat{f}_{\text{Lap}}$ yields underconfident estimators for p/n large enough. On the other hand, temperature scaling yields a well-calibrated estimator, whether we apply it on $\hat{f}_{\text{erm}}(\lambda = 0)$ or $\hat{f}_{\text{erm}}(\lambda_{\text{error}})$

E.2 Additional setting 2 : $\tau_0^2 = 0, n/d = 20, \|\theta_*\|^2 = 50$

In the previous plots, we defined $\theta_* = 1$. This is of course not a limitation of our model and we can assume any norm for the teacher. In this section, we will assume $\|\theta_*\|^2 = 50$. This allows us to significantly reduce the noise in the data. Indeed, as $\|\theta_*\|^2 \to \infty$, the label becomes deterministic in the input. As before, figures 6 and 5 show the test error, calibration and variance for the different estimators. In the left panel of Figure 7, we observe that $\hat{f}_{\text{Lap}}$ with $\lambda \in \{\lambda_{\text{error}}, \lambda_{\text{loss}}, 10^{-4}\}$ systematically underconfident for p/n large enough. As with the previous settings, we also note that $\hat{f}_{\text{erm}}(\lambda_{\text{error}})$ used in combination with temperature scaling is the most competitive estimator as it yields very good test error and calibration.

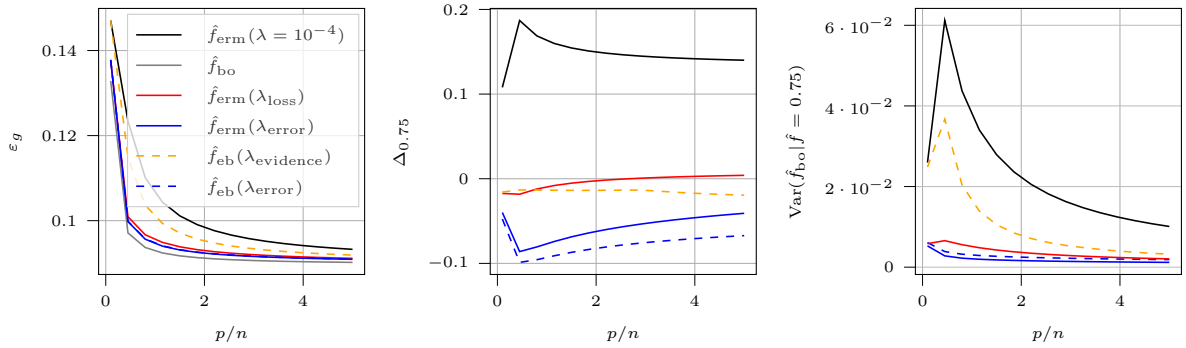

Figure 6: (**Left**) Test error of the estimators as a function of $1/\alpha$ in the setting described in section E.2. (**Middle**) Calibration of the estimators. (**Right**) Variance of $\hat{f}_{\text{bo}}$ conditioned on $\hat{f} = 0.75$ for the different estimators.

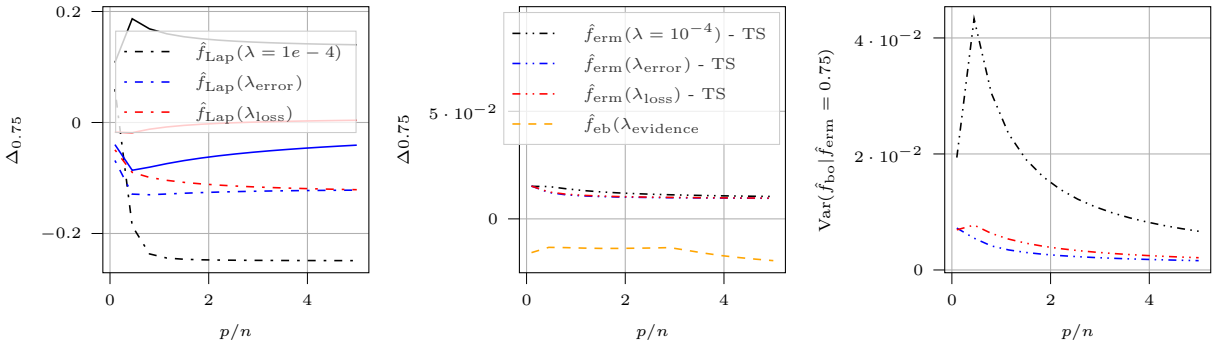

Figure 7: (**Left**) Calibration of $\hat{f}_{\text{Lap}}$ and $\hat{f}_{\text{erm}}$ with the setting described in section E.2. (**Middle**) Calibration of $\hat{f}_{\text{erm}}$ after temperature scaling. Solid red line is $\hat{f}_{\text{erm}}(\lambda_{\text{loss}})$ before temperature scaling. (**Right**) Variance of $\hat{f}_{\text{bo}}$ conditioned on the confidence of temperature scaling.