Fluctuations, Bias, Variance & Ensemble of Learners: Exact Asymptotics for Convex Losses in High-Dimension

Bruno Loureiro [1 *] **Cédric Gerbelot** [2 *] **Maria Refinetti** [2] **Gabriele Sicuro** [3 *] **Florent Krzakala** [1]

Abstract

From the sampling of data to the initialisation of parameters, randomness is ubiquitous in modern Machine Learning practice. Understanding the statistical fluctuations engendered by the different sources of randomness in prediction is therefore key to understanding robust generalisation. In this manuscript we develop a quantitative and rigorous theory for the study of fluctuations in an ensemble of generalised linear models trained on different, but correlated, features in high-dimensions. In particular, we provide a complete description of the asymptotic joint distribution of the empirical risk minimiser for generic convex loss and regularisation in the high-dimensional limit. Our result encompasses a rich set of classification and regression tasks, such as the lazy regime of over-parametrised neural networks, or equivalently the random features approximation of kernels. While allowing to study directly the mitigating effect of ensembling (or bagging) on the bias-variance decomposition of the test error, our analysis also helps disentangle the contribution of statistical fluctuations, and the singular role played by the interpolation threshold that are at the roots of the "double-descent" phenomenon.

1. Introduction

Randomness is ubiquitous in Machine Learning. It is present in the data (e.g., noise in acquisition and annotation), in commonly used statistical models (e.g., random features (Rahimi & Recht, 2007)), or in the algorithms used to train them (e.g., in the choice of initialisation of weights of neural networks (Narkhede et al., 2021), or when sampling a mini-batch in Stochastic Gradient Descent (Bottou, 2012)). Strikingly, fluctuations associated to different sources of randomness can have a major impact in the generalisation performance of a model. For instance, this is the case in least-squares regression with random features, where it has been shown (Geiger et al., 2020; D'Ascoli et al., 2020; Jacot et al., 2020) that the variance associated with the random projections matrix is responsible for poor generalisation near the interpolation peak (Advani & Saxe, 2017; Spigler et al., 2019; Belkin et al., 2020). As a consequence, this *double-descent* behaviour can be mitigated by averaging over a large *ensemble* of learners, effectively suppressing this variance. Indeed, considering an ensemble (sometimes also refereed to as a committee (Drucker et al., 1994)) of independent learners provide a natural framework to study the contribution of the variance of prediction in the estimation accuracy. In this manuscript we leverage this idea to provide an exact asymptotic characterisation of the statistics of fluctuations in empirical risk minimisation with generic convex losses and penalties in high-dimensional models. We focus on the case of synthetic datasets, and we apply our results to random feature learning in particular.

Setting Let $(\boldsymbol{x}^\mu, y^\mu) \in \mathbb{R}^d \times \mathcal{Y}$, $\mu \in [n] \coloneqq \{1, \dots, n\}$, denote a labelled data set composed of n independent samples from a joint density $p(\boldsymbol{x}, y)$ (e.g., $\mathcal{Y} = \{-1, 1\}$ for a binary classification problem). In this manuscript we are interested in studying an ensemble of K parametric predictors, each of them depending on a vector of parameters $\boldsymbol{w}_k \in \mathbb{R}^p$, $k \in [K]$, and independently trained on the dataset $\{(\boldsymbol{x}^\mu, y^\mu)\}_{\mu \in [n]}$. Note that even if the vectors of parameters $\{\boldsymbol{w}_k\}_{k \in [K]}$ are trained independently, they correlate through the training data. Statistical fluctuations in the learnt parameters can then arise for different reasons. For instance, a common practice is to initialise the parameters randomly during optimisation, which will induce statistical variability between the different predictors. Alternatively, each predictor could be trained on a subsample of the data, as it is commonly done in bagging (Breiman, 1996). The statistical model can also be inherently stochastic, e.g., the

[*]Equal contribution [1]IdePHICS lab, École Polytechnique Fédérale de Lausanne, Rte Cantonale 1015 Lausanne, Switzerland [2]Laboratoire de Physique de l'ENS, Université PSL, CNRS, Sorbonne Université, 24 Rue Lhomond, 75005 Paris, France [3]Department of Mathematics, King's College London, Strand WC2R 2LS, London, United Kingdom. Correspondence to: Bruno Loureiro <bruno.loureiro@epfl.ch>, Gabriele Sicuro <gabriele.sicuro@kcl.ac.uk>.

Proceedings of the 39^{th} International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022. Copyright 2022 by the author(s).

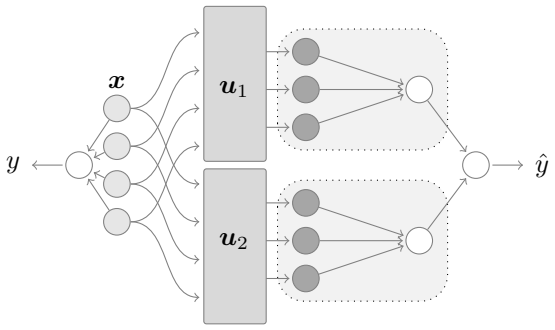

Figure 1. Pictorial representation of the model considered in the paper for $K = 2$. Two learners with the same architecture (in gray) receive a correlated input generated from the same vector $\boldsymbol{x} \sim \mathcal{N}(\mathbf{0}_d, \boldsymbol{I}_d)$. The output $\hat{y}$ is an average of their outputs. While the study of an ensemble of learners is already interesting *per se*, it is also pivotal to study the fluctuation between learners, and the error steaming from the difference in the weights in random features and lazy training.

random features approximation for kernel methods (Rahimi & Recht, 2007). Finally, the predictors could also be jointly trained, e.g., coupling them through the loss or penalty as it is done in boosting (Schapire, 1990).

Our goal in this work is to provide a sharp characterisation of the statistical fluctuations of the ensemble of parameters $\{\boldsymbol{w}_k\}_{k\in[K]}$ in a particular, mathematically tractable, class of predictors: *generalised linear models* (GLMs),

$$\hat{y}(\boldsymbol{x}) = \hat{f}\left(\frac{\hat{\boldsymbol{w}}_1^\top \boldsymbol{u}_1(\boldsymbol{x})}{\sqrt{p}}, \dots, \frac{\hat{\boldsymbol{w}}_K^\top \boldsymbol{u}_K(\boldsymbol{x})}{\sqrt{p}}\right) \tag{1}$$

where $\boldsymbol{u}_k : \mathbb{R}^d \to \mathbb{R}^p$, $k \in [K]$ is an ensemble of possibly correlated features and $\hat{f}\colon \mathbb{R}^K \to \mathcal{Y}$ is an activation function. For most of this work, we discuss the case in which the predictors are *independently trained* through regularised empirical risk minimisation:

$$\hat{\boldsymbol{w}}_k = \underset{\boldsymbol{w}\in\mathbb{R}^p}{\operatorname{argmin}} \left[\frac{1}{n}\sum_{\mu=1}^{n} \ell\left(y^\mu, \frac{\boldsymbol{w}^\top \boldsymbol{u}_k(\boldsymbol{x}^\mu)}{\sqrt{p}}\right) + \frac{\lambda}{2}\|\boldsymbol{w}\|_2^2\right] \tag{2}$$

with a convex loss function $\ell\colon \mathcal{Y}\times\mathbb{R} \to \mathbb{R}$ (e.g., the logistic loss) and ridge penalty whose strength is given by $\lambda \in \mathbb{R}^+$. However, our analysis also includes the case in which the learners are jointly trained with a generic convex penalty. This case will be further discussed in Sec. 4. In what follows we will also concentrate in the random features case where $\boldsymbol{u}_k(\boldsymbol{x}) = \phi(\boldsymbol{F}_k\boldsymbol{x})$ with $\phi : \mathbb{R} \to \mathbb{R}$ an activation function acting component-wise and $\boldsymbol{F}_k \in \mathbb{R}^{p\times d}$ a family of independently sampled random matrices. Besides being an efficient approximation for kernels (Rahimi & Recht, 2007), random features are often studied as a simple model for neural networks in the lazy and neural tangent kernel regimes of deep neural networks (Chizat et al., 2019; Jacot et al., 2018), in which case the matrices $\boldsymbol{F}_k$ correspond to different random initialisation of hidden-layer weights. Moreover, the random features model displays some of the exotic behaviours of high-dimensional overparametrised models, such as double-descent (Mei & Montanari, 2021; Gerace et al., 2020) and benign overfitting (Bartlett et al., 2020), therefore providing an ideal playground to study the interplay between fluctuations and overparametrisation. A broader class of features maps is also discussed in Sec. 4.

To provide an exact characterisation of the statistics of the estimators in eq. (2), we shall assume data is generated from a target

$$y = f_0\left(\frac{\boldsymbol{\theta}^\top \boldsymbol{x}}{\sqrt{d}}\right) \tag{3}$$

with $f_0 : \mathbb{R} \to \mathcal{Y}$ and $\boldsymbol{\theta} \sim \mathcal{N}(\mathbf{0}_d, \rho\boldsymbol{I}_d)$, $\rho \in \mathbb{R}_0^+$, $\boldsymbol{I}_d$ d-dimensional identity matrix. The dataset is then constructed generating *i.i.d.* n vectors $\boldsymbol{x}^\mu \sim \mathcal{N}(\mathbf{0}_d, \boldsymbol{I}_d)$, $\mu \in [n]$.

An illustration summary of the setting considered here in given in Figure 1. Note that such architecture can be interpreted as a two-layer tree neural network, also known in some contexts as the *tree-committee* or *parity machine* (Schwarze & Hertz, 1992). Although our focus on generalised linear models might appear restrictive, the investigation of this hypothesis class has proven fruitful in the current understanding of overparametrisation in learning, capturing the phenomenology of more complex architectures which is the ultimate goal of our contribution. Indeed, the popularity of the considered setting in theoretical works is justified by its connection with the lazy training limit of neural networks (Chizat et al., 2019; Jacot et al., 2018) and to kernel methods more generally, together with the fact that it displays a similar phenomenology to many non-convex networks (for instance, benign overfitting (Bartlett et al., 2020)). From this perspective, our analysis highlights the role played by overfitting the fluctuations due to random initialisation in the degradation of the generalisation of lazy deep neural networks (Geiger et al., 2020).

Main contributions The results in this manuscript can be listed as follows.

• We provide a sharp asymptotic characterisation of the joint statistics of the ensemble of empirical risk minimisers $\{\hat{\boldsymbol{w}}_k\}_{k\in[K]}$ in the high-dimensional limit where $p, n \to +\infty$ with n/p kept constant, for any convex loss and penalty. In particular, we show that the pre-activations $\{\hat{\boldsymbol{w}}_k^\top \boldsymbol{u}_k\}_{k\in[K]}$ are jointly Gaussian, with sufficient statistics obeying a set of explicit closed-form equations. Note that the analysis of ensembling with non-square losses is out of the grasp of the most commonly adopted theoretical tools (e.g., random matrix theory). Therefore, our proof method based on recent progress on Approximate Message Passing techniques (Berthier et al., 2020; Gerbelot & Berthier, 2021)

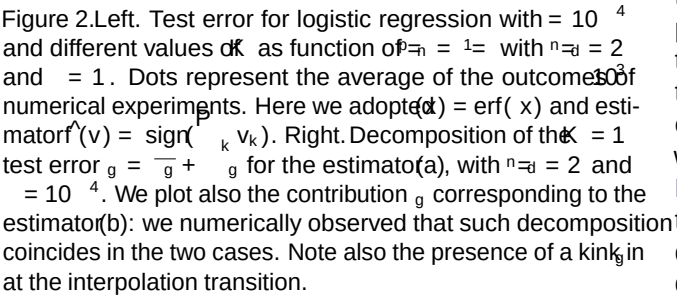

Figure 2. Left. Test error for logistic regression with $= 10^{-4}$ and different values of K as function of $= = ^{-1} =$ with $= = 2$ and $= 1$. Dots represent the average of the outcomes of numerical experiments. Here we adopted $(x) = \mathrm{erf}(x)$ and estimator $\hat{f}(v) = \mathrm{sign}(\sum_k v_k)$. Right. Decomposition of the $K = 1$ test error $_g = \bar{}_g + \tilde{}_g$ for the estimator (a), with $= = 2$ and $= 10^{-4}$. We plot also the contribution $_g$ corresponding to the estimator (b): we numerically observed that such decomposition coincides in the two cases. Note also the presence of a kink in $_g$ at the interpolation transition.

is of independent interest. Different versions of our theorem are discussed throughout the manuscript. First, in Sec. 2 for the particular case of independently trained learners on random features (Theorem 2.1). Later, in Sec. 4 for the general case of jointly trained learners on correlated Gaussian covariates (Theorem 4.1).

• We discuss the role played by uctuations in the non-monotonic behaviour of the generalisation performance of interpolators (a.k.a. double-descent behaviour). In particular —as discussed in (Geiger et al., 2020; d'Ascoli et al., 2021) for the ridge case— the interpolation peak arises from the model over tting the particular realisation of the random weights. We show the test error can be decomposed $_g(K = 1) = \bar{}_g + \tilde{}_g$ in terms of a uctuation-free term $\bar{}_g$ and a uctuation term $\tilde{}_g$ responsible for the double-descent behavior, see Fig. 2 for the case of max-margin classi cation.

• In the context of classi cation, we discuss how majority vote and score averaging two popular ensembling procedures, compare in terms of generalisation performance. More speci cally, we show that in the setting we study score averaging consistently outperforms the majority vote predictor. However, for a large number of learners K 1 these two predictors agree, see Fig. 6 (right).

• Finally, we discuss how ensembling can be used as a tool for uncertainty quanti cation. In particular, we connect the correlation between two learners to the probability of disagreement, and show that it decreases with overparametrisation, see Fig. 6 (left). We provide a full characterisation of the joint probability density of the con dence score between two independent learners, see Fig. 5.

Related works The idea of reducing the variance of a predictor by averaging over independent learners is quite old in Machine Learning (Hansen & Salamon, 1990; Perrone & Cooper, 1993; Perrone, 1994; Krogh & Vedelsby, 1995; Krogh & Sollich, 1997), and a variety of ensembling methods have been proposed (Opitz & Maclin, 1999). In a very inspiring work, Geiger et al. (2020) carried out an extensive series of experiments in order to shed light on the generalisation properties of neural networks, and reported many observations and empirical arguments about the role of the variance due to the random initialisation of the weights in the double-descent curve using an ensemble of learners. This was a major motivation for the present work. Closest to our setting is the work of Neal et al. (2018); D'Ascoli et al. (2020); Jacot et al. (2020) which disentangles the various sources of variance in the process of training deep neural networks. Indeed, here we adopt the model de ned by D'Ascoli et al. (2020), and provide a rigorous justi cation of their results for the case of ridge regression. A slightly ner decomposition of the variance in terms of the different sources of randomness in the problem was later proposed by Adlam & Pennington (2020a). Lin & Dobriban (2021) show that such decomposition is not unique, and can be more generally understood from the point of view of the analysis of variance (ANOVA) framework. Interestingly, subsequent papers were able to identity a series of triple (and more) descent, e.g., (d'Ascoli et al., 2021; Adlam & Pennington, 2020b; Chen et al., 2020). (LeJeune et al., 2020) studies ensembles of ordinary least-squares learned from subsamples of a common data matrix, and shows its equivalence to an implicit ridge regularization.

The Random Features (RF) model was introduced in the seminal work of Rahimi & Recht (2007) as an ef cient approximation for kernel methods. Drawing from early ideas of Karoui (2010), Pennington & Worah (2017) showed that the empirical distribution of the Gram matrix of RF is asymptotically equivalent to a linear model with matched second statistics, and characterised in this way memorisation with RF regression. The learning problem was rst analysed by Mei & Montanari (2021), who provided an exact asymptotic characterisation of the training and generalisation errors of RF regression. This analysis was later extended to generic convex losses by Gerace et al. (2020) using the heuristic replica method, and later proved by Dhifallah & Lu (2020) using convex Gaussian inequalities.

The aforementioned asymptotic equivalence between the RF model and a Gaussian model with matched moments

has been named the Gaussian Equivalence Principle (GEP) (Goldt et al., 2020). Rigorous proofs in the memorisation and learning setting with square loss appeared in (Pennington & Worah, 2017; Mei & Montanari, 2021), and for general convex penalties in (Goldt et al., 2021; Hu & Lu, 2020). Goldt et al. (2021) and Loureiro et al. (2021b) provided extensive numerical evidence that the GEP holds for more generic feature maps, including features stemming from trained neural networks.

Most of the previously mentioned works deriving exact asymptotics for the RF model in the proportional limit use either Random Matrix Theory techniques or Convex Gaussian inequalities. While these tools have been recently used in many different contexts, they ultimately fall short when considering an ensemble of predictors with generic convex loss and regularisation, along with structured design matrices. Therefore, to prove the results herein we employ an Approximate Message Passing (AMP) proof technique (Bayati & Montanari, 2011a; Donoho & Montanari, 2016), leveraging on recently introduced progresses in (Loureiro et al., 2021b; Gerbelot & Berthier, 2021) which enables to capture the full complexity of the problem and obtain the asymptotic joint distribution of the ensemble of predictors.

2. Learning with an ensemble of random features

In this section give a first formulation of our main result, namely the exact asymptotic characterisation of the statistics of the ensembling estimator introduced in (1). We prove that, in the proportional high dimensional limit, the statistics of the arguments of the activation function in eq. (1) is simply given by a multivariate Gaussian, whose covariance matrix we can completely specify. This result holds for any convex loss, any convex regularisation, and for all models of generative networks $u_k : \mathbb{R}^d \to \mathbb{R}^p$, as we will show in full generality in Sec. 4. However, for simplicity, in this section and in the following we focus on the setting described in Sec. 1, in which the statistician averages over an independent ensemble of random features, i.e. $u_k(x) = \phi(F_k x)$. In this case, our result can be formulated as follows:

Theorem 2.1 (Simplified version). *Assume that in the high-dimensional limit where $d, p, n \to +\infty$ with $\alpha := n/p$ and $\gamma := d/p$ kept $\Theta(1)$ constants, the Wishart matrix $\frac{1}{p} F^\intercal F$ has a well-defined asymptotic spectral distribution. Then in this limit, for any pseudo-Lispchitz function of order 2 $\psi : \mathbb{R} \times \mathbb{R}^K \to \mathbb{R}$, we have*

$$\mathbb{E}_{(x,y)} \psi\left(y, \frac{w_1^\intercal u_1}{\sqrt{p}}, \dots, \frac{w_K^\intercal u_K}{\sqrt{p}}\right) \to \mathbb{E}_{(\lambda,\nu)}[\psi(f_0(\lambda), \nu)], \tag{4}$$

where $(\lambda, \nu) \in \mathbb{R}^{K+1}$ is a jointly Gaussian vector $(\lambda, \nu) \sim \mathcal{N}(0_{K+1}, \Omega)$ with covariance

$$\Omega = \begin{pmatrix} \rho & m\, 1_K^\intercal \\ m\, 1_K & Q \end{pmatrix}, \qquad Q := (q_0 - q_1) I_K + q_1 1_{K,K}, \tag{5}$$

with $1_{K,K} \in \mathbb{R}^{K \times K}$ and $1_K \in \mathbb{R}^K$ are a matrix and a vector of ones respectively. The entries of Ω are solutions of a set of self-consistent equations given in Corollary 4.3.

Note that pseudo-Lipschitz functions of order 2 roughly amounts to functions with polynomially bounded increment, see Definition B.6 in the Appendix. As discussed in the introduction, the asymptotic statistics of the single learner has been studied in (Gerace et al., 2020; Dhifallah & Lu, 2020; Loureiro et al., 2021b). Their result amounts to the analysis of the estimator solving the empirical risk minimisation problem in eq. (2) and it is recovered imposing $K = 1$ in the theorem above. For $K = 1$, $(\lambda, \nu) \in \mathbb{R}^2$ is jointly Gaussian with zero mean and covariance $\Omega = \begin{pmatrix} \rho & m \\ m & q_0 \end{pmatrix}$.

However, such result is not enough to quantify the correlation between different learners, induced by the training on the same dataset, which is required to compute, e.g., the test error associated with an ensembling predictor as in eq. (1). For example, in the simple case where $f_0(u) = u$ and $\hat{f}(v) = \frac{1}{K}\sum_k v_k$, the mean-squared error on the labels is given by $\epsilon_g = \mathbb{E}_{(x,y)}[(y - \hat{y}(x))^2] = \rho + (q_0 - q_1)K^{-1} + q_1 - 2m$, which crucially depends on the average correlation between two independent learners $q_1 := \frac{1}{p}\mathbb{E}[w_1^\intercal w_2]$. Our main result is precisely an exact asymptotic characterisation of this correlation in the proportional limit of the previous theorem. Once m, q_0 and q_1 have been determined, the generalisation error can be computed as

$$\epsilon_g := \mathbb{E}_{(x,y)}[\ell(y, \hat{y}(x))] \to \mathbb{E}_{(\lambda,\nu)}\left[\ell(f_0(\lambda), \hat{f}(\nu))\right] \tag{6}$$

for any error measure $\ell : Y \times Y \to \mathbb{R}^+$.

Suppose now that

$$\hat{f}(v) \to \hat{f}_0\left(\frac{1}{K}\sum_k v_k\right) \tag{7}$$

for some $\hat{f}_0 : \mathbb{R} \to Y$ activation function of the single learner. In this case we can introduce the random variable $\hat{\nu} \overset{d}{=} \lim_{K\to+\infty} \frac{1}{K}\sum_k \nu_k$. It is not difficult to see that the joint probability $p(\lambda, \hat{\nu}) \to \mathcal{N}(0_2, \hat{\Omega})$ where $\hat{\Omega} = \begin{pmatrix} \rho & m \\ m & q_1 \end{pmatrix}$. This formally coincides with the joint distribution for the activation fields for $K = 1$ (Gerace et al., 2020), but with q_0 replaced by q_1. The smaller variance is due to the fact that the fluctuations of the activation fields are averaged out by the ensembling process. The test error in the $K \to +\infty$ limit is then

$$\bar{\epsilon}_g := \mathbb{E}_{(\lambda,\hat{\nu})}[\ell(f_0(\lambda), \hat{f}_0(\hat{\nu}))]; \tag{8}$$

so that the fluctuation contribution to the test error for $K = 1$

can be de ned as

$$_g := E_{(\ ;\ )}[\ (\ f_0(\ );\hat{f}_0(\ ))] \quad \overline{\ }_g: \qquad (9)$$

The term $_g$ is by de nition the contribution suppressed by ensembling and corresponds to the ambiguity introduced by Krogh & Vedelsby (1995) for the square loss. This contribution expresses the variance in the ensemble and it is responsible for the non-monotonic behaviour in the test error of interpolators, also known as the double-descent behavior.

3. Applications

We will consider now two relevant examples of separable losses, namely a ridge loss and a logistic loss. In both cases it is possible to derive the explicit expression of the training loss and generalisation error in terms of the elements of the correlation matrix introduced above.

3.1. Ridge regression

If we assume $f_0(x) = x$, $\hat{f}(v) = \frac{1}{K}\sum_k v_k$, and a quadratic loss of the type $`(y;x) = \frac{1}{2}(y \quad x)^2$, it is possible to write down simple recursive equations for q_0 and q_1 (see Appendix A.3). Taking $(\ y;\hat{y}) = (\ y \quad \hat{y})^2$, the generalisation error is easily computed as

$$_g = \ + \frac{q_0 \quad q_1}{K} + q_1 \quad 2m \xrightarrow{K\ !\ +1} \ + q_1 \quad 2m \quad \overline{\ }_g: \qquad (10)$$

Note that in this case the $\ !\ 0^+$ limit gives the minimum $`_2$-norm interpolator. In Fig. 3 we compare our theoretical prediction with numerical results for $= 10^{\ 6}$ and various values of K. It is evident that the divergence of the generalisation error at $= 1$ is only due to the divergence of q_0, whereas the contribution $\overline{\ }_g$, which is independent on q_0, is smooth everywhere. Alongside with the interpolation divergence, $_g = q_0 \quad q_1$ has an additional bump at $p=n = d=n$, which corresponds to the "linear peak" discussed by d'Ascoli et al. (2021). In the plot we present also the so-called kernel limit, corresponding to the limit $n=p = \ ! \ 0$ at $xed\ n=d$. An explicit manipulation (see Appendix A.3) shows that $q_1 = q_0 \quad q$ in this limit. This implies that in the kernel limit ^k_g does not depend on K, being equal to $^k_g \quad + q \quad 2m$. The generalisation error obtained in the kernel limit coincides with $\overline{\ }_g$ for $p > n$: this is expected as in $\overline{\ }_g$ the $uctuations$ amongst learners are averaged out, effectively recovering the cost obtained in the case of an in nite number of parameters.

3.2. Binary classi cation

Suppose now that we are considering a classi cation task such that $Y = f \ 1; 1g$. For this task we consider $f_0(x) = \text{sign}(x)$. A popular choice of loss in this classi cation task is the logistic loss,

$$`(y;x) = \ln \ \ 1 + e^{\ yx} \ ; \qquad (11)$$

although other choices, e.g, hinge loss, can be considered. Since both the logistic and hinge losses depend only on the margin $yw^{\mid}u$, the empirical risk minimiser for $\ !\ 0^+$ in both cases give the max-margin interpolator (Rosset et al., 2004). We write down the explicit saddle-point equations associated to the logistic and hinge loss in Appendix A.3, but we will focus our attention on the logistic case for the sake of brevity. For this choice of the loss, we obtained the values of m, q_0 and q_1 showed in Fig. 4. Using these values, a number of relevant questions can be addressed.

Figure 3. Left. Test error for ridge regression with $= 10^{\ 6}$ and different values of K as function of $p=n = \ 1= \ $ with $n=d = 2$ and $= 1$. Dots represent the average of the outcomes of 50 numerical experiments in which the parameters of the neurons are estimated using $\min(d;p) = 200$. Here we adopted $(x) = \text{erf}(\ x)$. Right. Decomposition of $_g = \overline{\ }_g + \ _g$ in the $K = 1$ case

Alignment of learners – Assuming that the predictor of the learner k is $\hat{y}_k(x) = \text{sign}(w_k^{\mid}u_k(x))$, in Fig. 6 (left) we estimate the probability that two learners give opposite classi cation. This is analytically given by

$$P[\hat{y}_1(x) \ 6= \hat{y}_2(x)] = P[\ _1 \ _2 < 0] = \frac{1}{\ }\arccos\frac{q_1}{q_0} \ : \qquad (12)$$

Note that by de nition the ratio $q_1=q_0$ is a cosine similarity between two learners in the norm induced by the feature space. Therefore, this provides an interesting interpretation of these suf cient statistics in terms of the probability of disagreement. In particular, as illustrated in Fig. 6 (left) over-parametrisation promotes agreement between the learners, therefore suppressing uncertainty. More generally, ensembling can be used as a technique for uncertainty estimation (Lakshminarayanan et al., 2017). In the context of logistic regression, the pre-activation to the sign function is often interpreted as a $con dence\ score$. Indeed, introducing the

logistic function $\varphi_k(x) = (1+\exp(-\sum_i^{1/2} w_{ki} u_k(x)))^{-1}$, it expresses the confidence of the kth classifier in associating $\hat{y}=1$ to the input x. Therefore, it is reasonable to ask how reliable is the logistic score as a confidence measure. For instance, what is the variance of the confidence among different learners? This can be quantified by the joint probability density $\rho(\varphi_1,\varphi_2) := \mathbb{E}_x[(\varphi_1 - \bar{\varphi}_1(x))(\varphi_2 - \bar{\varphi}_2(x))]$, which can be readily computed using our Theorem 2.1. Fig. 5 shows one example at fixed $\alpha = n/d$ and vanishing λ.

Figure 4. Analytical estimation of the order parameters r (left), q_0 and q_1 (right) in a classification task using logistic loss with ridge penalty with $\lambda = 10^{-4}$ at fixed $n/d = 2$ as function of p/n. In the inset, ratio q_1/q_0, quantifying the correlation between two learners. In all parameters the interpolation kink is clearly visible.

Figure 5. Joint probability density of the confidence scores $\varphi(x) = (1+\exp(-\sum_i^{1/2} w_i u_i(x)))^{-1}$ of two learners for $p/n = 0.13$.

Ensemble predictors – In the previous two points, we discussed how ensembling can be used as a tool to quantify fluctuations. However, ensembling methods are also used in practical settings in order to mitigate fluctuations, e.g., (Breiman, 1996). An important question in this context is: given an ensemble of predictors $\{\hat{f}_k\}_{k\in[K]}$, what is the best way of combining them to produce a point estimate? In our setting, this amounts to choosing the function $\hat{f}: \mathbb{R}^K \to Y$. Let us consider two popular choices for the estimator $\hat{f}$ in eq. (1) used in practice:

$$\text{(a)} \quad \hat{f}(v) = \operatorname{sign}\Big(\sum_k v_k\Big); \tag{13a}$$

$$\text{(b)} \quad \hat{f}(v) = \operatorname{sign}\Big(\sum_k \operatorname{sign}(v_k)\Big). \tag{13b}$$

In a sense, (a) provides an estimator based on the average of the output fields, whereas (b), which corresponds to a majority rule if K is odd (Hansen & Salamon, 1990), is a function of the average of the estimators of the single learners. For both choices of the estimator we use $\epsilon(\hat{y},y) = \mathbb{1}_{\hat{y}\neq y}$ to measure the test error. In Fig. 6 (right) we compare the test error obtained using (a) and (b) for $K=3$ with vanishing regularisation $\lambda = 10^{-4}$. It is observed that the estimator (a) has better performances than the estimator (b). As previously discussed, in this case logistic regression is equivalent to max-margin estimation, and in this case the error (a) can be intuitively understood in terms of a robust max-margin estimation obtained by averaging the margins associated to different draws of the random features. In the case (a) it is easy to show that the generalisation error takes the form

$$\epsilon_g = \frac{1}{\pi}\arccos\left(\frac{\sqrt{K}m}{\sqrt{(q_0 - q_1 + Kq_1)}}\right)^{K\to 1} \to \frac{1}{\pi}\arccos\left(\frac{m}{\sqrt{q_1}}\right) = \bar{\epsilon}_g. \tag{14}$$

This formula is in agreement with numerical experiments, see Fig. 2 (left). Unfortunately, we did not find a similar closed-form expression in case (b). However, we can observe that in the $K \to +\infty$ limit the generalisation error in case (a) coincides with the generalisation error in case (b), see Fig. 2 (right). By comparing with the results in Fig. 6 (left), it is evident that the benefit of ensembling in reducing the test error correlates with the tendency of learners to disagree, i.e., for small values of p/n, as stressed by Krogh & Vedelsby (1995). Finally, we observe a constant value of

Figure 6. Left. Probability that two learners give discordant predictions using logistic regression as function of $p/n = 1/\alpha$ with $n/d = 2$, $\gamma = 1$, and $\lambda = 10^{-4}$. Right. Test error for logistic regression using the estimators in eq. (13) and $K = 3$, with the same parameters. We adopted $\sigma(x) = \operatorname{erf}(x)$. We observe that the test error obtained using (a) is always smaller than the one obtained using (b). (Left and right) Dots represent the average of the outcomes of 10^3 numerical experiments.