The role of regularization in classification of high-dimensional noisy Gaussian mixture

Francesca Mignacco
Université Paris-Saclay, CNRS, CEA, Institut de physique théorique, 91191, Gif-sur-Yvette, France

Florent Krzakala
Laboratoire de Physique de lEcole normale supérieure, ENS, Université PSL, CNRS, Sorbonne Université, Université de Paris, F-75005 Paris, France

Yue M. Lu
John A. Paulson School of Engineering and Applied Sciences, Harvard University, Cambridge, MA 02138, USA

Lenka Zdeborová
Université Paris-Saclay, CNRS, CEA, Institut de physique théorique, 91191, Gif-sur-Yvette, France

We consider a high-dimensional mixture of two Gaussians in the noisy regime where even an oracle knowing the centers of the clusters misclassifies a small but finite fraction of the points. We provide a rigorous analysis of the generalization error of regularized convex classifiers, including ridge, hinge and logistic regression, in the high-dimensional limit where the number n of samples and their dimension d go to infinity while their ratio is fixed to $\alpha = n/d$. We discuss surprising effects of the regularization that in some cases allows to reach the Bayes-optimal performances. We also illustrate the interpolation peak at low regularization, and analyze the role of the respective sizes of the two clusters.

I. INTRODUCTION

High-dimensional statistics where both the dimensionality d, and number of samples n are large with a fixed ratio $\alpha = n/d$ has largely non-intuitive behaviour. A number of the associated statistical surprises are for example presented in the recent, yet already rather influential papers [1, 2] that analyze high-dimensional regression for rather simple models of data. The present paper subscribes to this line of work and studies high-dimensional classification in one of the simplest models considered in statistics — the mixture of two Gaussian clusters in d-dimensions, one of size ρn and the other $(1-\rho)n$ points. The labels reflect the memberships in the clusters. In particular, there are two centroids localized at $\pm\frac{\mathbf{v}^*}{\sqrt{d}} \in \mathbb{R}^d$, and we are given data points $\mathbf{x}_i, i = 1 \dots n$ generated as

$$\mathbf{x}_i = \frac{\mathbf{v}^*}{\sqrt{d}} y_i + \sqrt{\Delta}\mathbf{z}_i, \tag{1}$$

where both $\mathbf{z}_i$ and $\mathbf{v}^*$ have components taken in $\mathcal{N}(0,1)$. The labels $y_i \in \pm 1$ are generated randomly with a fraction ρ of $+1$ (and $1-\rho$ of -1). We focus on the high-dimensional limit where $n, d \to \infty$ while $\alpha = n/d$, ρ and Δ are fixed. The factor $\sqrt{d}$ in (1) is such that a classification better than random is possible, yet even the oracle-classifier that knows exactly the centroid $\frac{\mathbf{v}^*}{\sqrt{d}}$ only achieves a classification error bounded away from zero. We focus on ridge regularized learning performed by the empirical risk minimization of the loss:

$$\mathcal{L}(\mathbf{w}, b) = \sum_{i=1}^{n} \ell\left[y_i(\tfrac{1}{\sqrt{d}}\mathbf{x}_i^\top \mathbf{w} + b)\right] + \frac{1}{2}\lambda\|\mathbf{w}\|_2^2, \tag{2}$$

where $\mathbf{w}$ and b are, respectively, the weight vector and the bias to be learned, and λ is the tunable strength of the regularization. While our result holds for any convex loss function $\ell(.)$, we will mainly concentrate on the following classic ones: the square $\ell(v) = \frac{1}{2}(1-v)^2$, the logistic $\ell(v) = \log(1+e^{-v})$, and the hinge $\ell(v) = \max_v\{0, 1-v\}$. We shall also study the Bayes-optimal estimator, i.e. the one achieving the lowest possible test error on classification given the n samples $y_i, \mathbf{x}_i$ and the model, including the constants ρ and Δ. Crucially, the position of the centroid is *not known* and can only be estimated from the data.

Our contributions and related works — The unsupervised version of the problem is the standard Gaussian mixture modeling problem in statistics [3]. For the supervised model considered here, [4] recently computed rigorously the performance of the Bayes-optimal estimator (that knows the generative model of the data, but does not have access to the vector $\mathbf{v}^*$) for the case of equally sized clusters. We generalize these results for arbitrary cluster sizes to provide a baseline for the estimators obtained by empirical risk minimization.

The model was recently under investigation in a number of papers. In [5], the authors study the same data generative model in the particular case of equally sized clusters, and analyze *non-regularized losses* under the assumption that the data are not linearly separable. They conclude that in that case the square loss is a universally optimal loss function. Our study of the regularized losses shows that the performance of the non-regularized square loss can be easily, and drastically improved. [6] studied the logistic loss, again without regularization and for two

clusters of equal size, and derive the linear separability condition in this case.

As a first contribution, we provide rigorous closed-form asymptotic formulas for the generalization and training error in the noisy high-dimensional regime, for any convex loss $\ell(.)$, that include the effects of regularization, and for arbitrary cluster size. Our proof technique uses Gordon's inequality technique [7–9], as in [6]. We show through numerical simulations that the formulas are extremely accurate even at moderately small dimensions.

Secondly, we present a systematic investigation of the effects of regularization and of the cluster size, discussing in particular how far estimators obtained by empirical risk minimization fall short of Bayes-optimal one, with surprising conclusions where we illustrate the effect of strong and weak regularizations. In particular, when data are linearly separable, Rosset *et al.* [10] proves that all monotone non-increasing loss functions depending on the margin find a solution maximizing the margin. This is indeed exemplified in our model by the fact that for $\alpha < \alpha^*(\Delta, \rho)$ (the location of transition for linear separability) the hinge, and logistic losses converge to the same test error as the regularization tends to zero. This is related to the implicit regularization of gradient descent for the non-regularized minimization [11], and we discuss this in connection with the "double-descent" phenomenon that is currently the subject of intense studies [1, 12–15].

The existence of a sharp transition for perfect separability in the model, with and without bias, is interesting in itself. Recently [16] analyzed the maximum likelihood estimate (MLE) in high-dimensional logistic regression. While they analyzed Gaussian data (whereas we study Gaussian mixture) their results on the existence of the MLE being related to the separability of the data and having a sharp phase transition are of the same nature as ours, and similar to earlier works in statistical physics [17–19].

Finally, we note that the formulas proven here can also be obtained from the heuristic replica theory from statistical physics. Indeed, a model closely related to ours was studied in this literature [20, 21] and our rigorous solution thus provides a further example of a rigorous proof of a result obtained by this technique.

All these results show that the Gaussian mixtures model studied here allows to discuss, illustrate, and clarify in a unified fashion many phenomena that are currently the subject of intense scrutiny in high-dimensional statistics and machine learning.

II. MAIN THEORETICAL RESULTS

A. Performance of empirical risk minimization

Our first result is a rigorous analytical formula for the generalization classification error obtained by the empirical risk minimization of (2). Define q as the length of the vector $\mathbf{w}$ and m as its overlap with $\mathbf{v}^*$, both rescaled by the dimensionality d

$$q \equiv \frac{1}{d}\|\mathbf{w}\|_2^2, \quad m \equiv \frac{1}{d}\mathbf{v}^{*\top}\mathbf{w}, \tag{3}$$

then we have the following:

Theorem 1 (Asymptotics of q and m) *In the high dimensional limit when $n, d \to \infty$ with a fixed ratio $\alpha = n/d$, the length q and overlap m of the vector $\mathbf{w}$ obtained by the empirical risk minimization of (2) with a convex loss converge to deterministic quantities given by the unique fixed point of the system:*

$$m = \frac{\hat{m}}{\lambda + \hat{\gamma}}, \tag{4}$$

$$q = \frac{\hat{q} + \hat{m}^2}{(\lambda + \hat{\gamma})^2}, \tag{5}$$

$$\gamma = \frac{\Delta}{\lambda + \hat{\gamma}}, \tag{6}$$

$$\hat{m} = \frac{\alpha}{\gamma}\mathbb{E}_{y,h}\left[v(y, h, \gamma) - h\right], \tag{7}$$

$$\hat{q} = \frac{\alpha\Delta}{\gamma^2}\mathbb{E}_{y,h}\left[(v(y, h, \gamma) - h)^2\right], \tag{8}$$

$$\hat{\gamma} = \frac{\alpha\Delta}{\gamma}\left(1 - \mathbb{E}_{y,h}\left[\partial_h v(y, h, \gamma)\right]\right), \tag{9}$$

where $h \sim \mathcal{N}(m + yb, \Delta q)$, $\rho \in (0, 1)$ is the probability with which $y_i = 1$, and v is the solution of

$$v \equiv \arg\min_{\omega} \frac{(\omega - h(y, m, q, b))^2}{2\gamma} + \ell(\omega), \tag{10}$$

and the bias b, defined in (2), is the solution of the equation

$$\mathbb{E}_{y,h}\left[y(v - h)\right] = 0. \tag{11}$$

This is proven in the next section using Gordon's minimax approach. Once the fixed point values of the overlap m and length q are known, then we can express the asymptotic values for the generalization error and the training loss:

Theorem 2 (Generalization and training error) *In the same limit as in theorem 1, the generalization error expressed as fraction of wrong labeled instances is given by*

$$\varepsilon_{gen} = \rho Q\Big(\frac{m + b}{\sqrt{\Delta q}}\Big) + (1 - \rho)Q\Big(\frac{m - b}{\sqrt{\Delta q}}\Big), \tag{12}$$

where $Q(x) = \frac{1}{\sqrt{2\pi}}\int_x^\infty e^{-t^2/2}dt$ is the Gaussian tail function. The value of the training loss rescaled by the data dimension reads

$$\mathrm{L}_{\mathrm{train}} \equiv \lim_{d\to\infty}\frac{\mathcal{L}}{d} = \frac{\lambda q}{2} + \alpha\mathbb{E}_{y,h}\left[l(v(y, h, \gamma))\right]. \tag{13}$$

The details on (12) and (13) are provided in Appendices A and C.

B. MLE and Bayes-optimal estimator

The maximum likelihood estimation (MLE) for the considered model corresponds to the optimization of the non-regularized logistic loss. This follows directly from the Bayes formula:

$$\begin{aligned}\log \mathrm{p}(y|\mathrm{x}) &= \log \frac{\mathrm{p}(\mathrm{x}|y)\mathrm{p}_y(y)}{\sum\limits_{y=\pm 1} \mathrm{p}(\mathrm{x}|y)\mathrm{p}_y(y)} \\ &= -\log\left(1+\exp(-c)\right),\end{aligned} \tag{14}$$

where $c = \frac{2}{\Delta} y \left(\frac{1}{\sqrt{d}} \mathrm{v}^\top \mathrm{x} + \frac{\Delta}{2} \log \frac{\rho}{1-\rho}\right)$, therefore a simple redefinition of the variables leads to the logistic cost function that turns out to be the MLE (or rather the maximum a posteriori estimator if one allows the learning of a bias to account for the possibility of different cluster sizes).

The Bayes-optimal estimator is the "best" possible one in the sense that it minimizes the number of errors for new labels. It can be computed as

$$\hat{y}_{\mathrm{new}} = \arg\max_{y\in\pm 1} \log \mathrm{p}\left(y|\{\mathbf{X},\mathbf{y}\},\mathbf{x}_{\mathrm{new}}\right), \tag{15}$$

where $\{\mathbf{X},\mathbf{y}\}$ is the training set and $\mathbf{x}_{\mathrm{new}}$ is a previously unseen data point. In the Bayes-optimal setting, the model generating the data (1) and the prior distributions p_y, $\mathrm{p}_{\mathbf{z}}$, $\mathrm{p}_{\mathbf{v}^*}$ are known. Therefore, we can compute the posterior distribution in (15):

$$\mathrm{p}\left(y_{\mathrm{new}}|\mathbf{x}_{\mathrm{new}},\mathbf{X},\mathbf{y}\right) = \mathbb{E}_{\mathbf{v}|\mathbf{X},\mathbf{y}}\left[\mathrm{p}\left(y_{\mathrm{new}}|\mathbf{x}_{\mathrm{new}},\mathbf{v}\right)\right], \tag{16}$$

and applying Bayes theorem

$$\begin{aligned}&\mathrm{p}\left(y_{\mathrm{new}}|\mathbf{x}_{\mathrm{new}},\mathbf{v}\right) \propto \mathrm{p}\left(\mathbf{x}_{\mathrm{new}}|y_{\mathrm{new}},\mathbf{v}\right)\mathrm{p}_y\left(y_{\mathrm{new}}\right) \\ &\propto \exp\left(-\frac{1}{2\Delta}\sum_{i=1}^{d}\left(x^i_{\mathrm{new}} - \frac{y_{\mathrm{new}}\mathrm{v}^i}{\sqrt{d}}\right)^2\right)\mathrm{p}_y(y_{\mathrm{new}}).\end{aligned} \tag{17}$$

Hence, we can compute the Bayes-optimal generalization error using

$$\varepsilon_{\mathrm{gen}} = \mathbb{P}\left(\hat{y}_{\mathrm{new}} \neq y_{\mathrm{new}}\right). \tag{18}$$

This computation yields

$$\varepsilon^{\mathrm{BO}}_{\mathrm{gen}} = \rho Q\Big(\frac{m_{\mathrm{BO}}+b_{\mathrm{BO}}}{\sqrt{\Delta q_{\mathrm{BO}}}}\Big) + (1-\rho)Q\Big(\frac{m_{\mathrm{BO}}-b_{\mathrm{BO}}}{\sqrt{\Delta q_{\mathrm{BO}}}}\Big), \tag{19}$$

where $m_{BO} = q_{\mathrm{BO}} = \frac{\alpha}{\Delta+\alpha}$ and $b_{BO} = \frac{\Delta}{2}\log\frac{\rho}{1-\rho}$. This formula is derived in the Appendix B. The case $\rho = 1/2$ was also discussed in [4, 22].

Finally, it turns out that in this problem, one can reach the performances of the Bayes-optimal estimator, usually difficult to compute, efficiently using a simple plug-in estimator akin to applying the Hebb's rule [23]. Consider indeed the weight vector averaged over the training samples, each multiplied by its label and rescaled by $\sqrt{d}$

$$\hat{\mathbf{w}}_{\mathrm{Hebb}} = \frac{\sqrt{d}}{n}\sum_{\mu=1}^{n} y_\mu \mathbf{x}_\mu. \tag{20}$$

It is straightforward to check that, for $\hat{\mathbf{w}}_{\mathrm{Hebb}}$, one has in large dimension $m = 1$ and $q = (1 + \frac{\Delta}{\alpha})$. If one further optimizes the bias (for instance by cross validation) and uses its optimal value $b = \frac{\Delta q}{2m}\log\frac{\rho}{1-\rho}$, plugging these in eq. (12) one reaches Bayes-optimal performance $\varepsilon^{\mathrm{Hebb}}_{\mathrm{gen}} = \varepsilon^{\mathrm{BO}}_{\mathrm{gen}}$. Since there exists a plug-in estimator that reaches the Bayes-optimal performance, it is particularly interesting to see how the ones obtained by empirical risk minimization compare with the optimal results.

C. High-Dimensional Landscapes of Training Loss

Our analysis also leads to an analytical characterization of the high-dimensional landscapes of the training loss. First, we let

$$\begin{aligned}\mathcal{L}_\lambda(q,m,b) \overset{\mathrm{def}}{=} &\min_{\mathbf{w}} \frac{1}{d}\sum_{i=1}^{n}\ell[y_i(\tfrac{1}{\sqrt{d}}\mathbf{x}_i^\top\mathbf{w}+b)] + \frac{\lambda}{2d}\|\mathbf{w}\|^2 \\ &\text{subject to } \; q = \frac{1}{d}\|\mathbf{w}\|^2 \text{ and } m = \frac{1}{d}\mathbf{w}^\top\mathbf{v}^*\end{aligned} \tag{21}$$

to denote the normalized training loss when we restrict the weight vector to satisfy the two conditions in (21). In what follows, we refer to $\mathcal{L}_\lambda(q,m,b)$ as the "local training loss" at fixed values of q, m and b. The "global training loss" can then be obtained as

$$\mathcal{L}^*_\lambda \overset{\mathrm{def}}{=} \min_{m^2\leq q,b} \mathcal{L}_\lambda(q,m,b), \tag{22}$$

where the constraint that $m^2 \leq q$ is due to the Cauchy-Schwartz inequality: $|m| = \frac{|\mathbf{w}^\top\mathbf{v}^*|}{d} \leq \frac{\|\mathbf{w}\|}{\sqrt{d}}\frac{\|\mathbf{v}^*\|}{\sqrt{d}} = \sqrt{q}$.

In the high-dimensional limit when $n, d \to \infty$ with a fixed ratio $\alpha = n/d$, many properties of the local training loss can be characterized by a deterministic function, defined as

$$\mathcal{E}_\lambda(q,m,b) \overset{\mathrm{def}}{=} \alpha\mathbb{E}[\ell(v_{\gamma^*})] + \frac{\lambda q}{2}. \tag{23}$$

Here, for any $\gamma \geq 0$, v_γ denotes a random variable whose cumulative distribution function is given by

$$\begin{aligned}\mathbb{P}(v_r \leq v) = &\rho Q\left(\frac{\gamma\ell'(v)+v-m-b}{\sqrt{\Delta q}}\right) \\ &+ (1-\rho)Q\left(\frac{\gamma\ell'(v)+v-m+b}{\sqrt{\Delta q}}\right).\end{aligned} \tag{24}$$

Moreover, γ^* in (23) is the unique solution to the equation

$$\alpha\gamma^2\mathbb{E}[(\ell'(v_\gamma))^2] = \Delta(q-m^2). \tag{25}$$

Proposition 1 *Let Ω be an arbitrary compact subset of $\{(q,m,b) : m^2 \leq q\}$. We define*

$$\mathcal{L}_\lambda(\Omega) = \inf_{(q,m,b)\in\Omega} \mathcal{L}_\lambda(q,m,b)$$

and

$$\mathcal{E}_\lambda(\Omega) = \inf_{(q,m,b)\in\Omega} \mathcal{E}_\lambda(q,m,b).$$

For any constant $\delta > 0$ and as $n, d \to \infty$ with $\alpha = n/d$ fixed, it holds that

$$\mathbb{P}\Big(\mathcal{L}_\lambda(\Omega) \geq \mathcal{E}_\lambda(\Omega) - \delta\Big) \to 1. \tag{26}$$

Moreover,

$$\mathcal{L}_\lambda^* \to \mathcal{E}_\lambda^* \stackrel{def}{=} \inf_{m^2\leq q,b} \mathcal{E}_\lambda(q,m,b), \tag{27}$$

where $\mathcal{L}_\lambda^*$ is the global training loss defined in (22).

The characterization in (27) shows that the global training loss will concentrate around the fixed value $\mathcal{E}_\lambda^*$. Meanwhile, (26) implies that the deterministic function $\mathcal{E}_\lambda(q,m,b)$ serves as a high-probability *lower bound* of the local training loss $\mathcal{L}_\lambda(\Omega)$ over any given compact subset Ω. This latter property allows us to study the high-dimensional landscapes of the training loss as we move along the 3-dimensional space of the parameters q, m and b.

In particular, by studying $\mathcal{E}_\lambda(q,m,b)$, we can obtain the phase transition boundary characterizing the critical value of α below which the training data become perfectly separable.

Proposition 2 *Let $\lambda = 0$. Then*

$$\mathcal{E}_\lambda^* = \begin{cases} > 0, & \text{if } \alpha > \alpha^* \\ 0, & \text{if } \alpha < \alpha^*, \end{cases}$$

where

$$\alpha^* \stackrel{def}{=} \max_{0\leq r\leq 1, b} \eta(r,b) \tag{28}$$

$$\eta(r,b) = \frac{1-r^2}{\int_0^\infty u^2[\rho f(u+\frac{r}{\sqrt{\Delta}}-b)+(1-\rho)f(u+\frac{r}{\sqrt{\Delta}}+b)]du}$$

and $f(x)$ is the probability density function of $\mathcal{N}(0,1)$.

III. PROOF SKETCHES

In this section, we sketch the proof steps behind our main results presented in Section II. The full technical details are given in the Appendix C.

Roughly speaking, our proof strategy consists of three main ingredients: (1) Using Gordon's minimax inequalities [7–9], we can show that the random optimization problem associated with the local training loss in (21) can be compared against a much simpler optimization problem (see (32) in Section III A) that is essentially decoupled over its coordinates; (2) we show in Section III B that the aforementioned simpler problem concentrates around a well-defined deterministic limit as $n, d \to \infty$; and (3) by studying properties of the deterministic function, we reach the various characterizations given in Theorem 1, Proposition 1 and Proposition 2.

A. The dual formulation and Gordon's inequalities

The central object in our analysis is the local training loss $\mathcal{L}_\lambda(q,m,b)$ defined in (21). The challenge in directly analyzing (21) lies in the fact that it involves a d-dimensional (random) optimization problem where all the coordinates of the weight vector $\mathbf{w}$ are fully coupled. Fortunately, we can bypass this challenge via Gordon's inequalities, which allow us to characterize $\mathcal{L}_\lambda(q,m,b)$ by studying a much simpler problem. To that end, we first need to rewrite (21) as a minimax problem, via a Legendre transformation of the convex loss function $\ell(v)$:

$$\ell(v) = \max_u \Big\{ vu - \widetilde{\ell}(u) \Big\}, \tag{29}$$

where $\widetilde{\ell}(u)$ is the convex conjugate, defined as

$$\widetilde{\ell}(u) = \max_v \left\{ uv - \ell(v) \right\}.$$

For example, for the square, logistic, and hinge losses defined in Section I, their corresponding convex conjugates are given by

$$\widetilde{\ell}_{\text{square}}(u) = \frac{u^2}{4} + u \tag{30}$$

$$\widehat{\ell}_{\text{logistic}}(u) = \begin{cases} -H(-u), & \text{for } -1 \leq u \leq 0 \\ \infty, & \text{otherwise} \end{cases}, \tag{31}$$

where $H(u) \stackrel{\text{def}}{=} -u\log u - (1-u)\log(1-u)$ is the binary entropy function, and

$$\widehat{\ell}_{\text{hinge}}(u) = \begin{cases} u, & \text{for } -1 \leq u \leq 0 \\ \infty, & \text{otherwise,} \end{cases}$$

respectively.

Substituting (29) into (21) and recalling the data model (1), we can rewrite (21) as the following minimax problem

$$\mathcal{L}_\lambda(q,m,b) = \frac{\lambda q}{2} + \min_{\mathbf{w}\in\mathcal{S}_{q,m}} \max_{\mathbf{u}} \frac{1}{d}\sum_{i=1}^n u_i \left(\frac{\mathbf{w}^\top \mathbf{v}^*}{d} + \sqrt{\Delta}\frac{y_i \mathbf{z}_i^\top \mathbf{w}}{\sqrt{d}} + b y_i \right) - \widehat{\ell}(u_i),$$

where $\mathcal{S}_{q,m} \stackrel{\text{def}}{=} \big\{ \mathbf{w} : q = \frac{1}{d}\|\mathbf{w}\|^2 \text{ and } m = \frac{1}{d}\mathbf{w}^\top \mathbf{v}^* \big\}$.

Proposition 3 *For every (q,m,b) satisfying $q > m^2$, let*

$$\mathcal{E}_\lambda^{(d)}(q,m,b) \stackrel{def}{=} \frac{\lambda q}{2} + \max_{\boldsymbol{u}\in\mathbb{R}^n} \left\{ -\sqrt{\frac{\Delta_d \|\boldsymbol{u}\|^2 (q-m^2)}{d}} + \frac{\boldsymbol{u}^\top \boldsymbol{h}}{d} - \frac{1}{d}\sum_{i=1}^n \widetilde{\ell}(u_i) \right\}, \tag{32}$$

where $\Delta_d \stackrel{def}{=} (Q_d/d)\Delta$ with $Q_d \sim \chi_d^2$,

$$\boldsymbol{h} = \sqrt{\Delta q}\boldsymbol{s} + m\mathbf{1} + b[y_1, y_2, \ldots, y_n]^\top \tag{33}$$

and $\boldsymbol{s} \sim \mathcal{N}(0, \boldsymbol{I}_n)$ *is an i.i.d. Gaussian random vector. Then for any constant* c *and* $\delta > 0$*, we have*

$$\mathbb{P}(\mathcal{L}_\lambda(q,m,b) < c) \leq 2\mathbb{P}(\mathcal{E}_\lambda^{(d)}(q,m,b) < c) \tag{34}$$

and

$$\mathbb{P}(|\mathcal{L}_\lambda^* - c| > \delta) \leq 2\mathbb{P}(|\inf_{q,m,b} \mathcal{E}_\lambda^{(d)}(q,m,b) - c| > \delta). \tag{35}$$

The proof of Proposition 3, which can be found in the Appendix C 1, is based on an application of Gordon's comparison inequalities for Gaussian processes [7–9]. Similar techniques have been used by the authors of [6] to study the Gaussian mixture model for the non-regularized logistic loss for two clusters of the same size.

B. Asymptotic Characterizations

The definition of $\mathcal{E}_\lambda^{(d)}(q,m,b)$ in (32) still involves an optimization with an n-dimensional vector $\boldsymbol{u}$, but it can be simplified to a one-dimensional optimization problem with respect to a Lagrange multiplier γ:

Lemma 1

$$\mathcal{E}_\lambda^{(d)}(q,m,b) = \frac{\lambda q}{2} + \max_{\gamma>0} \Big\{ -\sqrt{\frac{\Delta_d(q-m^2)\|\boldsymbol{u}_\gamma\|^2}{d}} + \frac{\boldsymbol{u}_\gamma^\top h}{d} - \frac{1}{d}\sum_{i=1}^n \widetilde{\ell}(u_{\gamma,i}) \Big\}, \tag{36}$$

where $\boldsymbol{u}_\gamma \in \mathbb{R}^n$ *is the solution to*

$$\nabla \widetilde{\ell}(\boldsymbol{u}_\gamma) + \gamma \boldsymbol{u}_\gamma = \boldsymbol{h}, \tag{37}$$

with $\boldsymbol{h}$ *defined as in* (33).

One can show that the problem in (36) reaches its maximum at a point γ^* that is the unique solution to

$$\alpha\gamma^2 \frac{\|\boldsymbol{u}_\gamma\|^2}{n} = \Delta_d(q-m^2). \tag{38}$$

Moreover,

$$\mathcal{E}_\lambda^{(d)}(q,m,b) = \frac{\sum_{i=1}^n \left[u_{\gamma^*,i}\widetilde{\ell}'(u_{\gamma^*,i}) - \widetilde{\ell}(u_{\gamma^*,i})\right]}{d} + \frac{\lambda q}{2}. \tag{39}$$

In the asymptotic limit, as $n, d \to \infty$, both (38) and (39) converge towards their deterministic limits:

$$\alpha\gamma^2 \mathbb{E}[u_\gamma^2] = \Delta_d(q-m^2) \tag{40}$$

and

$$\mathcal{E}_\lambda^{(d)}(q,m,b) \to \alpha\mathbb{E}[u_\gamma \widetilde{\ell}'(u_\gamma) - \widetilde{\ell}(u_\gamma)] + \frac{\lambda q}{2}, \tag{41}$$

where u_γ is a random variable defined through the implicit equation $\widetilde{\ell}'(u_\gamma) + \gamma u_\gamma = h$.

Note that (40) and (41) already resemble their counterparts (25) and (23) given in our main results. The precise connection can be made by introducing the following scalar change of variables: $v = \widetilde{\ell}'(u)$. It is easy to verify from properties of Legendre transformations that

$$u = \ell'(v) \quad \text{and} \quad u\widetilde{\ell}'(u) - \widetilde{\ell}(u) = \ell(v).$$

Substituting these identities into (40) and (41) then gives us the characterizations (25) and (23) as stated in Section II.

Finally, the fixed point characterizations given in Theorem 1 can be obtained by taking derivatives of $\mathcal{E}_\lambda(q,m,b)$ with respect to q, m, b and setting them to 0. Similarly, the phase transition curve given in Proposition 2 can be obtained by quantifying the conditions under which the deterministic function $\mathcal{E}_\lambda(q,m,b)$ reaches its minimum at a finite point. We give more details in Appendix C 4 - C 5.

C. Interpretation from the replica method

These same equations can be independently derived from the non-rigorous replica methods from statistical physics [24], a technique that has proven useful in the study of high-dimensional statistical models, for instance following [21, 25]. Alternatively, these equations can also be seen as a special case of the State Evolution equation of the Approximate Message Passing algorithm [25–27]. Both interpretations can be useful, since the various quantities enjoy additional heuristic interpretations that allow us to obtain further insight. For instance, the parameter γ in (6) is connected to the rescaled variance of the estimator $\mathbf{w}$:

$$V = \lim_{d\to\infty} \frac{\mathbb{E}_{\mathbf{X},\mathbf{y}}\left[\|\mathbf{w}\|^2\right] - \mathbb{E}_{\mathbf{X},\mathbf{y}}\left[\|\mathbf{w}\|\right]^2}{d}. \tag{42}$$

The zero temperature limit of the fixed point equations obtained with the replica method corresponds to the loss minimization [24, 28]. In this limit, the behaviour of the rescaled variance V at zero penalty ($\lambda = 0$) is an indicator of data separability. In the non-separable regime, the minimizer of the loss is unique and $V \to 0$ at temperature $T = 0$. The parameter γ turns out to be simply $\gamma = \frac{V}{T}$. However, in the regime where data are separable there is a degeneracy of solutions at $\lambda = 0$, and the variance is finite: $V > 0$. Hence the parameter γ has a divergence at the transition, and this provides a very easy way to compute the location of the phase transition.

IV. CONSEQUENCES OF THE FORMULAS

In this section we evaluate the above formulas and investigate how does the test error depend on the regularization parameter λ, the fraction taken by the smaller cluster ρ, the ratio between the number of samples and

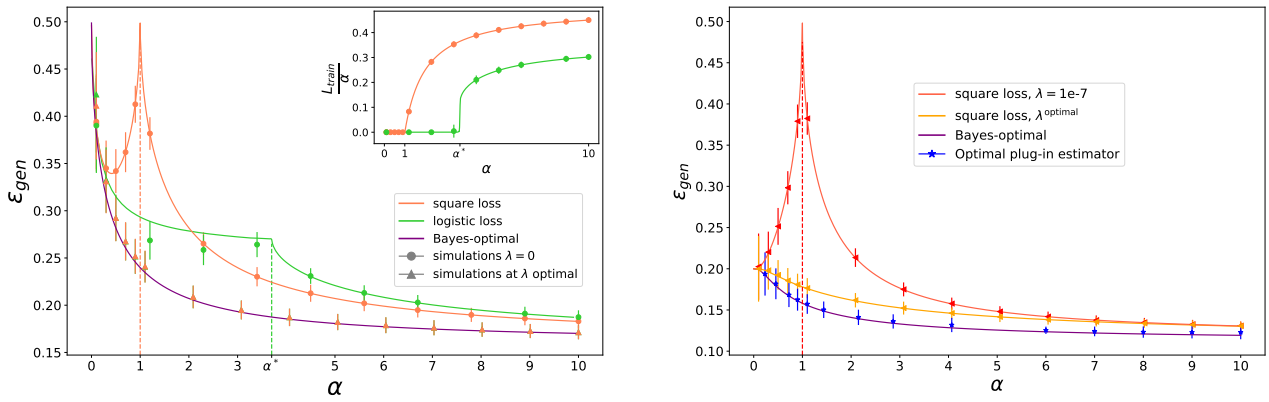

Figure 1. **Left (equal cluster size)**. Generalization error as a function of α at low regularization ($\lambda = 10^{-7}$) and fixed $\Delta = 1$, $\rho = 0.5$. The dashed vertical lines mark the interpolation thresholds. The generalization error achieved by the square and logistic losses is compared to the Bayes-optimal one. In this symmetric clusters case, it is possible to tune λ in order to reach the optimal performance. In the inset, the training loss as a function of α. The training loss is close to zero up to the interpolation transition. We compare our theoretical findings with simulations, at $d = 1000$. **Right (unequal cluster size)** Generalization error as a function of α at fixed $\Delta = 1$, $\rho = 0.2$. The performance of the square loss at low ($\lambda = 10^{-7}$) and optimal regularization is compared to the Bayes-optimal performance. In this non-symmetric case $\rho \neq 0.5$, the Bayes-optimal error is not achieved by the optimally regularized losses under consideration. We compare our results with numerical simulations at $d = 1000$. Additionally, we illustrate that the Bayes-optimal performance can be reached by the optimal plug-in estimator defined in eq. (20) (here with $d = 5000$).

the dimension α and the cluster variance Δ. The details on the evaluation and iteration of the fixed point equations in Theorem 1 are provided in Appendices D and F respectively. Keeping in mind that minimization of the non-regularized logistic loss corresponds in the considered model to the maximum likelihood estimation (MLE), we thus pay a particular attention to it as a benchmark of what the most commonly used method in statistics would achieve in this problem. Another important benchmark is the Bayes-optimal performance that provides a threshold that no algorithm can improve.

Weak and strong regularization — Fig. 1 summarizes how the regularization parameter λ and the cluster size ρ influence the generalization performances. The left panel of Fig. 1 is for the symmetric case $\rho{=}0.5$, the right panel for the non-symmetric case $\rho = 0.2$. Let us define as α^* the value of α such that for $\alpha < \alpha^*$ the training loss for hinge and logistic goes to zero (in other words, the data are linearly separable [16]. In the left part of Fig. 1 we depict (in green) the performance of the non-regularized logistic loss a.k.a. the maximum likelihood. For $\alpha > \alpha^*(\rho, \Delta)$ the training data are not linearly separable and the minimum training loss is bounded away from zero. For $\alpha < \alpha^*(\rho, \Delta)$ the data are linearly separable, in which case properly speaking the maximum likelihood is ill-defined [2], the curve that we depict is the limiting value reached as $\lambda \to 0^+$. The points are results of simulations with a standard scikitlearn [29] package. As shown in [11], even though the logistic estimator does not exist, gradient descent actually converges to the max-margin solution in this case, or equivalently to the least norm solution corresponding to $\lambda{\to}0^+$, a phenomenon coined "implicit regularization", which is well illustrated here.

Another interesting phenomenon is the non-monotonicity of the curve. This is actually an avatar of the so-called "double descent" phenomenon where the generalization "peaks" to a bad value and then decays again. This was observed and discussed recently in several papers [1, 12–15], but similar observations appeared as early as 1996 in Opper and Kinzel [30]. Indeed, we observed that the generalization error of the non-regularized square loss (in red) has a peak at $\alpha = 1$ at which point the data matrix in the non-regularized square loss problem becomes invertible. It is interesting that for $\alpha > \alpha^*$ the generalization performance of the non-regularized square loss is better than the one of the maximum likelihood. This has been proven recently in [5], who showed that among all the convex *non-regularized losses*, the square loss is optimal.

Fig. 1 further depicts (in purple) the Bayes-optimal error eq. (19). We have also evaluated the performance of both the logistic and square loss at optimal value of the regularization parameter λ. This is where the symmetric case (left panel) differs *crucially* from the non-symmetric one (right panel). While in the high-dimensional limit of the symmetric case the optimal regularization $\lambda_{\text{opt}} \to \infty$ and the corresponding error matches exactly the Bayes-optimal error, for the non-symmetric case $0 < \lambda_{\text{opt}} < \infty$ and the error for both losses is bounded away from the Bayes-optimal one for any $\alpha{>}0$.

We give a fully analytic argument in the Appendix E for the perhaps unexpected property of achieving the

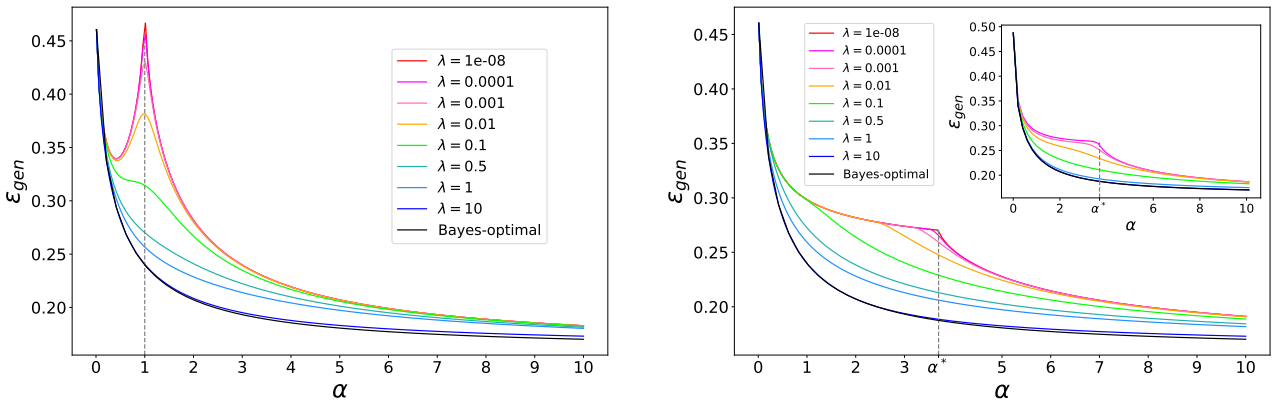

Figure 2. Generalization error as a function of α for different values of λ, at fixed $\Delta = 1$ and $\rho = 0.5$, for the square loss (left), the hinge loss (right) and the logistic loss (inset), compared to the Bayes-optimal error. If the two clusters have the same size, the Bayes-optimal error can be reached by increasing the regularization. Notice how regularization smooths the curves and makes the "peak" or "kink" disappear in all cases.

Bayes-optimal generalization at $\lambda_{\text{opt}} \to \infty$ and $\rho = 0.5$ for any loss that has a finite 2nd derivative at the origin. In simulations for finite value of d we use a large but finite value of λ, details on the simulation are provided in the Appendix F.

Regularization and the interpolation peak — In Fig. 2 we depict the dependence of the generalization error on the regularization λ for the symmetric $\rho = 0.5$ case for the square, hinge and logistic loss. The curves at small regularization show the interpolation peak/cusp at $\alpha = 1$ for the square loss and α^* for all the losses that are zero whenever the data are linearly separable. We observe a smooth disappearance of the peak/cusp as regularization is added, similarly to what has been observed in other models that present the interpolation peak [1, 15] in the case of the square loss. Here we thus show that a similar phenomena arises with the logistic and hinge losses as well; this is of interest as this effect has been observed in deep neural networks using a logistic/cross-entropy loss [12, 31]. In fact, as the regularization increases, the error gets better in this model with equal-size cluster, and one reaches the Bayes-optimal values for large regularization.

Max-margin and weak regularization — Fig. 3 illustrates the generic property that all non-regularized monotone non-increasing loss functions converge to the max-margin solution for linearly separable data [10]. Fig. 3 depicts a very slow convergence towards this result as a function of regularization parameter λ for the logistic loss. While for $\alpha > \alpha^*$ both the hinge and logistic losses performance is basically indistinguishable from the asymptotic one already at $\log \lambda \approx -3$, for $\alpha < \alpha^*$ the convergence of the logistic loss still did not happen even at $\log \lambda \approx -10$.

Cluster sizes and regularization — In Fig. 4 we study in greater detail the dependence of the generalization error both on the regularization λ and ρ as $\rho \to 0.5$.

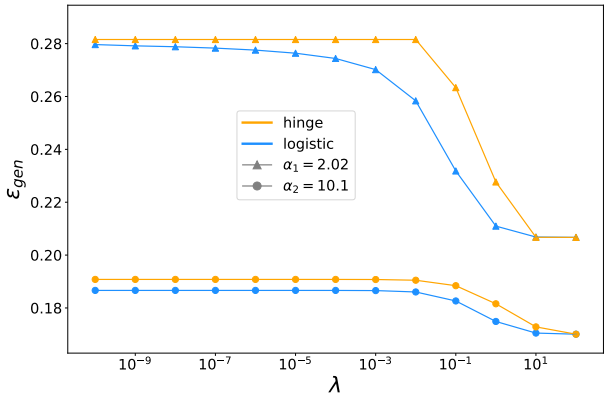

Figure 3. Generalization error as a function of λ for the hinge and logistic losses, at fixed $\Delta = 1$, $\rho = 0.5$ and two different values of α: $\alpha_1 = 2, \alpha_2 = 10$. As $\lambda \to 0^+$, the error of the two losses approaches the same value if the data are separable ($\alpha_1 < \alpha^*$). This is not true if the data are not separable ($\alpha_2 > \alpha^*$). At large λ, the error of both losses reaches the Bayes-optimal, for all α.

We see that the optimality of $\lambda \to \infty$ holds only strictly at $\rho = 0.5$ and at any ρ only close to 0.5 the error at $\lambda \to \infty$ is very large and there is a well delimited region of λ for which the error is close to (but strictly above) the Bayes-optimal error. As $\rho \to 0.5$ this interval is getting longer and longer until it diverges at $\rho = 0.5$. It needs to be stressed that this result is asymptotic, holding only when $n, d \to \infty$ while $n/d = \alpha$ is fixed. The finite size fluctuations cause that finite size system behaves rather as if ρ was close but not equal to 0.5, and at finite size if we set λ arbitrarily large then we reach a high generalization error. We instead need to optimize the value of λ for finite sizes either by cross-validation or otherwise.

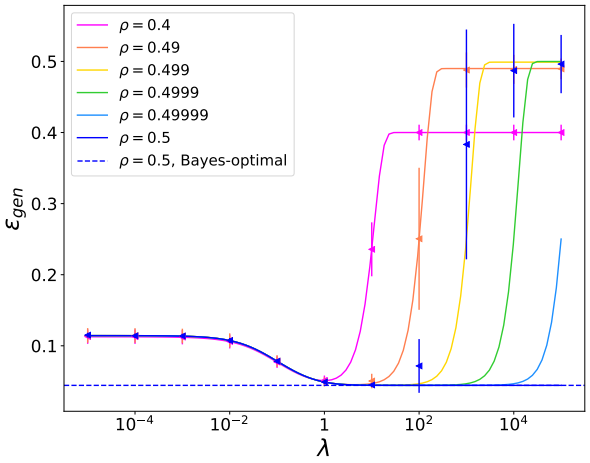

Figure 4. Generalization error as a function of λ for different values of ρ close to 0.5, at fixed $\Delta = 0.3$ and $\alpha = 2$, for the square loss. At all $\rho < 0.5$, the error exhibits a minimum at finite $\lambda = \lambda^*$, and reaches a plateau at $\lambda > \lambda^*$. The value of the error at the plateau is $\varepsilon_{\rm gen} = \min\{\rho, 1-\rho\}$, which is the error attained by the greedy strategy of assigning all points to the larger cluster. We compare our analytical results with simulations for $\rho = 0.4, 0.49, 0.5$. Simulations for $\rho = 0.4$ are done at $d = 1000$. Simulations for $\rho = 0.5, 0.49$ are done at $d = 10000$. Since the dimensionality d is finite in the simulations, effectively $\rho < 0.5$ in the numerics. Therefore, simulations always reach a plateau at large λ.

Separability phase transition — The position of the "interpolation" threshold when data become linearly separable has a well defined limit in the high-dimensional regime as a function of the ratio between the number of samples n and the dimension d. The kink in generalization indeed occurs at a value α^* when the training loss of logistic and hinge losses goes to zero (while for the square loss the peak appears at $d = n$ when the system of n linear equations with d parameters becomes solvable). The position of α^*, given by Proposition 2 is shown in Fig. 5 as a function of the cluster variance for different values of ρ. For very large cluster variance, the data become random and hence $\alpha = 2$ for equal-sized cluster, as famously derived in classical work by [32]. When $\rho < 1/2$, however, it is easier to separate linearly the data points and the limiting value of α^* gets larger and differ from Cover's. For finite Δ, the two Gaussian distributions become distinguishable, and the data acquires structure. Consequently, the α^* is growing as the correlations make data easier to linearly separate again, similarly as described [16]. This phenomenology of the separability phase transition, or equivalently of the existence of the maximum likelihood estimator, thus seems very generic.

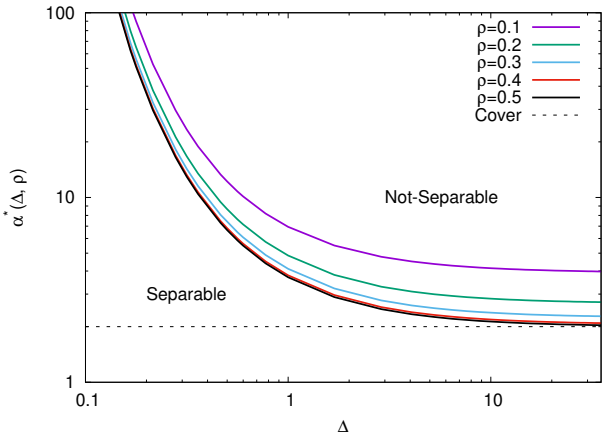

Figure 5. Critical value $\alpha = \alpha^*$, define by Proposition 2, at which the linear separability transition occurs as a function of Δ, for different values of ρ. Similarly as for what happens for Gaussian data [16], the MLE does not exists on the left the curve. The line indicates the location of the transition from linearly separable to non-linearly separable data, that depends on the data structure (the variance Δ and the fraction ρ).

ACKNOWLEDGEMENTS

We thank Pierfrancesco Urbani, Federica Gerace, and Bruno Loureiro for many clarifying discussions related to this project. This work is supported by the ERC under the European Unions Horizon 2020 Research and Innovation Program 714608-SMiLe, by the French Agence Nationale de la Recherche under grant ANR-17-CE23-0023-01 PAIL and ANR-19-P3IA-0001 PRAIRIE, and by the US National Science Foundation under grants CCF-1718698 and CCF-1910410. We also acknowledge support from the chaire CFM-ENS "Science des données". Part of this work was done when Yue Lu was visiting Ecole Normale as a CFM-ENS "Laplace" invited researcher. We thank Google Cloud for providing us access to their platform through the Research Credits Application program.

[1] T. Hastie, A. Montanari, S. Rosset, and R. J. Tibshirani, Surprises in high-dimensional ridgeless least squares interpolation, arXiv preprint arXiv:1903.08560 (2019).
[2] P. Sur and E. J. Candès, A modern maximum-likelihood theory for high-dimensional logistic regression, Proceedings of the National Academy of Sciences **116**, 14516 (2019).
[3] J. Friedman, T. Hastie, and R. Tibshirani, *The elements of statistical learning*, Vol. 1 (Springer series in statistics New York, 2001).
[4] M. Lelarge and L. Miolane, Asymptotic Bayes risk for Gaussian mixture in a semi-supervised setting, arXiv preprint arXiv:1907.03792 (2019).

[5] X. Mai and Z. Liao, High dimensional classification via empirical risk minimization: Improvements and optimality, arXiv preprint arXiv:1905.13742 (2019).

[6] Z. Deng, A. Kammoun, and C. Thrampoulidis, A model of double descent for high-dimensional binary linear classification, arXiv preprint arXiv:1911.05822 (2019).

[7] Y. Gordon, Some inequalities for Gaussian processes and applications, Israel Journal of Mathematics **50**, 265 (1985).

[8] Y. Gordon, On milman's inequality and random subspaces which escape through a mesh in rn, in *Geometric Aspects of Functional Analysis*, Lecture Notes in Mathematics No. 1317, edited by J. Lindenstrauss and V. D. Milman (Springer Berlin Heidelberg, 1988) pp. 84–106.

[9] C. Thrampoulidis, S. Oymak, and B. Hassibi, Regularized linear regression: A precise analysis of the estimation error, in *Proceedings of The 28th Conference on Learning Theory*, Vol. 40 (PMLR, Paris, France, 2015) pp. 1683–1709.

[10] S. Rosset, J. Zhu, and T. J. Hastie, Margin maximizing loss functions, in *Advances in neural information processing systems* (2004) pp. 1237–1244.

[11] D. Soudry, E. Hoffer, M. S. Nacson, S. Gunasekar, and N. Srebro, The implicit bias of gradient descent on separable data, The Journal of Machine Learning Research **19**, 2822 (2018).

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

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

[14] P. P. Mitra, Understanding overfitting peaks in generalization error: Analytical risk curves for l_2 and l_1 penalized interpolation, arXiv preprint arXiv:1906.03667 (2019).

[15] S. Mei and A. Montanari, The generalization error of random features regression: Precise asymptotics and double descent curve, arXiv preprint arXiv:1908.05355 (2019).

[16] E. J. Candès and P. Sur, The phase transition for the existence of the maximum likelihood estimate in high-dimensional logistic regression, arXiv preprint arXiv:1804.09753 (2018).

[17] E. Gardner, The space of interactions in neural network models, Journal of physics A: Mathematical and general **21**, 257 (1988).

[18] E. Gardner and B. Derrida, Three unfinished works on the optimal storage capacity of networks, Journal of Physics A: Mathematical and General **22**, 1983 (1989).

[19] W. Krauth and M. Mézard, Storage capacity of memory networks with binary couplings, Journal de Physique **50**, 3057 (1989).

[20] P. Del Giudice, S. Franz, and M. Virasoro, Perceptron beyond the limit of capacity, Journal de Physique **50**, 121 (1989).

[21] S. Franz, D. J. Amit, and M. A. Virasoro, Prosopagnosia in high capacity neural networks storing uncorrelated classes, Journal de Physique **51**, 387 (1990).

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

[23] D. O. Hebb, *The organization of behavior: A neuropsychological theory* (Psychology Press, 2005).

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

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

[26] D. L. Donoho, A. Maleki, and A. Montanari, Message-passing algorithms for compressed sensing, Proceedings of the National Academy of Sciences **106**, 18914 (2009).

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

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

[29] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, *et al.*, Scikit-learn: Machine learning in python, Journal of machine learning research **12**, 2825 (2011).

[30] M. Opper and W. Kinzel, Statistical mechanics of generalization, in *Models of neural networks III* (Springer, 1996) pp. 151–209.

[31] P. Nakkiran, G. Kaplun, Y. Bansal, T. Yang, B. Barak, and I. Sutskever, Deep double descent: Where bigger models and more data hurt, arXiv preprint arXiv:1912.02292 (2019).

[32] T. M. Cover, Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition, IEEE transactions on electronic computers , 326 (1965).

Appendix A: Derivation of the generalization error formula

The generalization error is defined as the average fraction of mislabeled instances

$$\varepsilon_{\rm gen} = \frac{1}{4}\mathbb{E}_{y_{\rm new},\mathbf{x}_{\rm new},\mathbf{X},\mathbf{y}}\left[(y_{\rm new}-\hat{y}_{\rm new})^2\right], \tag{A.1}$$

where $y_{\rm new}$ is the label of a new observation $\mathbf{x}_{\rm new}$, and the estimator $\hat{y}_{\rm new}$ is computed as

$$\hat{y}_{\rm new} = {\rm sign}\left(\frac{\mathbf{w}^\top \mathbf{x}_{\rm new}}{\sqrt{d}} + b\right). \tag{A.2}$$

Eq. (A.2) holds for every vector $\mathbf{w} = \mathbf{w}(\mathbf{X},\mathbf{y})$ and bias $b = b(\mathbf{X},\mathbf{y})$ computed on the training set $\{\mathbf{X},\mathbf{y}\}$. Using the fact that $y_{\rm new}, \hat{y}_{\rm new} = \pm 1$, it is easy to show that (A.1) can be rewritten as

$$\varepsilon_{\rm gen} = \frac{1}{2}\left(1 - \mathbb{E}_{y_{\rm new},\mathbf{x}_{\rm new},\mathbf{X},\mathbf{y}}\left[y_{\rm new}\hat{y}_{\rm new}\right]\right) = \frac{1}{2}\left(1 - \mathbb{E}_{y_{\rm new},\mathbf{x}_{\rm new},\mathbf{X},\mathbf{y}}\left[y_{\rm new}{\rm sign}\left(\frac{\mathbf{w}^\top \mathbf{x}_{\rm new}}{\sqrt{d}} + b\right)\right]\right). \tag{A.3}$$

Let us consider the last term in (A.3). Using again $y_{\rm new} = \pm 1$, we can move $y_{\rm new}$ inside the argument of the sign function and rewrite

$$\mathbb{E}_{y_{\rm new},\mathbf{x}_{\rm new},\mathbf{X},\mathbf{y}}\left[y_{\rm new}{\rm sign}\left(\frac{\mathbf{w}^\top \mathbf{x}_{\rm new}}{\sqrt{d}} + b\right)\right] = \mathbb{E}_{y_{\rm new},\mathbf{x}_{\rm new},\mathbf{X},\mathbf{y}}\left[{\rm sign}\left(\frac{y_{\rm new}\mathbf{w}^\top \mathbf{x}_{\rm new}}{\sqrt{d}} + y_{\rm new}b\right)\right]. \tag{A.4}$$

The term $y_{\rm new}\mathbf{x}_{\rm new}$ can be rewritten as

$$y_{\rm new}\mathbf{x}_{\rm new} = y_{\rm new}\left(y_{\rm new}\frac{\mathbf{v}^*}{\sqrt{d}} + \sqrt{\Delta}\mathbf{z}_{\rm new}\right) = \frac{\mathbf{v}^*}{\sqrt{d}} + \sqrt{\Delta}\mathbf{z}'_{\rm new}, \tag{A.5}$$

where $\mathbf{z}'_{\rm new} = y_{\rm new}\mathbf{z}_{\rm new} \sim \mathcal{N}(\mathbf{0},\mathbf{I}_d)$ has the same distribution as $\mathbf{z}_{\rm new}$, since $y_{\rm new}$ and $\mathbf{z}_{\rm new}$ are independent. Hence

$$\mathbb{E}_{y_{\rm new},\mathbf{x}_{\rm new},\mathbf{X},\mathbf{y}}\left[{\rm sign}\left(\frac{\mathbf{w}^\top y_{\rm new}\mathbf{x}_{\rm new}}{\sqrt{d}} + y_{\rm new}b\right)\right] = \mathbb{E}_{y_{\rm new},\mathbf{z}'_{\rm new},\mathbf{v}^*,\mathbf{X},\mathbf{y}}\left[{\rm sign}\left(\frac{\mathbf{w}^\top \mathbf{v}^*}{d} + \sqrt{\frac{\Delta}{d}}\mathbf{w}^\top \mathbf{z}'_{\rm new} + y_{\rm new}b\right)\right]. \tag{A.6}$$

The estimator $\mathbf{w}$ only depends on the training set, hence $\mathbf{w}$ and $\mathbf{z}'_{\rm new}$ are independent. We call their rescaled scalar product ς, a random variable distributed as a standard normal

$$\varsigma = \frac{1}{\|\mathbf{w}\|}\mathbf{w}^\top \mathbf{z}'_{\rm new} \sim \mathcal{N}(0,1). \tag{A.7}$$

By averaging over ς, we obtain

$$\begin{aligned} &\mathbb{E}_{y_{\rm new},\mathbf{v}^*,\mathbf{X},\mathbf{y},\varsigma}\left[{\rm sign}\left(\frac{\mathbf{w}^\top \mathbf{v}^*}{d} + \sqrt{\frac{\Delta}{d}}\|\mathbf{w}\|\varsigma + y_{\rm new}b\right)\right] \\ &= \mathbb{E}_{y_{\rm new},\mathbf{v}^*,\mathbf{X},\mathbf{y},\varsigma}\left[{\rm sign}\left(\frac{1}{\sqrt{\Delta}}\frac{\mathbf{w}}{||\mathbf{w}||}^\top \frac{\mathbf{v}^*}{\sqrt{d}} + \varsigma + y_{\rm new}b\frac{\sqrt{d}}{\sqrt{\Delta}||\mathbf{w}||}\right)\right], \end{aligned} \tag{A.8}$$

where we have used that $\sqrt{\frac{\Delta}{d}}\|\mathbf{w}\| > 0$ to rescale the argument of the sign function. Finally, we obtain

$$\varepsilon_{\rm gen} = \frac{1}{2}\left(1 - \mathbb{E}_{y_{\rm new},\mathbf{v}^*,\mathbf{X},\mathbf{y}}\left[\mathbb{P}\left(\varsigma > -\tau\right) - \mathbb{P}\left(\varsigma < -\tau\right)\right]\right) = \mathbb{E}_{y_{\rm new},\mathbf{v}^*,\mathbf{X},\mathbf{y}}\left[Q(\tau)\right]. \tag{A.9}$$

where $Q(x) = \frac{1}{\sqrt{2\pi}}\int_x^\infty e^{-t^2/2}dt$ is the Gaussian tail function, and we have defined

$$\tau = \frac{\sqrt{d}}{\sqrt{\Delta}||\mathbf{w}||}\left(\frac{\mathbf{w}^\top \mathbf{v}^*}{d} + y_{\rm new}b\right). \tag{A.10}$$

In the large d limit, the overlaps concentrate to deterministic quantities:

$$\frac{\mathbf{w}^\top \mathbf{v}^*}{d} \underset{d\to\infty}{\longrightarrow} m, \tag{A.11}$$

$$\frac{||\mathbf{w}||}{\sqrt{d}} \underset{d\to\infty}{\longrightarrow} \sqrt{q}. \tag{A.12}$$

Hence the generalization error reads

$$\varepsilon_{\rm gen} = \rho Q\Big(\frac{m+b}{\sqrt{\Delta q}}\Big) + (1-\rho) Q\Big(\frac{m-b}{\sqrt{\Delta q}}\Big), \tag{A.13}$$

where $\rho \in (0,1)$ is the probability that $y_{\rm new} = +1$.

Appendix B: Derivation of the Bayes-optimal error

In order to compute the Bayes-optimal error, we consider the posterior distribution of a new label $y_{\rm new}$, given the corresponding new data point $\mathbf{x}_{\rm new}$ and the estimate $\mathbf{v}$ of the true centroid $\mathbf{v}^*$

$$\mathrm{p}\left(y_{\rm new}|\mathbf{x}_{\rm new},\mathbf{v}\right) \propto \mathrm{p}\left(\mathbf{x}_{\rm new}|y_{\rm new},\mathbf{v}\right) \mathrm{p}_y\left(y_{\rm new}\right) \propto \exp\left(-\frac{1}{2\Delta}\sum_{i=1}^{d}\left(x^i_{\rm new} - \frac{y_{\rm new}\mathrm{v}^i}{\sqrt{d}}\right)^2\right) \mathrm{p}_y(y_{\rm new}), \tag{B.1}$$

where "$\propto$" takes into account the normalization over $y_{\rm new}$. Similarly, the posterior on $\mathbf{v}$ given the training data is

$$\mathrm{p}\left(\mathbf{v}|\mathbf{X},\mathbf{y}\right) \propto \mathrm{p}\left(\mathbf{X}|\mathbf{v},\mathbf{y}\right) \mathrm{p}_{\mathbf{v}}\left(\mathbf{v}\right) \propto \left[\prod_{\mu=1}^{n} \exp\left(-\frac{1}{2\Delta}\sum_{i=1}^{d}\left(x^i_\mu - \frac{y_\mu \mathrm{v}^i}{\sqrt{d}}\right)^2\right)\right] \exp\left(-\frac{1}{2}\sum_{i=1}^{d}(\mathrm{v}^i)^2\right), \tag{B.2}$$

where we remind that $\mathbf{v}$ has i.i.d. components taken in $\mathcal{N}(0,1)$, and "$\propto$" takes into account the normalization over $\mathbf{v}$. We would like to find an explicit expression for

$$\mathrm{p}\left(y_{\rm new}|\mathbf{x}_{\rm new},\mathbf{X},\mathbf{y}\right) = \mathbb{E}_{\mathbf{v}|\mathbf{X},\mathbf{y}}\left[\mathrm{p}\left(y_{\rm new}|\mathbf{x}_{\rm new},\mathbf{v}\right)\right], \tag{B.3}$$

in order to estimate the new label as

$$\hat{y}_{\rm new} = \arg\max_{y'=\pm 1} \log \mathrm{p}\left(y'|\mathbf{x}_{\rm new},\mathbf{X},\mathbf{y}\right). \tag{B.4}$$

Therefore, we have to compute

$$\mathbb{E}_{\mathbf{v}|\mathbf{X},\mathbf{y}}\left[\mathrm{p}\left(y_{\rm new}|\mathbf{x}_{\rm new},\mathbf{v}\right)\right] \propto \mathrm{p}_y\left(y_{\rm new}\right) \int \left(\prod_{i=1}^{d} \mathrm{dv}^i \ \mathrm{e}^{-\frac{1}{2}(\mathrm{v}^i)^2}\right) \prod_{\mu=0}^{n} \mathrm{e}^{-\frac{1}{2\Delta}\sum_{i=1}^{d}\left(x^i_\mu - \frac{y_\mu \mathrm{v}^i}{\sqrt{d}}\right)^2}, \tag{B.5}$$

where in the product over μ on the right-hand side we have used the notation $y_0 = y_{\rm new}$, $\mathbf{x}_0 = \mathbf{x}_{\rm new}$. Let us call $I_{\rm v}$ the integral over $\mathbf{v}$ in (B.5).

$$I_{\rm v} = \int \prod_{i=1}^{d} \mathrm{dv}^i \ \mathrm{e}^{-\sum_{i=1}^{d}\left[\frac{1}{2\Delta}\sum_{\mu=0}^{n}\left(x^i_\mu - \frac{y_\mu \mathrm{v}^i}{\sqrt{d}}\right)^2 + \frac{1}{2}(\mathrm{v}^i)^2\right]} = \prod_{i=1}^{d} \int \mathrm{dv} \ \mathrm{e}^{-\frac{1}{2\Delta}\sum_{\mu=0}^{n}\left(x^i_\mu - \frac{y_\mu \mathrm{v}}{\sqrt{d}}\right)^2 - \frac{1}{2}\mathrm{v}^2}, \tag{B.6}$$

where in the last equality we have dropped the index i from the components of $\mathbf{v}$ for simplicity, since they are all independent. Computing the integral over v, we obtain

$$\begin{aligned} I_{\mathrm{v}} &= C\left(\alpha,\Delta,d\right)\prod_{i=1}^{d}\prod_{\mu=0}^{n}\exp\left(-\frac{1}{2\Delta\left(\alpha+\Delta+\frac{1}{d}\right)}\left((\alpha+\Delta)(x_\mu^i)^2-\frac{\alpha}{n}y_\mu x_\mu^i\sum_{\substack{\nu=0\\\nu\neq\mu}}^{n}y_\nu x_\nu^i\right)\right) \\ &= C\left(\alpha,\Delta,d\right)\exp\left(-\frac{1}{2\Delta\left(\alpha+\Delta+\frac{1}{d}\right)}\sum_{i=1}^{d}\left((\alpha+\Delta)(x_{\mathrm{new}}^i)^2-\frac{\alpha}{n}y_{\mathrm{new}}x_{\mathrm{new}}^i\sum_{\nu=1}^{n}y_\nu x_\nu^i\right)\right) \\ &\times\exp\left(-\frac{1}{2\Delta\left(\alpha+\Delta+\frac{1}{d}\right)}\sum_{\mu=1}^{n}\sum_{i=1}^{d}\left((\alpha+\Delta)(x_\mu^i)^2-\frac{\alpha}{n}y_\mu x_\mu^i\sum_{\substack{\nu=1\\\nu\neq\mu}}^{n}y_\nu x_\nu^i-\frac{\alpha}{n}y_\mu x_\mu^i y_{\mathrm{new}}x_{\mathrm{new}}^i\right)\right) \\ &= C\left(\alpha,\Delta,d\right)\tilde{C}\left(\mathbf{X},\mathbf{y},\mathbf{x}_{\mathrm{new}},\alpha,\Delta,d\right)\exp\left(\frac{\alpha}{\Delta\left(\alpha+\Delta+\frac{1}{d}\right)}y_{\mathrm{new}}\mathbf{x}_{\mathrm{new}}^\top\frac{1}{n}\sum_{\mu=1}^{n}y_\mu\mathbf{x}_\mu\right), \end{aligned} \tag{B.7}$$

where the first two factors C and $\tilde{C}$ contain all the terms that do not depend on y_{new}. Therefore

$$\hat{y}_{\mathrm{new}}=\arg\max_{y=\pm1}\left[\frac{\alpha}{\Delta\left(\alpha+\Delta+\frac{1}{d}\right)}y\mathbf{x}_{\mathrm{new}}^\top\frac{1}{n}\sum_{\mu=1}^{n}y_\mu\mathbf{x}_\mu+\log\mathrm{p}_y\left(y\right)\right]. \tag{B.8}$$

Using the fact that $y_\mu\mathbf{x}_\mu=\frac{\mathbf{v}^*}{\sqrt{d}}+\sqrt{\Delta}\mathbf{z}_\mu$, $\mathbf{z}_\mu\sim\mathcal{N}(0,\mathbf{I}_d)$ and $\mathbf{v}^*$ is the true realization of $\mathbf{v}$, the first term in (B.8) in the limit where $n,d\to\infty$ can be rewritten as

$$\frac{1}{n}\sum_{\mu=1}^{n}\mathbf{x}_{\mathrm{new}}^\top y_\mu\mathbf{x}_\mu\underset{n,d\to\infty}{\longrightarrow}y_{\mathrm{new}}+\sqrt{\Delta\left(1+\frac{\Delta}{\alpha}\right)}z'_{\mathrm{new}}, \tag{B.9}$$

where $z'_{\mathrm{new}}\sim\mathcal{N}(0,1)$. Therefore, in the large d limit we find that

$$\hat{y}_{\mathrm{new}}=\arg\max_{y=\pm1}\left[\frac{\alpha}{\Delta\left(\alpha+\Delta\right)}y\left(y_{\mathrm{new}}+\sqrt{\Delta\left(1+\frac{\Delta}{\alpha}\right)}z'_{\mathrm{new}}\right)+\log\mathrm{p}_y\left(y\right)\right]. \tag{B.10}$$

It is useful to rewrite the generalization error as

$$\varepsilon_{\mathrm{gen}}=\frac{1}{4}\mathbb{E}_{\boldsymbol{X},\mathbf{y},\mathbf{x}_{\mathrm{new}},y_{\mathrm{new}}}\left[(\hat{y}_{\mathrm{new}}-y_{\mathrm{new}})^2\right]=\sum_{y_{\mathrm{new}}=-1,1}\mathbb{P}\left(\hat{y}_{\mathrm{new}}\neq y_{\mathrm{new}}\right)\mathrm{p}_y(y_{\mathrm{new}}). \tag{B.11}$$

Using (B.10), we can compute

$$\mathbb{P}\left(\hat{y}_{\mathrm{new}}\neq y_{\mathrm{new}}\right)=\mathbb{P}\left(y_{\mathrm{new}}z'_{\mathrm{new}}<-\sqrt{\frac{\alpha}{\Delta(\alpha+\Delta)}}\left(1+\left(1+\frac{\Delta}{\alpha}\right)\frac{\Delta}{2}\log\frac{\mathrm{p}_y(y_{\mathrm{new}})}{\mathrm{p}_y(-y_{\mathrm{new}})}\right)\right). \tag{B.12}$$

If $y_{\mathrm{new}}=1$, (B.12) gives

$$\mathbb{P}\left(\hat{y}_{\mathrm{new}}\neq1\right)=Q\left(\frac{\frac{\alpha}{\Delta+\alpha}+\frac{\Delta}{2}\log\frac{\rho}{1-\rho}}{\sqrt{\Delta\frac{\alpha}{\Delta+\alpha}}}\right), \tag{B.13}$$

where $Q(x)=\frac{1}{\sqrt{2\pi}}\int_x^\infty e^{-t^2/2}dt$ is the Gaussian tail function. If $y_{\mathrm{new}}=-1$, (B.12) gives

$$P\left(\hat{y}_{\mathrm{new}}\neq-1\right)=Q\left(\frac{\frac{\alpha}{\Delta+\alpha}-\frac{\Delta}{2}\log\frac{\rho}{1-\rho}}{\sqrt{\Delta\frac{\alpha}{\Delta+\alpha}}}\right). \tag{B.14}$$

Using the fact that $\rho=\mathrm{p}_y(1)$ and $1-\rho=\mathrm{p}_y(-1)$, we get that

$$\varepsilon_{\mathrm{gen}}^{\mathrm{BO}}=\rho Q\left(\frac{\frac{\alpha}{\Delta+\alpha}+\frac{\Delta}{2}\log\frac{\rho}{1-\rho}}{\sqrt{\Delta\frac{\alpha}{\Delta+\alpha}}}\right)+(1-\rho)Q\left(\frac{\frac{\alpha}{\Delta+\alpha}-\frac{\Delta}{2}\log\frac{\rho}{1-\rho}}{\sqrt{\Delta\frac{\alpha}{\Delta+\alpha}}}\right). \tag{B.15}$$

1. Bayes-optimal estimator

It is worth noting that the optimal error in (B.15) can be achieved by the plug-in estimator

$$\hat{\mathbf{w}} = \frac{\sqrt{d}}{n}\sum_{\mu=1}^{n} y_\mu \mathbf{x}_\mu. \tag{B.16}$$

This result was already shown in [4] for the case of symmetric clusters. The optimal bias is obtained from the minimization of the generalization error (A.13) with respect to b, at fixed m, q. This yields:

$$\hat{b} = \underset{b}{\text{argmin}}\ \varepsilon_{\text{gen}}(q,m) = \frac{q}{m}\frac{\Delta}{2}\log\left(\frac{\rho}{1-\rho}\right). \tag{B.17}$$

Substituting (B.16) in the definition of the overlaps (3) in the main text, we obtain that the values of m and q associated to the plugin estimator are

$$m = 1, \qquad q = \left(1 + \tfrac{\Delta}{\alpha}\right). \tag{B.18}$$

Hence, the generalization error of the plug-in estimator is

$$\begin{aligned} \varepsilon_{\text{gen}}^{\text{plugin}} &= \mathbb{P}\left(y_{\text{new}}\left(\frac{1}{\sqrt{d}}\hat{\mathbf{w}}^\top \mathbf{x}_{\text{new}} + \hat{b}\right) < 0\right) \\ &= \mathbb{P}\left(y_{\text{new}} z'_{\text{new}} < -\sqrt{\frac{\alpha}{\Delta(\alpha+\Delta)}}\left(1 + y_{\text{new}}\left(1+\frac{\Delta}{\alpha}\right)\frac{\Delta}{2}\log\frac{\rho}{1-\rho}\right)\right), \end{aligned} \tag{B.19}$$

where we have used (B.9) in the last equality. The probability in (B.19) is the same as in (B.12). Hence, the plug-in estimator achieves the Bayes-optimal error.

Appendix C: Details of proofs

In what follows, we provide more technical details for several key results stated in Section III. They serve as the basis of the proof of Proposition 1.

1. Proof of Proposition 3

Recall from the main text that

$$\begin{aligned} \mathcal{L}_\lambda(q,m,b) &= \frac{\lambda q}{2} + \min_{\mathbf{w}\in\mathcal{S}_{q,m}}\max_{\mathbf{u}} \frac{1}{d}\sum_{i=1}^{n}\Big[u_i\Big(\frac{\mathbf{w}^\top\mathbf{v}^*}{d} + \sqrt{\Delta}\frac{y_i\mathbf{z}_i^\top\mathbf{w}}{\sqrt{d}} + b y_i\Big) - \widehat{\ell}(u_i)\Big] \\ &= \frac{\lambda q}{2} + \min_{\mathbf{w}\in\mathcal{S}_{q,m}}\max_{\mathbf{u}} \frac{1}{d}\sum_{i=1}^{n}\Big[u_i(m + b y_i) - \widehat{\ell}(u_i) + \sqrt{\frac{\Delta}{d}}u_i y_i \mathbf{z}_i^\top\mathbf{w}\Big], \end{aligned}$$

where in reaching the second equality we have used the fact that any $\mathbf{w}\in\mathcal{S}_{q,m}$ satisfies the equality $m = \frac{1}{d}\mathbf{w}^\top\mathbf{v}^*$. Introduce an auxiliary problem

$$\begin{aligned} \widetilde{\mathcal{L}}_\lambda(q,m,b) &= \frac{\lambda q}{2} + \min_{\mathbf{w}\in\mathcal{S}_{q,m}}\max_{\mathbf{u}}\left\{\frac{1}{d}\sum_{i=1}^{n}\Big[u_i(m+by_i) - \widehat{\ell}(u_i)\Big] + \sqrt{\frac{\Delta}{d}}\|\mathbf{u}\|\frac{\mathbf{g}^\top\mathbf{w}}{d} + \sqrt{\Delta q}\Big(\frac{1}{d}\sum_{i=1}^{n}u_i y_i s_i\Big)\right\} \\ &= \frac{\lambda q}{2} + \min_{\mathbf{w}\in\mathcal{S}_{q,m}}\max_{\mathbf{u}}\left\{\frac{1}{d}\sum_{i=1}^{n}\Big[u_i h_i - \widehat{\ell}(u_i)\Big] + \sqrt{\frac{\Delta}{d}}\|\mathbf{u}\|\frac{\mathbf{g}^\top\mathbf{w}}{d}\right\}, \end{aligned}$$

where $\mathbf{g} = (g_1, g_2, \ldots, g_d)^\top$ and $\mathbf{s} = (s_1, s_2, \ldots, s_n)$ are two independent random vectors whose entries are drawn from the i.i.d. standard normal distribution, and $h_i = \sqrt{\Delta q}(y_i s_i) + m + b y_i$. As $y_i \in \{\pm 1\}$, independent of s_i, we note that h_i has the same probability distribution as the quantity defined in (33) in the main text.

Gordon's minimax inequalities [7–9] allow us to make the following comparison: For any constants c and $\delta > 0$, we have

$$\mathbb{P}(\mathcal{L}_\lambda(q,m,b) < c) \le 2\,\mathbb{P}(\widetilde{\mathcal{L}}_\lambda(q,m,b) < c). \tag{C.1}$$

To connect this to the statements in Proposition 3, we note that

$$\begin{aligned}
\widetilde{\mathcal{L}}_\lambda(q,m,b) &\ge \frac{\lambda q}{2} + \max_{\mathbf{u}} \min_{\mathbf{w}\in\mathcal{S}_{q,m}} \left\{ \frac{1}{d}\sum_{i=1}^n \Big[u_i h_i - \widehat{\ell}(u_i)\Big] + \sqrt{\frac{\Delta}{d}}\|\mathbf{u}\|\frac{\mathbf{g}^\top \mathbf{w}}{d} \right\} \\
&= \frac{\lambda q}{2} + \max_{\mathbf{u}} \left\{ \frac{1}{d}\sum_{i=1}^n \Big[u_i h_i - \widehat{\ell}(u_i)\Big] - \sqrt{\frac{\Delta\|\mathbf{u}\|^2(q-m^2)}{d}}\frac{\|\mathbf{g}\|}{\sqrt{d}} \right\} \\
&= \mathcal{E}_\lambda^{(d)}(q,m,b).
\end{aligned}$$

It follows that

$$\mathbb{P}(\widetilde{\mathcal{L}}_\lambda(q,m,b) < c) \le \mathbb{P}(\mathcal{E}_\lambda^{(d)}(q,m,b) < c).$$

Combining this inequality with (C.1) gives us the first inequality in Proposition 3. To obtain the second inequality in the proposition, we use the fact that the unconstrained optimization problem in (22) for the global training loss $\mathcal{L}^*$ is convex. Following exactly the same strategy as used in [9], we can interchange the order of min and max in the dual formulation of (22), which then allows us to reach the two-sided inequality in (35).

2. Proof of Lemma 1

We first rewrite the optimization problem in (32) as

$$\max_{\mu\ge 0} \max_{\|\boldsymbol{u}\|^2/d=\mu} \left\{ -\sqrt{\Delta_d \mu(q-m^2)} + \frac{\boldsymbol{u}^\top \boldsymbol{h}}{d} - \frac{1}{d}\sum_{i=1}^n \widetilde{\ell}(u_i) \right\}. \tag{C.2}$$

For the inner maximization, the constraint on the squared norm $\|\mathbf{u}\|^2$ weakly couples different coordinates of $\boldsymbol{u}$ together. To fully decouple these coordinates, we introduce a Lagrangian function

$$\frac{\boldsymbol{u}^\top \boldsymbol{h}}{d} - \frac{1}{d}\sum_{i=1}^n \widetilde{\ell}(u_i) - \frac{\gamma}{2d}(\|\mathbf{u}\|^2 - \mu d),$$

where $\gamma > 0$ is the Lagrange multiplier. For any fixed γ, the optimal solution $\boldsymbol{u}_\gamma \in \mathbb{R}^n$ can be obtained by setting the gradient of the Lagrangian function to zero, which gives us

$$\nabla\widetilde{\ell}(\boldsymbol{u}_\gamma) + \gamma \boldsymbol{u}_\gamma = \boldsymbol{h}.$$

Since there is a one-to-one correspondence between the Lagrange multiplier γ and the normalized squared norm $\mu = \|\boldsymbol{u}_\gamma\|^2/d$, it is thus equivalent to solve (C.2) in terms of

$$\max_{\gamma>0} \Big\{ -\sqrt{\frac{\Delta_d(q-m^2)\|\boldsymbol{u}_\gamma\|^2}{d}} + \frac{\boldsymbol{u}_\gamma^\top h}{d} - \frac{1}{d}\sum_{i=1}^n \widetilde{\ell}(u_{\gamma,i}) \Big\}$$

and thus we get (36).

3. Proof of Proposition 1

We first establish (26) for the special case where the subset Ω is a singleton. In this case, we just need to show

$$\mathbb{P}\Big(\mathcal{L}_\lambda(q,m,b) \ge \mathcal{E}_\lambda(q,m,b) - \delta\Big) \to 1. \tag{C.3}$$

for any fixed q, m and b.

Recall the characterization of $\mathcal{E}_\lambda^{(d)}(q,m,b)$ given in Lemma 1. The problem in (36) reaches its maximum at a point γ_d^* where the derivative of the function to be maximized is equal to 0. In calculating this derivative, we need the quantity $\frac{du_{\gamma,i}}{d\gamma}$, which can be obtained as

$$\widetilde{\ell}''(u_{\gamma,i})\frac{du_{\gamma,i}}{d\gamma} + u_{\gamma,i} + \gamma\frac{du_{\gamma,i}}{d\gamma} = 0$$

and thus $\frac{du_{\gamma,i}}{d\gamma} = \frac{-u_{\gamma,i}}{\widetilde{\ell}''(u_{\gamma,i})+\gamma}$. Using this expression and after some simple manipulations, we get

$$\alpha(\gamma_d^*)^2\frac{\|\boldsymbol{u}_{\gamma_d^*}\|^2}{n} = \Delta_d(q-m^2). \tag{C.4}$$

Moreover,

$$\mathcal{E}_\lambda^{(d)}(q,m,b) = \frac{\sum_{i=1}^n\left[u_{\gamma_d^*,i}\widetilde{\ell}'(u_{\gamma_d^*,i}) - \widetilde{\ell}(u_{\gamma_d^*,i})\right]}{d} + \frac{\lambda q}{2}. \tag{C.5}$$

Next, we introduce the following scalar change of variables: $v_{\gamma,i} = \widetilde{\ell}'(u_{\gamma,i})$. It is easy to verify from properties of Legendre transformations that

$$u_{\gamma,i} = \ell'(v_{\gamma,i}) \quad \text{and} \quad u_{\gamma,i}\widetilde{\ell}'(u_{\gamma,i}) - \widetilde{\ell}(u_{\gamma,i}) = \ell(v_{\gamma,i}).$$

Substituting these identities, we can characterize $v_{\gamma,i}$ via the implicit equation

$$v_{\gamma,i} + \gamma\ell'(v_{\gamma,i}) = h_i. \tag{C.6}$$

Moreover, (C.4) can now be rewritten as

$$\alpha(\gamma_d^*)^2\frac{1}{n}\sum_{i=1}^n[\ell'(v_{\gamma_d^*,i})]^2 = \Delta_d(q-m^2) \tag{C.7}$$

and more importantly, (C.5) can be simplified as

$$\mathcal{E}_\lambda^{(d)}(q,m,b) = \frac{\alpha}{n}\sum_{i=1}^n\ell(v_{\gamma_d^*,i}) + \frac{\lambda q}{2}.$$

Let v_γ be a random variable defined via the implicit equation

$$v_\gamma + \gamma\ell'(v_\gamma) = h, \tag{C.8}$$

where $h = \sqrt{\Delta q}s + m + by$ with $S \sim \mathcal{N}(0,1)$ and y being a random variable independent of s such that

$$\mathbb{P}(y=1) = \rho \quad \text{and} \quad \mathbb{P}(y=-1) = 1-\rho.$$

Since the loss function $\ell(\cdot)$ is convex, the function $v + \gamma\ell'(v)$ is strictly increasing. It follows that the distribution function of v_γ is given as in (24). As $n, d \to \infty$ with d/n fixed at α, we have $\Delta_d \to \Delta$ and

$$\frac{1}{n}\sum_{i=1}^n[\ell'(v_{\gamma,i})]^2 \to \mathbb{E}[(\ell'(v_\gamma))^2]$$

uniformly over any compact subset of γ. It follows that γ_d^* as defined in (C.7) converges to γ^*, which is the unique solution of (25). Moreover, we have

$$\mathcal{E}_\lambda^{(d)}(q,m,b) \to \mathcal{E}_\lambda(q,m,b) = \alpha\mathbb{E}[\ell(v_{\gamma^*})] + \frac{\lambda q}{2}. \tag{C.9}$$

For any $\delta > 0$, we can apply Proposition 3 to get

$$\mathbb{P}(\mathcal{L}_\lambda(q,m,b) < \mathcal{E}_\lambda(q,m,b) - \delta) \le 2\mathbb{P}(\mathcal{E}_\lambda^{(d)}(q,m,b) < \mathcal{E}_\lambda(q,m,b) - \delta).$$

As the right-hand side tends to 0 due to (C.9), we have (C.3).

Let Ω be an arbitrary compact subset of $\big\{(q,m,b) : m^2 \le q\big\}$. We denote by Ω_K a finite subset of Ω consisting of K points, *i.e.*, $\Omega_K = \{(q_k, m_k, b_k) \in \Omega : 1 \le k \le K\}$.

$$\begin{aligned}\mathbb{P}(\mathcal{L}_\lambda(\Omega_K) < \mathcal{E}_\lambda(\Omega) - \delta) &= \mathbb{P}(\cup_{k=1}^K \{\mathcal{L}_\lambda(q_k,m_k,b_k) < \mathcal{E}_\lambda(\Omega) - \delta\}) \\ &\le \sum_{k=1}^K \mathbb{P}(\mathcal{L}_\lambda(q_k,m_k,b_k) < \mathcal{E}_\lambda(\Omega) - \delta) \\ &\le \sum_{k=1}^K \mathbb{P}(\mathcal{L}_\lambda(q_k,m_k,b_k) < \mathcal{E}_\lambda(q_k,m_k,b_k) - \delta).\end{aligned}$$

As $n \to \infty$, the right-hand side of the inequality tends to 0. It follows that $\mathbb{P}(\mathcal{L}_\lambda(\Omega_K) \ge \mathcal{E}_\lambda(\Omega) - \delta) \to 1$. Note that this characterization holds for any finite K. From the smoothness of the optimization problem (21), one can construct a family of subsets $\{\Omega_K\}$ such that $\mathcal{L}_\lambda(\Omega_K) \to \mathcal{L}_\lambda(\Omega)$ as $K \to \infty$, and thus we have (26). This strategy follows closely the approach used in [9]. Finally, to get (27), we first note that (26) implies that

$$\mathbb{P}\Big(\mathcal{L}_\lambda^* \ge \mathcal{E}_\lambda^* - \delta\Big) \to 1.$$

The "other direction", *i.e.*, $\mathbb{P}\Big(\mathcal{L}_\lambda^* \le \mathcal{E}_\lambda^* + \delta\Big) \to 1$ can be obtained by exploiting the convexity of the loss function $\ell(\cdot)$, which allows us to interchange the order of min and max in the dual formulation of (22). We omit the details as they follow exactly the same strategy as used in [9].

4. Proof of Proposition 2

We start with the fixed-point equation for the Lagrange multiplier given in (25). For our proof, it will be more convenient to rewrite this equation in terms of the random variable $u_\gamma \stackrel{\text{def}}{=} \ell'(v_\gamma)$. It is a well-known property of Legendre transformations that we can write the "symmetric equation" $v_\gamma = \widetilde{\ell}'(u_\gamma)$. Since v_γ is determined via the implicit equation (C.8), we have

$$\widetilde{\ell}'(u_\gamma) + \gamma u_\gamma = h.$$

It follows that the cumulant distribution function of u_γ is given by

$$\mathbb{P}(u_\gamma \le u) = \rho Q\left(\frac{\widetilde{\ell}'(u) + \gamma u - m - b}{\sqrt{\Delta q}}\right) + (1-\rho) Q\left(\frac{\widetilde{\ell}'(u) + \gamma u - m + b}{\sqrt{\Delta q}}\right),$$

where $Q(\cdot)$ is the distribution function of a standard normal random variable. Writing (25) in terms of u_γ, we have

$$\alpha\gamma^2 \mathbb{E}[u_\gamma^2] = \Delta(q - m^2). \tag{C.10}$$

Our assumption of the loss function $\ell(\cdot)$ is that it is convex and monotonically decreasing, with $\ell(+\infty) = \ell'(+\infty) = 0$. It follows that $\ell'(-\infty) < u_\gamma < 0$. Introducing the changes of variables $\theta \stackrel{\text{def}}{=} m/\sqrt{q}$, $\widetilde{b} \stackrel{\text{def}}{=} b/\sqrt{q}$ and $\widetilde{\gamma} = \gamma/\sqrt{q}$, and using the identity $\mathbb{E}[u_\gamma^2] = (-2)\int_{\ell'(-\infty)}^0 u \mathbb{P}(u_\gamma \le u) du$, we can rewrite (C.10) as

$$\alpha S(\widetilde{\gamma}, q, \theta) = \Delta(1 - \theta^2), \tag{C.11}$$

where

$$S(\widetilde{\gamma}, q, \theta) \stackrel{\text{def}}{=} \widetilde{\gamma}^2 \int_0^{-\ell'(-\infty)} (2u) \left(\rho Q\Big(\frac{\widetilde{\ell}'(-u)}{\sqrt{\Delta q}} + \frac{-\widetilde{\gamma} u - \theta - \widetilde{b}}{\sqrt{\Delta}}\Big) + (1-\rho) Q\Big(\frac{\widetilde{\ell}'(-u)}{\sqrt{\Delta q}} + \frac{-\widetilde{\gamma} u - \theta + \widetilde{b}}{\sqrt{\Delta}}\Big)\right) du.$$

We further denote by $\widehat{\gamma}^*(q,\theta)$ the solution to (C.11). We can show that, for any fixed $\widetilde{\gamma}$ and θ, the function $S(\widetilde{\gamma}, q, \theta)$ is monotonically decreasing as we increase q. Moreover,

$$\lim_{q\to\infty} S(\widetilde{\gamma}, q, \theta) = S^*(\widetilde{\gamma}, \theta) \stackrel{\text{def}}{=} \int_0^{-\widetilde{\gamma}\ell'(-\infty)} (2u) \Big[\rho Q\Big(\frac{-u - \theta - \widetilde{b}}{\sqrt{\Delta}}\Big) + (1-\rho) Q\Big(\frac{-u - \theta + \widetilde{b}}{\sqrt{\Delta}}\Big)\Big] du.$$

Clearly, $S^*(\widetilde{\gamma},\theta)$ is monotonic with respect to $\widetilde{\gamma}$, but it has a finite limit as $\widetilde{\gamma}\to\infty$, *i.e.*,

$$\lim_{\widetilde{\gamma}\to\infty} S^*(\widetilde{\gamma},\theta) = \Delta\int_0^\infty u^2\Big[\rho f\Big(u+\frac{\theta+\widetilde{b}}{\sqrt{\Delta}}\Big)+(1-\rho)f\Big(u+\frac{\theta-\widetilde{b}}{\sqrt{\Delta}}\Big)\Big],$$

where $f(\cdot)$ is the probability density function of $\mathcal{N}(0,1)$. An implication of this limit being finite is that, although the Lagrange multiplier $\widehat{\gamma}^*(q,\theta)$ remains finite for any fixed q, it tends to ∞ as $q\to\infty$ if

$$\alpha < \frac{1-\theta^2}{S^*(\infty,\theta)}. \tag{C.12}$$

It follows from (C.8) that, as $\gamma\to\infty$, $\ell'(v_\gamma)\to 0$ and thus $v_\gamma\to\infty$. Consequently,

$$\lim_{q\to\infty}\mathcal{E}_{\lambda=0}(q,m,b) = \lim_{q\to\infty}\alpha\mathbb{E}[\ell(v_{\gamma^*(q,\theta)})]\to 0.$$

This characterization can be interpreted as follows: If there exists a θ that satisfies (C.12), then as we move along the "ray" of constant slope $\theta = m/\sqrt{q}$, the training loss $\mathcal{E}_{\lambda=0}(q,m,b)$ will tend to 0. The critical threshold α^* can then be obtained by maximizing the right-hand side of (C.12), which gives us the final expression as stated in Proposition 2.

5. Derivation of Theorem 1 from Gordon's characterization

In this section, we show that the fixed point equations in Theorem 1 can be mapped to Gordon's characterization, namely (25) and (27) in the main text. First of all, we observe that (25) is trivially satisfied by the solution of system (4)-(9). Then, we consider the minimization of $\mathcal{E}_\lambda(q,m,b)$, derived in (C.9), with respect to q,m,b. This simply amounts to setting the derivatives to zero. Note that the partial derivatives of v and γ^* can be computed by taking the derivatives of both sides of (C.6) and (25) respectively. The minimization leads to the following system of equations:

$$\alpha\sqrt{\frac{\Delta}{q}}\mathbb{E}_{y,s}\left[\ell'(v_{\gamma^*})s\right]+\lambda = \frac{\Delta}{\gamma}, \tag{C.13}$$

$$m = -\alpha\frac{\gamma}{\Delta}\mathbb{E}_{y,s}\left[\ell'(v_{\gamma^*})\right], \tag{C.14}$$

$$\mathbb{E}_{y,s}\left[y\ell'(v_{\gamma^*})\right] = 0, \tag{C.15}$$

where $s\sim\mathcal{N}(0,1)$, $y=+1$ with probability $\rho\in(0,1)$ and $y=-1$ otherwise. We observe that (C.15) is the same as (11) and (C.14) is equivalent to (4) and (6). Using again (6), we can rewrite (C.13) as

$$\hat{\gamma} = \alpha\sqrt{\frac{\Delta}{q}}\mathbb{E}_{y,s}\left[\ell'(v_{\gamma^*})s\right]. \tag{C.16}$$

Note that $\ell'(v_{\gamma^*}(h(s)))$ is a function of s, and ℓ'' is well defined. Therefore, we can apply Stein's lemma and rewrite

$$\hat{\gamma} = \alpha\sqrt{\frac{\Delta}{q}}\mathbb{E}_{y,s}\left[\partial_s v_{\gamma^*}\ell''(v_{\gamma^*})\right], \tag{C.17}$$

which leads to an identity if we substitute the definition of $\hat{\gamma}$ provided in (9).

Appendix D: Evaluation of the fixed point equations

In this section we will compute the fixed-point equations for the square and hinge loss. The equations for the logistic loss cannot be computed analytically and require numerical integration.

1. Square loss

In this case, $\ell(\omega) = \frac{1}{2}(\omega - 1)^2$ and the fixed point equations (4)-(9) can be inverted analytically. The minimizer v, defined as

$$v \equiv \underset{\omega}{\operatorname{argmin}} \frac{(\omega - h(y, m, q, b))^2}{2\gamma} + \frac{1}{2}(\omega - 1)^2, \tag{D.1}$$

is simply

$$v = h - \gamma l'(v) = \frac{h + \gamma}{1 + \gamma}, \tag{D.2}$$

where $h \sim \mathcal{N}(m + yb, \Delta q)$. Hence, we obtain

$$\hat{m} = \frac{\alpha}{\gamma}\mathbb{E}_{y,h}\left[v(y, h, \gamma) - h\right] = \frac{\alpha}{1 + \gamma}\left(1 - m - (2\rho - 1)b\right), \tag{D.3}$$

$$\hat{q} = \frac{\alpha\Delta}{\gamma^2}\mathbb{E}_{y,h}\left[(v(y, h, \gamma) - h)^2\right] = \frac{\alpha\Delta}{(1 + \gamma)^2}\left(\Delta q + \mathbb{E}_y\left[(1 - m - yb)^2\right]\right), \tag{D.4}$$

$$\hat{\gamma} = \frac{\alpha\Delta}{\gamma}\left(1 - \mathbb{E}_{y,h}\left[\partial_h v(y, h, \gamma)\right]\right) = \frac{\alpha\Delta}{1 + \gamma}. \tag{D.5}$$

To compute the bias b, we have to solve

$$0 = \mathbb{E}_{y,h}\left[y(v - h)\right] = \frac{\gamma}{1 + \gamma}\mathbb{E}_{y,h}\left[y(1 - h)\right], \tag{D.6}$$

which simply gives

$$b = (2\rho - 1)(1 - m). \tag{D.7}$$

We can plug (D.3)-(D.5) in the equations for m, q, γ to obtain

$$\gamma = \frac{\Delta}{\lambda + \hat{\gamma}} = \frac{\Delta(1 - \alpha) - \lambda + \sqrt{(\Delta(1 - \alpha) - \lambda)^2 + 4\lambda\Delta}}{2\lambda}, \tag{D.8}$$

$$m = \frac{\hat{m}}{\lambda + \hat{\gamma}} = \frac{4\alpha\gamma\rho(1 - \rho)}{\Delta(1 + \gamma) + 4\alpha\gamma\rho(1 - \rho)}, \tag{D.9}$$

$$q = \frac{\hat{q} + \hat{m}^2}{(\lambda + \hat{\gamma})^2} = \frac{1}{(1 + \gamma)^2 - \alpha\gamma^2}\left(\frac{\alpha\gamma^2}{\Delta}((1 - m)^2 - b^2) + (1 + \gamma)^2 m^2\right). \tag{D.10}$$

2. Hinge loss

In this case, $\ell(\omega) = \max\{0, 1 - \omega\}$ and the minimizer

$$v \equiv \underset{\omega}{\operatorname{argmin}} \frac{(\omega - h(y, m, q, b))^2}{2\gamma} + \max\{0, 1 - \omega\}, \tag{D.11}$$

is piece-wise defined as

$$v = \begin{cases} h & \text{if} \quad h > 1 \\ 1 & \text{if} \quad 1 - \gamma < h < 1 \\ h + \gamma & \text{if} \quad h < 1 - \gamma \end{cases}. \tag{D.12}$$

From (4)-(9), it follows that

$$\gamma = \frac{\gamma}{K_\gamma}, \tag{D.13}$$

$$m = \frac{\alpha}{\Delta}\frac{K_m}{K_\gamma}, \tag{D.14}$$

$$q = \frac{\alpha}{\Delta K_\gamma^2}\left(K_q + \frac{\alpha}{\Delta}K_m^2\right), \tag{D.15}$$

where we have defined

$$K_\gamma = \frac{\lambda\gamma}{\Delta} + \alpha\left(1 - \mathbb{E}_y\left[Q\left(\frac{1-m-yb}{\sqrt{\Delta q}}\right) + Q\left(\frac{\gamma-(1-m-yb)}{\sqrt{\Delta q}}\right)\right]\right), \tag{D.16}$$

$$\begin{aligned} K_m = \sqrt{\frac{\Delta q}{2\pi}}\mathbb{E}_y\left[\exp\left(-\frac{(1-m-yb)^2}{2\Delta q}\right) - \exp\left(-\frac{(\gamma-(1-m-yb))^2}{2\Delta q}\right)\right] \\ +\mathbb{E}_y\left[(1-m-yb)\left(1-Q\left(\frac{1-m-yb}{\sqrt{\Delta q}}\right) - Q\left(\frac{\gamma-(1-m-yb)}{\sqrt{\Delta q}}\right)\right) + \gamma Q\left(\frac{\gamma-(1-m-yb)}{\sqrt{\Delta q}}\right)\right], \end{aligned} \tag{D.17}$$

$$\begin{aligned} K_q = \sqrt{\frac{\Delta q}{2\pi}}\mathbb{E}_y\left[(1-m-yb)\exp\left(-\frac{(1-m-yb)^2}{2\Delta q}\right) - (\gamma+1-m-yb)\exp\left(-\frac{(\gamma-(1-m-yb))^2}{2\Delta q}\right)\right] \\ +\mathbb{E}_y\left[\left(\Delta q + (1-m-yb)^2\right)\left(1-Q\left(\frac{1-m-yb}{\sqrt{\Delta q}}\right) - Q\left(\frac{\gamma-(1-m-yb)}{\sqrt{\Delta q}}\right)\right) + \gamma^2 Q\left(\frac{\gamma-(1-m-yb)}{\sqrt{\Delta q}}\right)\right]. \end{aligned} \tag{D.18}$$

The equation to determine the bias is

$$\begin{aligned} \sqrt{\frac{\Delta q}{2\pi}}\mathbb{E}_y\left[y\exp\left(-\frac{(1-m-yb)^2}{2\Delta q}\right) - y\exp\left(-\frac{(\gamma-(1-m-yb))^2}{2\Delta q}\right)\right] + \gamma\mathbb{E}_y\left[yQ\left(\frac{\gamma-(1-m-yb)}{\sqrt{\Delta q}}\right)\right] \\ +\mathbb{E}_y\left[y(1-m-yb)\left(1-Q\left(\frac{1-m-yb}{\sqrt{\Delta q}}\right) - Q\left(\frac{\gamma-(1-m-yb)}{\sqrt{\Delta q}}\right)\right)\right] = 0. \end{aligned} \tag{D.19}$$

Appendix E: Bayes-optimality at $\lambda = \infty$, for $\rho = \frac{1}{2}$

In this section we will show how the result on Bayes-optimality for balanced clusters at large regularization arises. First we start by considering the square loss. At $\rho = 1/2$, it is straightforward to check from (11) that $b = 0$ and the generalization error, given by (12) in the main text, is

$$\varepsilon_{\text{gen}} = Q\left(\frac{m}{\sqrt{\Delta q}}\right), \tag{E.1}$$

where m and q are given by (D.9)-(D.10), evaluated at $\rho = \frac{1}{2}$. The Bayes-optimal error for this problem is given by (19) in the main text and reads

$$\varepsilon_{\text{gen}}^{\text{BO}} = Q\left(\sqrt{\frac{\alpha}{\Delta(\Delta+\alpha)}}\right). \tag{E.2}$$

Therefore, in order to reach Bayes-optimality, we need a weight vector $\mathbf{w}$ with an overlap m and a length q such that

$$\sqrt{\frac{\alpha}{(\Delta+\alpha)}} = \frac{m}{\sqrt{q}} = \left(\sqrt{\frac{\hat{q}}{\hat{m}^2}+1}\right)^{-1}. \tag{E.3}$$

By using (D.3)-(D.4) evaluated at $\rho = \frac{1}{2}$, (E.3) can be rewritten as

$$\frac{\Delta q}{(1-m)^2} = 0. \tag{E.4}$$

Eq. (E.4) is verified by the fixed point equations only at $\lambda \to \infty$. Indeed in this limit we find that

$$\gamma = \frac{\Delta}{\lambda} + o\left(\lambda^{-1}\right),$$

hence

$$m = \frac{\alpha}{\lambda} + o\left(\lambda^{-1}\right)$$

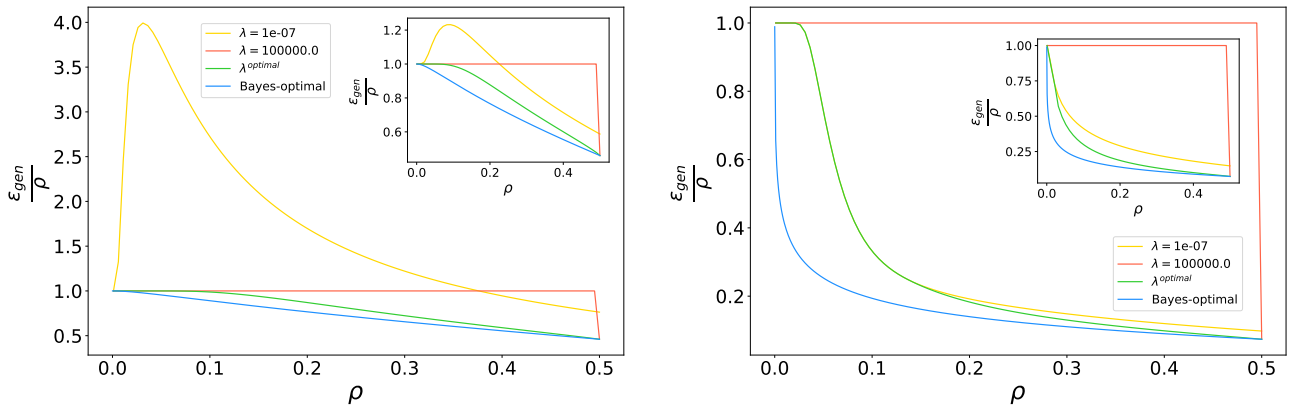

Figure 6. Generalization error as a function of ρ, at fixed $\alpha = 1.2$ and $\Delta = 1$ (left) and $\alpha = 7$ and $\Delta = 0.3$ (right), for the square loss compared to the Bayes-optimal performance. In the inset, the same figure for the hinge loss. The vertical axis is rescaled by ρ for convenience. The error is computed at low ($\lambda = 10^{-7}$), high ($\lambda = 10^{5}$) and optimal regularization. We observe that Bayes-optimality at infinite regularization holds strictly at $\rho = 1/2$.

and

$$q = \frac{\alpha}{\lambda^2}(\Delta + \alpha) + o(\lambda^{-2}),$$

so that

$$\frac{m}{\sqrt{q}} \to \sqrt{\frac{\alpha}{(\Delta + \alpha)}}.$$

Therefore, as λ grows and while the ℓ_2 norm of the vector goes to zero, the vector aligns itself optimally to the hidden one and the generalization error becomes optimal.

It is then easy to see why this remains correct for any differentiable loss: as long as the ℓ_2 norm vanishes when $\lambda \to \infty$, then one can expand

$$\ell(\mathbf{w}^\top \mathbf{x}) = \ell(\mathbf{0}) + \mathbf{w}^\top \mathbf{x}\ell'(\mathbf{0}) + \mathbf{o}(\mathbf{q})$$

so that any loss will behave like the square one. This is the origin of the peculiar behavior of Bayes optimally observed at $\lambda \to \infty$ for the symmetric case $\rho = 1/2$. We observed numerically that this result is not valid anymore as soon as $\rho \neq 1/2$. This peculiar behaviour is shown in Fig. 6, which depicts the generalization error, computed from the solution of (4)-(11) in the main text, as a function of ρ at zero, infinite and optimal regularization for the square and hinge losses.

Appendix F: Details on the numerics

1. Iteration of the fixed point equations

The solution (q, m, b, γ) of the fixed point equations (4)-(9) can be obtained analytically only in the case of square loss. For the hinge and logistic loss, the equations have to be iterated until convergence. In our codes we used initialization $(q^{t=0}, \gamma^{t=0}, m^{t=0}, b^{t=0}) = (0.5, 0.5, 0.01, 0)$. The stopping criterion for convergence consists in checking if the values of the generalization error at two consecutive iterations differ less than a threshold *eps*. In all figures, we used $eps \leq 10^{-8}$.

2. Simulations

In order to check the validity of the fixed point equations (4)-(9) we computed numerically the solution of the optimization problem defined in (2), and we averaged over multiple realizations of the noise. In the case of square loss,

the solution is simply

$$\mathbf{w}^{\text{square}} = \left(\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}_d\right)^{-1} \mathbf{X}^\top \mathbf{y}. \tag{F.1}$$

In the case of logistic and hinge loss, the solution can be computed by a standard gradient descent algorithm. In particular, in Fig. 1 we used the Logistic Regression classifier provided by the scikitlearn package *linear_model* [29]. In particular, we used the "lbfgs" solver, with L2-penalty, tolerance $tol = 10^{-5}$ for the stopping criterion and maximum number of iterations $max\_iter = 10^{-5}$. It is important to remind that all our analytic results are computed in the infinite-dimensional limit $d, n \to \infty$, while the ratio $\alpha = n/d$ remains finite. Therefore, all the simulations involve errors due to finite size effects. However, we found a very good agreement bewteen theory and simulations already at relatively small dimensionality ($d \leq 5000$). The only case in which finite size effects prevent simulations to match our theoretical predictions is the behaviour of the generalization error at large regularization λ, at $\rho = 1/2$. Since at all finite dimensions d the effective clusters size is $\rho \neq 1/2$, the result of reaching Bayes-optimality at $\lambda \to \infty$ cannot be obtained in simulations, since it holds strictly at $\rho = 1/2$. However, we obtain greater and greater precision, i.e. the minimum of the generalization error moving towards higher values of λ (see Fig. 4), as d increases.