The Role of Regularization in Classification of High-dimensional Noisy Gaussian Mixture

Francesca Mignacco [1] **Florent Krzakala** [2] **Yue M. Lu** [3] **Pierfrancesco Urbani** [1] **Lenka Zdeborová** [1]

Abstract

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.

1. Introduction

High-dimensional statistics where both the dimensionality d, and the 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 (Hastie et al., 2019; Sur & Candès, 2019) 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 $\pm\frac{\mathbf{v}^*}{\sqrt{d}}$, $\mathbf{v}^* \in \mathbb{R}^d$, 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 $\mathbf{v}^*$ 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 (Friedman et al., 2001). For the supervised model considered here, (Lelarge & Miolane, 2019) 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 (Mai & Liao, 2019), the authors study the same data generative model in the particular case of equally sized clusters, and analyze *non-regularized losses* under the

[1]Université Paris-Saclay, CNRS, CEA, Institut de Physique théorique, 91191, Gif-sur-Yvette, France [2]Laboratoire de physique de l'École normale supérieure, ENS, Université PSL, CNRS, Sorbonne Université, Université de Paris, F-75005, Paris, France [3]John A. Paulson School of Engineering and Applied Sciences, Harvard University, Cambridge, MA 02138, USA . Correspondence to: Francesca Mignacco <francesca.mignacco@ipht.fr>.

Proceedings of the 37^{th} International Conference on Machine Learning, Online, PMLR 119, 2020.

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. (Deng et al., 2019) studied the logistic loss, again without regularization and for two clusters of equal size, and derive the linear separability condition in this case. (Kini & Thrampoulidis, 2020) carried out the same analysis for the case of square loss. The effect of data structure on learning a linearly separable rule defined by a perceptron teacher was studied in a model of two clusters of binary input data already in (Maeangi et al., 1995).

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 (Gordon, 1985; 1988; Thrampoulidis et al., 2015), as in (Deng et al., 2019; Kini & Thrampoulidis, 2020) for the case of Gaussian mixture, and in (Salehi et al., 2019; Montanari et al., 2019) for the case of i.i.d. Gaussian input data and labels generated by a single-layer teacher network. 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 the 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. (2004) 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 (Soudry et al., 2018), and we discuss this in connection with the "double-descent" phenomenon that is currently the subject of intense studies (Belkin et al., 2019).

The existence of a sharp transition for perfect separability in the model, with and without bias, is interesting in itself. Recently (Candès & Sur, 2018) 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 (Gardner, 1988; Gardner & Derrida, 1989; Krauth & Mézard, 1989).

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 (Del Giudice et al., 1989; Franz et al., 1990) 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 mixture 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.

2. Main theoretical results

2.1. 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 $y_i = 1$, 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 wrongly 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

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

2.2. 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:

$$\begin{aligned}\log \mathrm{p}(y|\mathrm{x}) &= \log\frac{\mathrm{p}(\mathrm{x}|y)\mathrm{p}_y(y)}{\sum_{y=\pm1}\mathrm{p}(\mathrm{x}|y)\mathrm{p}_y(y)} \\ &= -\log\left(1+\exp(-c)\right),\end{aligned} \tag{14}$$

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

$$\hat{y}_{\mathrm{new}} = \arg\max_{y\in\pm1}\log\mathrm{p}\left(y|\{\mathbf{X},\mathbf{y}\},\mathbf{x}_{\mathrm{new}}\right), \tag{15}$$

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):

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

and we have

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

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 supplementary material. The case $\rho = 1/2$ was also discussed in (Dobriban et al., 2018; Lelarge & Miolane, 2019).

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 (Hebb, 2005). Consider indeed the weight vector averaged over the training samples, each multiplied by its label and rescaled by $\sqrt{d}$

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

It is straightforward to check that, for $\hat{\mathbf{w}}_{\mathrm{Hebb}}$, one has in large dimension limit $m = 1$ and $q = (1+\frac{\Delta}{\alpha})$. In the case of balanced clusters, the Hebb estimator is unbiased by definition, since the noise has zero mean. In the more interesting case of non balanced mixture of Gaussians, one further needs to optimize the intercept b in the linear fit $a\mathbf{x}+b$. Since the minimizer of the generalization error with respect to the bias is unique, this parameter can be optimized in a number of ways, including gradient descent or cross validation. For instance, we obtain that k-fold cross validation can get extremely close to the optimal value $b = \frac{\Delta q}{2m}\log\frac{\rho}{1-\rho}$. If one then plugs the optimized value for the bias in eq. (12), one reaches the 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.

2.3. 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

$$\mathcal{L}_\lambda^* \stackrel{\text{def}}{=} \min_{m^2 \le q, b} \mathcal{L}_\lambda(q, m, b), \tag{22}$$

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

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

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

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

$$\mathbb{P}(v_\gamma \le v) = \rho Q\big(\frac{\gamma \ell'(v) + v - m - b}{\sqrt{\Delta q}}\big) + (1-\rho) Q\big(\frac{\gamma \ell'(v) + v - m + b}{\sqrt{\Delta q}}\big).$$

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

$$\alpha \gamma^2 \mathbb{E}[(\ell'(v_\gamma))^2] = \Delta(q - m^2). \tag{24}$$

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

$$\mathbb{P}\Big(\mathcal{L}_\lambda(\Omega) \ge \mathcal{E}_\lambda(\Omega) - \delta\Big) \to 1. \tag{25}$$

Moreover,

$$\mathcal{L}_\lambda^* \to \mathcal{E}_\lambda^* \stackrel{def}{=} \inf_{m^2 \le q, b} \mathcal{E}_\lambda(q, m, b), \tag{26}$$

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

The characterization in (26) shows that the global training loss will concentrate around the fixed value $\mathcal{E}_\lambda^*$. Meanwhile, (25) 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 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

$$\alpha^* \stackrel{def}{=} \max_{0 \le r \le 1, b} \eta(r, b) \tag{27}$$

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

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

3. Proof Sketches

In this section, we sketch the proof steps behind our main results presented in Section 2. The full technical details are given in the supplementary materials.

Roughly speaking, our proof strategy consists of three main ingredients: (1) Using Gordon's minimax inequalities (Gordon, 1985; 1988; Thrampoulidis et al., 2015), 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 (31) in Section 3.1) that is essentially decoupled over its coordinates; (2) we show in Section 3.2 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.

3.1. 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 $\ell(v)$:

$$\ell(v) = \max_u \Big\{ vu - \widetilde{\ell}(u) \Big\}, \tag{28}$$

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

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

For example, for the square, logistic, and hinge losses defined in Section 1, their corresponding convex conjugates

are given by

$$\widetilde{\ell}_{\text{square}}(u) = \frac{u^2}{4} + u \tag{29}$$

$$\widetilde{\ell}_{\text{logistic}}(u) = \begin{cases} -H(-u), & \text{for } -1 \le u \le 0 \\ \infty, & \text{otherwise} \end{cases}, \tag{30}$$

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

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

respectively.

Substituting (28) into (21) and recalling the data model (1), we can rewrite (21) as the following minimax problem

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

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

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

$$\mathcal{E}_\lambda^{(d)}(q,m,b) \stackrel{\text{def}}{=} \frac{\lambda q}{2} + \max_{\boldsymbol{u}\in\mathbb{R}^n}\left\{-\sqrt{\frac{\Delta\|\boldsymbol{u}\|^2(q-m^2)}{d}} + \frac{\boldsymbol{u}^\top\boldsymbol{h}}{d} - \frac{1}{d}\sum_{i=1}^n \widetilde{\ell}(u_i)\right\}, \tag{31}$$

where

$$\boldsymbol{h} = \sqrt{\Delta q}\boldsymbol{s} + m\mathbf{1} + b[y_1, y_2, \ldots, y_n]^\top \tag{32}$$

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

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

and

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

The proof of Proposition 3, which can be found in the supplementary material, is based on an application of Gordon's comparison inequalities for Gaussian processes (Gordon, 1985; 1988; Thrampoulidis et al., 2015). Similar techniques have been used by the authors of (Deng et al., 2019) to study the Gaussian mixture model for the non-regularized logistic loss for two clusters of the same size.

3.2. Asymptotic Characterizations

The definition of $\mathcal{E}_\lambda^{(d)}(q,m,b)$ in (31) 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

$$\mathcal{E}_\lambda^{(d)}(q,m,b) = \frac{\lambda q}{2} + \max_{\gamma>0}\Big\{-\sqrt{\frac{\Delta(q-m^2)\|\boldsymbol{u}_\gamma\|^2}{d}} + \frac{\boldsymbol{u}_\gamma^\top h}{d} - \frac{1}{d}\sum_{i=1}^n \widetilde{\ell}(u_{\gamma,i})\Big\}, \tag{33}$$

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

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

with $\boldsymbol{h}$ defined as in (32).

One can show that the problem in (33) reaches its maximum at a point γ^* that is the unique solution to

$$\alpha\gamma^2\frac{\|\boldsymbol{u}_\gamma\|^2}{n} = \Delta(q-m^2). \tag{34}$$

Moreover,

$$\mathcal{E}_\lambda^{(d)}(q,m,b) = \frac{\sum_{i=1}^n\left[u_{\gamma^*,i}\widetilde{\ell}'(u_{\gamma^*,i}) - \widetilde{\ell}(u_{\gamma^*,i})\right]}{d} + \frac{\lambda q}{2}. \tag{35}$$

In the asymptotic limit, as $n, d \to \infty$, both (34) and (35) converge towards their deterministic limits:

$$\alpha\gamma^2\mathbb{E}[u_\gamma^2] = \Delta(q-m^2) \tag{36}$$

and

$$\mathcal{E}_\lambda^{(d)}(q,m,b) \to \alpha\mathbb{E}[u_\gamma\widetilde{\ell}'(u_\gamma) - \widetilde{\ell}(u_\gamma)] + \frac{\lambda q}{2}, \tag{37}$$

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

Note that (36) and (37) already resemble their counterparts (24) 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 (36) and (37) then gives us the characterizations (24) and (23) as stated in Section 2.

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 leave the details to the supplementary materials.

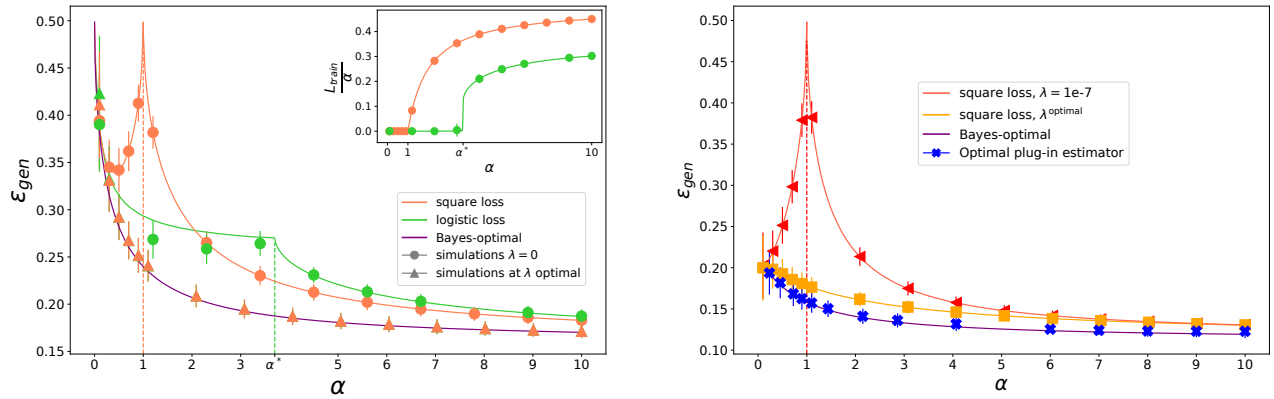

Figure 1. **Left (equal cluster size)**. Generalization error as a function of α at low regularization ($\lambda = 10^{-7}$) and fixed $\Delta = 1$, $\rho = 0.5$. The dashed vertical lines mark the interpolation thresholds. The generalization error achieved by the square and logistic losses is compared 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 $d = 5000$).

3.3. Interpretation from the replica method

These same equations can be independently derived from the non-rigorous replica method from statistical physics (Mézard et al., 1987), a technique that has proven useful in the study of high-dimensional statistical models, for instance following (Franz et al., 1990; Lesieur et al., 2016). Alternatively, these equations can also be seen as a special case of the State Evolution equation of the Approximate Message Passing algorithm (Donoho et al., 2009; Bayati & Montanari, 2011; Lesieur et al., 2016). 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 $\mathbf{w}$:

$$V = \lim_{d\to\infty} \frac{\mathbb{E}_{\mathbf{X},\mathbf{y}}\left[\|\mathbf{w}\|_2^2\right] - \mathbb{E}_{\mathbf{X},\mathbf{y}}\left[\|\mathbf{w}\|_2\right]^2}{d}. \tag{38}$$

The zero temperature limit of the fixed point equations obtained with the replica method corresponds to the loss minimization (Mézard et al., 1987; Mézard & Montanari, 2009). In this limit, the behaviour of the rescaled variance V at penalty going to zero ($\lambda \to 0$) is an indicator of data separability. In the non-separable regime, the minimizer of the loss is unique and $V \to 0$ at temperature $T \to 0$. The parameter γ turns out to be simply $\gamma = \lim_{T\to 0} V/T$. However, in the regime where data are separable there is a degeneracy of solutions at $\lambda = 0$, and the variance at zero temperature 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.

4. Applications

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 the dimension α and the cluster variance Δ. 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 recall that α^* is defined as 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 (Candès & Sur, 2018)), see (27). 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 (Sur & Candès, 2019), the curve

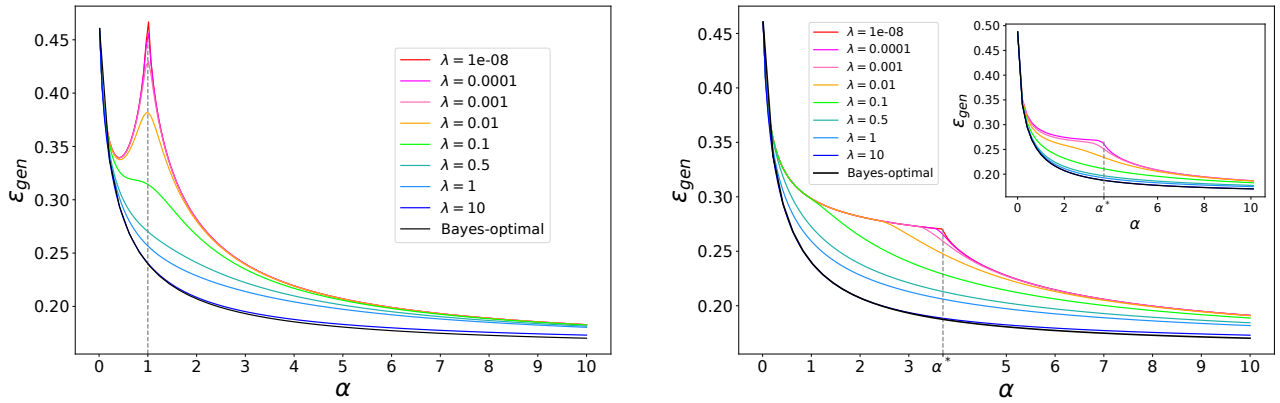

Figure 2. Generalization error as a function of α for different values of λ, at fixed $\Delta = 1$ and $\rho = 0.5$, for the square loss (left), the hinge loss (right) and the logistic loss (inset), compared to the Bayes-optimal error. If the two clusters have the same size, the Bayes-optimal error can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

that we depict is the limiting value reached as $\lambda \to 0^+$. The points are results of simulations with a standard scikitlearn (Pedregosa et al., 2011) package. As shown in (Soudry et al., 2018), even though the logistic estimator does not exist, gradient descent actually converges to the max-margin solution in this case, or equivalently to the minimal norm solution classifying all samples correctly, a phenomenon called "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 (Geiger et al., 2019; Belkin et al., 2019; Hastie et al., 2019; Mitra, 2019; Mei & Montanari, 2019), but similar observations appeared as early as 1996 in (Opper & Kinzel, 1996). 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 (Mai & Liao, 2019), 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 in 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 $\lambda_{\rm opt} \to \infty$ and the corresponding error matches exactly the Bayes-optimal error, for the non-symmetric case $0 < \lambda_{\rm opt} < \infty$ 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 supplementary material for the perhaps unexpected property of achieving the Bayes-optimal generalization at $\lambda_{\rm opt} \to \infty$ 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 supplementary material.

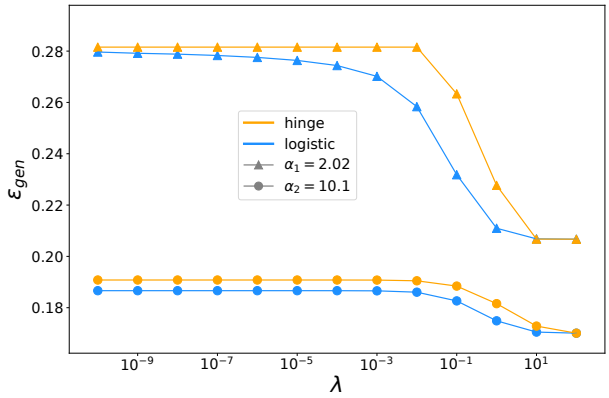

Figure 3. Generalization error as a function of λ for the hinge and logistic losses, at fixed $\Delta = 1$, $\rho = 0.5$ and two different values of α: $\alpha_1 = 2, \alpha_2 = 10$. As $\lambda \to 0^+$, the error of the two losses approaches the same value if the data are separable ($\alpha_1 < \alpha^*$). 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.

Regularization and the interpolation peak — In Fig. 2 we depict the dependence of the generalization error on

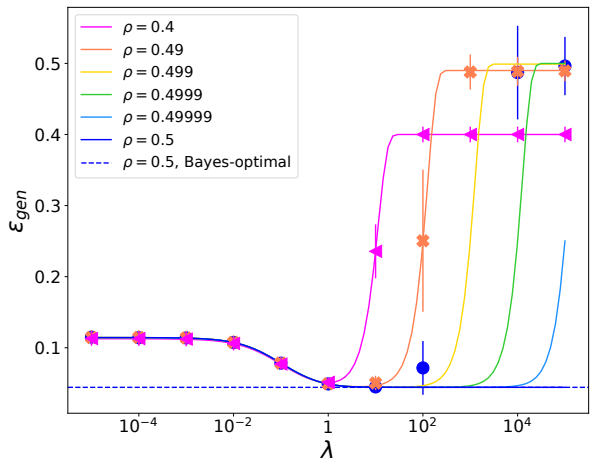

Figure 4. Generalization error as a function of λ for different values of ρ close to 0.5, at fixed $\Delta = 0.3$ and $\alpha = 2$, for the square loss. At all $\rho < 0.5$, the error exhibits a minimum at finite $\lambda = \lambda^*$, and reaches a plateau at $\lambda > \lambda^*$. The value of the error at the plateau is $\varepsilon_{\text{gen}} = \min\{\rho, 1-\rho\}$, 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, the error in simulations always reaches a plateau at large λ.

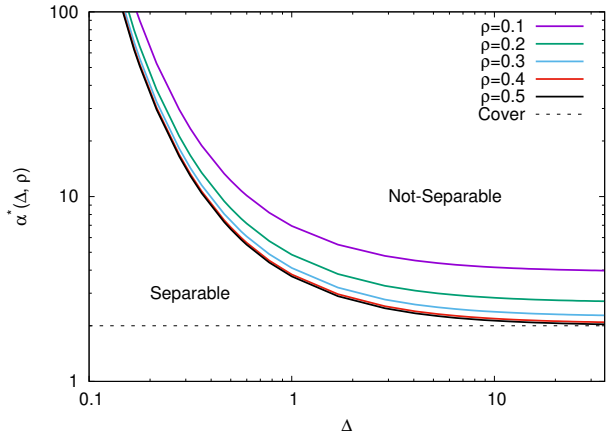

Figure 5. Critical value $\alpha = \alpha^*$, defined by Proposition 2, at which the linear separability transition occurs as a function of Δ, for different values of ρ. Similarly as for what happens for Gaussian data (Candès & Sur, 2018), the MLE does not exists on the left the curve. The line indicates 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).

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 (Hastie et al., 2019; Mei & Montanari, 2019) 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 (Geiger et al., 2019; Nakkiran et al., 2019). In fact, as the regularization increases, the error gets better in this model with equal-size clusters, and one reaches the Bayes-optimal values for large regularization.

Max-margin and weak regularization — Fig. 3 illustrates the generic property that all monotone non-increasing loss functions converge to the max-margin solution for linearly separable data as $\lambda \to 0^+$ (Rosset et al., 2004). 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 $\rho \to 0.5$. 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 the 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.

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 the classical work by (Cover, 1965). 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 (Candès & Sur, 2018). This phenomenology of the separability phase transition, or equivalently of the existence of the maximum likelihood estimator, thus seems very generic.

5. Conclusion

We have studied the performance of regularized convex classifiers at separating a mixture of two Gaussian clusters in the noisy regime when even an oracle knowing the centers of the clusters would make a finite fraction of mistakes. We have derived rigorous closed-form formulas for the generalization and training errors in the limit where the number of samples and dimensions go to infinity, while their ratio is a fixed control parameter. We have then applied our theoretical findings to shed light on the role of the different model parameters on the generalization performance. We have considered the setup with a generic bias b, two clusters with generic sizes, showing that the case $b = 0$, $\rho = 1/2$ is singular and the generic case, $\rho \neq 1/2$, has qualitatively different behavior when regularization is added. Finally, we have obtained that the linear separability transition explicitly depends on the cluster size.

In this work we focused on ridge regularization, however the analysis framework can be generalized to handle other separable convex regularizers. Moreover, our analysis can be extended to the following multi-class Gaussian mixture model:

$$\mathbf{x}_i = \mathbf{v}_{y_i} + \sqrt{\Delta}\mathbf{z}_i,$$

where $\{y_i\}$ are i.i.d. samples drawn from a finite set $\{1, 2, \ldots, M\}$, and $\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_M\}$ is a collection of M fixed vectors. The current model studied in the paper, eq. (1), is a special case of this more general setting.

Acknowledgements

We thank 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.

References

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

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

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

Cover, T. M. Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition. *IEEE transactions on electronic computers*, (3):326–334, 1965.

Del Giudice, P., Franz, S., and Virasoro, M. Perceptron beyond the limit of capacity. *Journal de Physique*, 50(2): 121–134, 1989.

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

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

Donoho, D. L., Maleki, A., and Montanari, A. Message-passing algorithms for compressed sensing. *Proceedings of the National Academy of Sciences*, 106(45):18914–18919, 2009.

Franz, S., Amit, D. J., and Virasoro, M. A. Prosopagnosia in high capacity neural networks storing uncorrelated classes. *Journal de Physique*, 51(5):387–408, 1990.

Friedman, J., Hastie, T., and Tibshirani, R. *The elements of statistical learning*, volume 1. Springer series in statistics New York, 2001.

Gardner, E. The space of interactions in neural network models. *Journal of physics A: Mathematical and general*, 21(1):257, 1988.

Gardner, E. and Derrida, B. Three unfinished works on the optimal storage capacity of networks. *Journal of Physics A: Mathematical and General*, 22(12):1983, 1989.

Geiger, M., Spigler, S., d'Ascoli, S., Sagun, L., Baity-Jesi, M., Biroli, G., and Wyart, M. Jamming transition as a paradigm to understand the loss landscape of deep neural networks. *Physical Review E*, 100(1):012115, 2019.

Gordon, Y. Some inequalities for Gaussian processes and applications. *Israel Journal of Mathematics*, 50(4):265–289, Dec 1985.

Gordon, Y. On milman's inequality and random subspaces which escape through a mesh in rn. In Lindenstrauss, J. and Milman, V. D. (eds.), *Geometric Aspects of Functional Analysis*, number 1317 in Lecture Notes in Mathematics, pp. 84–106. Springer Berlin Heidelberg, January 1988. ISBN 978-3-540-19353-1, 978-3-540-39235-4. URL http://link.springer.com/chapter/10.1007/BFb0081737.

Hastie, T., Montanari, A., Rosset, S., and Tibshirani, R. J. Surprises in high-dimensional ridgeless least squares interpolation. *arXiv preprint arXiv:1903.08560*, 2019.

Hebb, D. O. *The organization of behavior: A neuropsychological theory*. Psychology Press, 2005.

Kini, G. and Thrampoulidis, C. Analytic study of double descent in binary classification: The impact of loss. *arXiv preprint arXiv:2001.11572*, 2020.

Krauth, W. and Mézard, M. Storage capacity of memory networks with binary couplings. *Journal de Physique*, 50 (20):3057–3066, 1989.

Lelarge, M. and Miolane, L. Asymptotic bayes risk for gaussian mixture in a semi-supervised setting. *arXiv preprint arXiv:1907.03792*, 2019.

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

Maeangi, C., Biehl, M., and Solla, S. Supervised learning from clustered input examples. 30:117–122, 04 1995. doi: 10.1209/0295-5075/30/2/010.

Mai, X. and Liao, Z. High dimensional classification via empirical risk minimization: Improvements and optimality. *arXiv preprint arXiv:1905.13742*, 2019.

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

Mézard, M. and Montanari, A. *Information, physics, and computation*. Oxford University Press, 2009.

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

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

Montanari, A., Ruan, F., Sohn, Y., and Yan, J. The generalization error of max-margin linear classifiers: High-dimensional asymptotics in the overparametrized regime. *arXiv preprint arXiv:1911.01544*, 2019.

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

Opper, M. and Kinzel, W. Statistical mechanics of generalization. In *Models of neural networks III*, pp. 151–209. Springer, 1996.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. Scikit-learn: Machine learning in python. *Journal of machine learning research*, 12(Oct):2825–2830, 2011.

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

Salehi, F., Abbasi, E., and Hassibi, B. The impact of regularization on high-dimensional logistic regression. In *Advances in Neural Information Processing Systems*, pp. 11982–11992, 2019.

Soudry, D., Hoffer, E., Nacson, M. S., Gunasekar, S., and Srebro, N. The implicit bias of gradient descent on separable data. *The Journal of Machine Learning Research*, 19(1):2822–2878, 2018.

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

Thrampoulidis, C., Oymak, S., and Hassibi, B. Regularized linear regression: A precise analysis of the estimation error. In *Proceedings of The 28th Conference on Learning Theory*, volume 40, pp. 1683–1709, Paris, France, 03–06 Jul 2015. PMLR.