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

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

A. Derivation of the generalization error formula

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

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

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

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

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 $y_{\text{new}}, \hat{y}_{\text{new}} = \pm 1$, it is easy to show that (A.1) can be rewritten as

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

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

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

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

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

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

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

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

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

[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.

By averaging over ς, we obtain

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

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

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

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

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

In the large d limit, the overlaps concentrate:

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

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

Hence the generalization error reads

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

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

B. Derivation of the Bayes-optimal error

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

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

where "$\propto$" takes into account the normalization. Similarly, the posterior on $\mathbf{v}$ given the training set 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

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

in order to estimate the new label as

$$\hat{y}_{\text{new}}=\underset{y'=\pm1}{\text{argmax}}\log\mathrm{p}\left(y'|\mathbf{x}_{\text{new}},\mathbf{X},\mathbf{y}\right).\tag{B.4}$$

Therefore, we have to compute

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

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

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

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

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

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

$$\hat{y}_{\text{new}} = \underset{y=\pm 1}{\text{argmax}}\left[\frac{\alpha}{\Delta\left(\alpha+\Delta+\frac{1}{d}\right)} y\mathbf{x}_{\text{new}} \cdot \frac{1}{n}\sum_{\mu=1}^{n} y_\mu \mathbf{x}_\mu + \log \text{p}_y\left(y\right)\right]. \tag{B.8}$$

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

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

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

$$\hat{y}_{\text{new}} = \underset{y=\pm 1}{\text{argmax}}\left[\frac{\alpha}{\Delta\left(\alpha+\Delta\right)} y\left(y_{\text{new}} + \sqrt{\Delta\left(1+\frac{\Delta}{\alpha}\right)} z'_{\text{new}}\right) + \log \text{p}_y\left(y\right)\right]. \tag{B.10}$$

It is useful to rewrite the generalization error as

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

Using (B.10), we can compute

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

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

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

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

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

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

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

which is indeed the formula given in Eq. 19 of the main text.

B.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 (Lelarge & Miolane, 2019) 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

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

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.

C. Derivation of the training loss formula

In what follows, we provide more technical details for several key results stated in Section 3. They serve as the basis of the proof of Proposition 1.

C.1. Proof of Proposition 3

Recall from the main text that

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

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

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

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 (32) in the main text.

Gordon's minimax inequalities (Gordon, 1985; 1988; Thrampoulidis et al., 2015) allow us to make the following comparison: For any constants c and $\delta > 0$, we have

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

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

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

It follows that

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

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 (Thrampoulidis et al., 2015), we can interchange the order of $\min$ and $\max$ in the dual formulation of (22), which then allows us to reach the result in the main text. Indeed, in applying Gordon's inequalities to analyze high-dimensional random optimization problems, one can exchange the order of the min and the max and thus obtain a two sided inequality by using the following arguments. Let

$$\Phi = \min_{\mathbf{w} \in \mathcal{S}_w} \max_{\mathbf{u} \in \mathcal{S}_u} \left\{ \mathbf{w}^\top \mathbf{Z} \mathbf{u} + f(\mathbf{w}) - g(\mathbf{u}) \right\},$$

where $\mathbf{Z}$ is a random matrix with i.i.d. standard Gaussian entries, and $f(\cdot), g(\cdot)$ are two convex functions. (Note that the quantity $\mathcal{L}_\lambda(q, m, b)$ in our proof is a special case of the above formulation). In the proof of Proposition 3, we applied Gordon's inequality to show that Φ has a high-probability lower bound given by $\widetilde{L}_\lambda(q, m, b)$. Now to exchange the order of the min and the max, we consider

$$\begin{aligned}
-\Phi &= \max_{\mathbf{w} \in \mathcal{S}_w} \min_{\mathbf{u} \in \mathcal{S}_u} \left\{ \mathbf{w}^\top (-\mathbf{Z}) \mathbf{u} - f(\mathbf{w}) + g(\mathbf{u}) \right\} \\
&= \min_{\mathbf{u} \in \mathcal{S}_u} \max_{\mathbf{w} \in \mathcal{S}_w} \left\{ \mathbf{w}^\top (-\mathbf{Z}) \mathbf{u} - f(\mathbf{w}) + g(\mathbf{u}) \right\},
\end{aligned} \tag{C.2}$$

where the second equality holds *if* $\mathcal{S}_w$ and $\mathcal{S}_u$ are convex sets. Note that the constraint set $\mathbf{w} \in \mathcal{S}_{q,m}$ as used in the proof of Proposition 3 is not convex, and that's exactly the reason why we only state an inequality for $\mathcal{L}_\lambda(q, m, b)$ and $\varepsilon_\lambda^{(d)}$ in Proposition 3. However, when we study $\mathcal{L}_\lambda^*$, there is no longer any restriction on $\mathbf{w}$ given by q and m. It follows that the corresponding constraint set $\mathcal{S}_w$ is convex. Consequently, we can now apply Gordon's inequalities to (C.2) to get a high-probability lower bound for $-\Phi$ and thus a high-probability upper bound for Φ.

C.2. Proof of Lemma 1

We first rewrite the optimization problem in (31) as

$$\max_{\mu \geq 0} \max_{\|\boldsymbol{u}\|^2/d = \mu} \left\{ -\sqrt{\Delta \mu (q - m^2)} + \frac{\boldsymbol{u}^\top \boldsymbol{h}}{d} - \frac{1}{d} \sum_{i=1}^n \widetilde{\ell}(u_i) \right\}. \tag{C.3}$$

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.3) in terms of

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

and thus we get (33).

C.3. Proof of Proposition 1

We first establish (25) for the special case where the subset Ω is a singleton. In this case, we just need to show

$$\mathbb{P}\Big(\mathcal{L}_\lambda(q,m,b) \geq \mathcal{E}_\lambda(q,m,b) - \delta\Big) \to 1. \tag{C.4}$$

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 (33) 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

$$\alpha(\gamma_d^*)^2\frac{\|\boldsymbol{u}_{\gamma_d^*}\|^2}{n} = \Delta(q-m^2). \tag{C.5}$$

Moreover,

$$\mathcal{E}_\lambda^{(d)}(q,m,b) = \frac{\sum_{i=1}^n \big[u_{\gamma_d^*,i}\widetilde{\ell}'(u_{\gamma_d^*,i}) - \widetilde{\ell}(u_{\gamma_d^*,i})\big]}{d} + \frac{\lambda q}{2}. \tag{C.6}$$

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

$$v_{\gamma,i} + \gamma\ell'(v_{\gamma,i}) = h_i. \tag{C.7}$$

Moreover, (C.5) can now be rewritten as

$$\alpha(\gamma_d^*)^2\frac{1}{n}\sum_{i=1}^n[\ell'(v_{\gamma_d^*,i})]^2 = \Delta_d(q-m^2) \tag{C.8}$$

and more importantly, (C.6) 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

$$v_\gamma + \gamma\ell'(v_\gamma) = h, \tag{C.9}$$

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 the main text. As $n, d \to \infty$ with d/n fixed at α, we have

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

uniformly over any compact subset of γ. It follows that γ_d^* as defined in (C.8) converges to γ^*, which is the unique solution of (24). Moreover, we have

$$\mathcal{E}_\lambda^{(d)}(q,m,b) \to \mathcal{E}_\lambda(q,m,b) = \alpha\mathbb{E}[\ell(v_{\gamma^*})] + \frac{\lambda q}{2}. \tag{C.10}$$

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.10), we have (C.4).

Let Ω be an arbitrary compact subset of $\{(q,m,b) : m^2 \le q\}$. 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 (25). This strategy follows closely the approach used in (Thrampoulidis et al., 2015). Finally, to get (26), we first note that (25) 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 (Thrampoulidis et al., 2015).

C.4. Proof of Proposition 2

We start with the fixed-point equation for the Lagrange multiplier given in (24). For our proof, it will be more convenient to rewrite this equation in terms of the random variable $u_\gamma \overset{\text{def}}{=} \ell'(v_\gamma)$. 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.9), 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 (24) in terms of u_γ, we have

$$\alpha\gamma^2\mathbb{E}[u_\gamma^2] = \Delta(q - m^2). \tag{C.11}$$

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.11) as

$$\alpha S(\widetilde{\gamma}, q, \theta) = \Delta(1 - \theta^2), \tag{C.12}$$

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.12). 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.*,

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

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

$$\alpha < \frac{\Delta(1 - \theta^2)}{S^*(\infty, \theta)}. \tag{C.13}$$

It follows from (C.9) 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.13), 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.13), which gives us the final expression as stated in Proposition 2.

C.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 (24) and (26) in the main text. First of all, we observe that (24) 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.10), 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.7) and (24) respectively. The minimization leads to the following system of equations:

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

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

$$\mathbb{E}_{y,s}\left[y\ell'(v_{\gamma^*})\right] = 0, \tag{C.16}$$

where $s \sim \mathcal{N}(0,1)$, $y = +1$ with probability $\rho \in (0,1)$ and $y = -1$ otherwise. We observe that (C.16) is the same as (11) and (C.15) is equivalent to (4) and (6). Using again (6), we can rewrite (C.14) as

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

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

$$\hat{\gamma} = \alpha \sqrt{\frac{\Delta}{q}} \mathbb{E}_{y,s} \left[\partial_s v_{\gamma^*} \ell''(v_{\gamma^*}) \right], \tag{C.18}$$

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

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.

D.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

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

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

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

D.2. Hinge loss

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

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

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

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

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

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

The equation to determine the bias is

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

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

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

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

$$\ell(y\mathbf{w}^\top\mathbf{x}) = \ell(0) + y\mathbf{w}^\top\mathbf{x}\,\ell'(0) + \frac{1}{2}\left(\mathbf{w}^\top\mathbf{x}\right)^2\ell''(0) + o(q)$$

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. 1, 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.

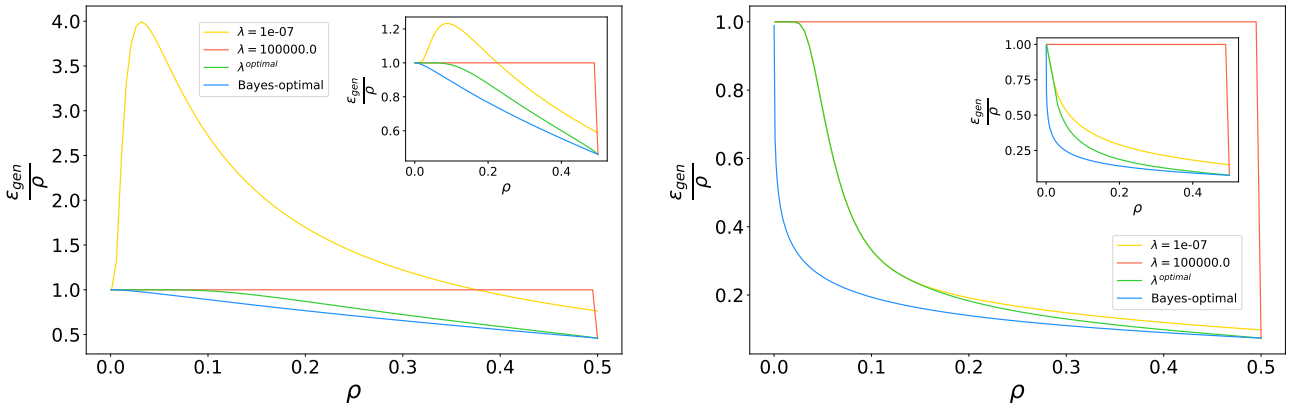

Figure 1. Generalization error as a function of ρ, at fixed $\alpha = 1.2$ and $\Delta = 1$ (left) and $\alpha = 7$ and $\Delta = 0.3$ (right), for the square loss compared to the Bayes-optimal performance. In the inset, the same figure for the hinge loss. The vertical axis is rescaled by ρ for convenience. 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 $\rho = 1/2$.

F. Details on the numerics

F.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 losses, the equations must 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 $eps \leq 10^{-8}$.

F.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 losses, the solution can be computed by a standard gradient descent algorithm. In Fig. 1 we used the Logistic Regression classifier provided by the scikitlearn package $linear\_model$ (Pedregosa et al., 2011). 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 behavior 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.

References

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.

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

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.

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.