A Replica derivation of the fixed-point equations

In this Appendix, we will derive the fixed point equations for the order parameters following the analysis by Loureiro et al. (2021); Pesce et al. (2023); Adomaityte et al. (2023) in the most general setting discussed in Result 4.3: the case in Result 2.4 is obtained by fixing $K = 1$ and $\boldsymbol{\mu}_1 \equiv \mathbf{0}$ below, and by assuming a ridge regularisation. The dataset $\mathcal{D} \coloneqq \{(\boldsymbol{x}_i, y_i)\}_{i\in[n]}$ consists of n independent datapoints $\boldsymbol{x}_i \in \mathbb{R}^d$ each associated to a label $y_i \in \mathcal{Y}$. The elements of the dataset are independently generated by using a law $P(\boldsymbol{x}, y)$ which we assume can be put in the form of a superstatistical mixture model (SMM) involving K clusters $\mathcal{C} = \{1, \dots, K\}$,

$$P(\boldsymbol{x}, y) \equiv P_0(y|\boldsymbol{\beta}_\star^\intercal \boldsymbol{x}) \sum_{c\in\mathcal{C}} p_c \mathbb{E}_{\sigma_c}\left[\mathcal{N}\left(\boldsymbol{x}; \boldsymbol{\mu}_c, \sigma_c^2/d \boldsymbol{I}_d\right)\right], \tag{23}$$

and $P_0(\bullet|\tau)$ is the distribution of the scalar label y produced via the "teacher" $\boldsymbol{\beta}_\star$. In the following, we assume that $\beta_\star^2 = {}^1/_d \|\boldsymbol{\beta}_\star\|_2^2 = \Theta(1)$. In the equation above, $\forall c \in \mathcal{C}$, $p_c \in [0, 1]$ and $\boldsymbol{\mu}_c \in \mathbb{R}^d$ with $\|\boldsymbol{\mu}_c\|_2^2 = \Theta({}^1/_d)$. It is assumed that $\sum_c p_c = 1$. The expectation is intended over σ_c, a positive random variable with density ϱ_c. We will perform our regression task searching for a set of *weights* $\hat{\boldsymbol{\beta}}_\lambda$, that will allow us to construct an estimator via a certain classifier $f \colon \mathbb{R} \to \mathcal{Y}$:

$$\boldsymbol{x} \mapsto f(\hat{\boldsymbol{\beta}}_\lambda^\intercal \boldsymbol{x}) = y, \tag{24}$$

which will provide us with our prediction for a datapoint $\boldsymbol{x}$. The weights will be chosen by minimising an empirical risk function in the form

$$\mathcal{R}(\boldsymbol{\beta}) \equiv \sum_{\nu=1}^n \ell\left(y_i, \boldsymbol{\beta}^\intercal \boldsymbol{x}_i\right) + \lambda r(\boldsymbol{\beta}), \tag{25}$$

i.e., they are given by

$$\hat{\boldsymbol{\beta}}_\lambda \coloneqq \arg \min_{\boldsymbol{\beta}\in\mathbb{R}^d} \mathcal{R}(\boldsymbol{\beta}). \tag{26}$$

We will assume that ℓ is a convex loss function with respect to its second argument, and r is a strictly convex regularisation function: the parameter $\lambda \geq 0$ will tune the strength of the regularisation. Note that this setting is slightly more general than the one given in the main text. The starting point is to reformulate the problem as an optimisation problem by introducing a Gibbs measure over the parameters $\boldsymbol{\beta}$ depending on a positive parameter $\upbeta$,

$$\mu_\upbeta(\boldsymbol{\beta}) \propto \mathrm{e}^{-\upbeta\mathcal{R}(\boldsymbol{\beta})} = \underbrace{\mathrm{e}^{-\beta r(\boldsymbol{\beta})}}_{P_w} \prod_{i=1}^n \underbrace{\exp\left[-\upbeta\ell\left(y_i, \boldsymbol{\beta}^\intercal \boldsymbol{x}_i\right)\right]}_{P_\ell}, \tag{27}$$

so that, in the $\upbeta \to +\infty$ limit, the Gibbs measure concentrates on $\hat{\boldsymbol{\beta}}_\lambda$. The functions P_y and P_w can be interpreted as (unnormalised) likelihood and prior distribution respectively. Our analysis will go through the computation of the average free energy density associated with such Gibbs measure in a specific proportional limit, i.e.,

$$f_\upbeta \coloneqq - \lim_{\substack{n,d\to+\infty\\ n/d=\alpha}} \mathbb{E}_\mathcal{D}\left[\frac{\ln \mathcal{Z}_\upbeta}{d\upbeta}\right] = \lim_{\substack{n,d\to+\infty\\ n/d=\alpha}} \lim_{s\to 0} \frac{1 - \mathbb{E}_\mathcal{D}[\mathcal{Z}_\upbeta^s]}{sd\upbeta}, \tag{28}$$

where $\mathbb{E}_\mathcal{D}[\bullet]$ is the average over the training dataset, and we have introduced the partition function

$$\mathcal{Z}_\upbeta \coloneqq \int \mathrm{e}^{-\upbeta\mathcal{R}(\boldsymbol{\beta})}\, \mathrm{d}\boldsymbol{\beta}. \tag{29}$$

A.1 Replica approach.

In our replica approach, we need to evaluate

$$\mathbb{E}_\mathcal{D}[\mathcal{Z}_\upbeta^s] = \prod_{a=1}^s \int \mathrm{d}\boldsymbol{\beta}^a P_w(\boldsymbol{\beta}^a) \left(\mathbb{E}_{(\boldsymbol{x},y)}\left[\prod_{a=1}^s P_\ell(y|\boldsymbol{x}^\intercal \boldsymbol{\beta}^a)\right]\right)^n. \tag{30}$$

Let us take the inner average introducing a new set of *local fields* η^a and τ,

$$\mathbb{E}_{(\boldsymbol{x},y)}\left[\prod_{a=1}^{s} P_\ell\left(y\middle|\boldsymbol{x}^\intercal\boldsymbol{\beta}^a\right)\right]=\sum_c p_c\mathbb{E}_{\sigma_c}\left[\int_{\mathcal{Y}}\mathrm{d}y\int_{\mathbb{R}^d}\mathrm{d}\boldsymbol{x}P_0(y|\boldsymbol{x}^\intercal\boldsymbol{\beta}_\star)\mathcal{N}(\boldsymbol{x};\boldsymbol{\mu}_c,\sigma_c^2/d\boldsymbol{I}_d)\prod_{a=1}^{s}P_\ell(y|\boldsymbol{x}^\intercal\boldsymbol{\beta}^a)\right]$$

$$=\sum_c p_c\mathbb{E}_{\sigma_c}\left[\int\mathrm{d}\boldsymbol{\eta}\int\mathrm{d}\tau\int_{\mathcal{Y}}\mathrm{d}yP_0(y|\tau)\prod_{a=1}^{s}P_\ell(y|\eta^a)\mathcal{N}\left(\left(\begin{smallmatrix}\tau\\ \boldsymbol{\eta}\end{smallmatrix}\right);\left(\begin{smallmatrix}\boldsymbol{\mu}_c^\intercal\boldsymbol{\beta}_\star\\ \boldsymbol{\mu}_c^\intercal\boldsymbol{\beta}^a\end{smallmatrix}\right),\frac{\sigma_c^2}{d}\left(\begin{smallmatrix}d\beta_\star^2 & \boldsymbol{\beta}_\star^\intercal\boldsymbol{\beta}^b\\ \boldsymbol{\beta}_\star^\intercal\boldsymbol{\beta}^a & \boldsymbol{\beta}^{a\intercal}\boldsymbol{\beta}^b\end{smallmatrix}\right)\right)\right]. \quad (31)$$

We can write then

$$\mathbb{E}_{\mathcal{D}}[\mathcal{Z}_\upbeta^s]=\prod_{a=1}^{s}\int\mathrm{d}\boldsymbol{\beta}^aP_w(\boldsymbol{\beta}^a)\times$$

$$\left(\sum_c p_c\mathbb{E}_{\sigma_c}\left[\int\mathrm{d}\boldsymbol{\eta}\int\mathrm{d}\tau\int_{\mathcal{Y}}\mathrm{d}yP_0(y|\tau)\prod_{a=1}^{s}P_\ell(y|\eta^a)\mathcal{N}\left(\left(\begin{smallmatrix}\tau\\ \boldsymbol{\eta}\end{smallmatrix}\right);\left(\begin{smallmatrix}\boldsymbol{\mu}_c^\intercal\boldsymbol{\beta}_\star\\ \boldsymbol{\mu}_c^\intercal\boldsymbol{\beta}^a\end{smallmatrix}\right),\frac{\sigma_c^2}{d}\left(\begin{smallmatrix}d\beta_\star^2 & \boldsymbol{\beta}_\star^\intercal\boldsymbol{\beta}^b\\ \boldsymbol{\beta}_\star^\intercal\boldsymbol{\beta}^a & \boldsymbol{\beta}^{a\intercal}\boldsymbol{\beta}^b\end{smallmatrix}\right)\right)\right]\right)^n$$

$$=\prod_c\left(\prod_{a\le b}\iint\mathcal{D}\boldsymbol{Q}^{ab}\mathcal{D}\hat{\boldsymbol{Q}}^{ab}\right)\left(\prod_a\int\mathcal{D}\boldsymbol{M}^a\mathcal{D}\hat{\boldsymbol{M}}^a\right)\left(\prod_a\int\mathrm{d}t^a\,\mathrm{d}\hat{t}^a\right)\mathrm{e}^{-d\upbeta\Phi^{(s)}}. \quad (32)$$

In the equation above we introduced the *order parameters*

$$Q_c^{ab}=\frac{\sigma_c^2}{d}\boldsymbol{\beta}^{a\intercal}\boldsymbol{\beta}^b\in\mathbb{R},\quad a,b=1,\dots,s, \quad (33)$$

$$M_c^a=\frac{\sigma_c^2}{d}\boldsymbol{\beta}_\star^\intercal\boldsymbol{\beta}^a\in\mathbb{R},\quad a=1,\dots,s, \quad (34)$$

$$t_c^a=\boldsymbol{\mu}_c^\intercal\boldsymbol{\beta}^a\in\mathbb{R},\quad a=1,\dots,s, \quad (35)$$

whilst the integration is over all possible order parameters, Q_c^{ab} and m_c^a to be intended as functions of σ_c. In the equation, we have also denoted the replicated free-energy

$$\upbeta\Phi^{(s)}(\boldsymbol{Q},\boldsymbol{M},\hat{\boldsymbol{Q}},\hat{\boldsymbol{M}})=\sum_c\sum_a\mathbb{E}_{\sigma_c}[\hat{M}_c^aM_c^a]+\sum_c\sum_{a\le b}\mathbb{E}_{\sigma_c}[\hat{Q}_c^{ab}Q_c^{ab}]+\frac{1}{d}\sum_{c,a}\hat{t}_c^at_c^a$$

$$-\frac{1}{d}\ln\prod_{a=1}^{s}\int P_w(\boldsymbol{\beta}^a)\mathrm{d}\boldsymbol{\beta}^a\prod_c\exp\left(\sum_{a\le b}\mathbb{E}_{\sigma_c}[\sigma_c^2\hat{Q}_c^{ab}]\boldsymbol{\beta}^{a\intercal}\boldsymbol{\beta}^b+\sum_a\mathbb{E}_{\sigma_c}[\sigma_c^2\hat{M}_c^a]\boldsymbol{\beta}^{a\intercal}\boldsymbol{\beta}_\star+\sum_a\hat{t}_c^a\boldsymbol{\beta}^{a\intercal}\boldsymbol{\mu}_c\right)$$

$$-\alpha\ln\sum_c p_c\mathbb{E}_{\sigma_c}\left[\int\mathrm{d}\boldsymbol{\eta}\int\mathrm{d}\tau\int_{\mathcal{Y}}\mathrm{d}y\,P_0\left(y|\tau\right)\prod_{a=1}^{s}P_\ell\left(y|\eta^a\right)\mathcal{N}\left(\left(\begin{smallmatrix}\tau\\ \boldsymbol{\eta}\end{smallmatrix}\right);\left(\begin{smallmatrix}t_c^0\\ t_c^a\end{smallmatrix}\right)\left(\begin{smallmatrix}\sigma_c^2\beta_\star^2 & M_c^b\\ M_c^a & Q_c^{ab}\end{smallmatrix}\right)\right)\right], \quad (36)$$

where, for the sake of brevity, $t_c^0:=\boldsymbol{\mu}_c^\intercal\boldsymbol{\beta}_\star$. At this point, the free energy f_β should be computed functionally extremisizing with respect to all the order parameters by virtue of the Laplace approximation,

$$f_\upbeta=\lim_{s\to0}\operatorname*{Extr}_{\substack{\boldsymbol{M},\hat{\boldsymbol{M}},\boldsymbol{t}\\ \boldsymbol{Q},\hat{\boldsymbol{Q}},\hat{\boldsymbol{t}}}}\frac{\Phi^{(s)}(\boldsymbol{Q},\boldsymbol{M},\hat{\boldsymbol{Q}},\hat{\boldsymbol{M}},\boldsymbol{t},\hat{\boldsymbol{t}})}{s}. \quad (37)$$

Replica symmetric ansatz. Before taking the $s\to0$ limit we make the replica symmetric assumptions

$$\begin{aligned}Q_c^{aa}&=\begin{cases}R_c, & a=b\\ Q_c & a\neq b\end{cases} & \hat{Q}_c^{aa}&=\begin{cases}-\frac{1}{2}\hat{R}_c, & a=b\\ \hat{Q}_c & a\neq b\end{cases}\\ M_c^a&=M_c & \hat{M}_c^a&=\hat{M}_c\quad\forall a\\ t_c^a&=t_c & \hat{t}_c^a&=\hat{t}_c\quad\forall a\end{aligned} \quad (38)$$

If we denote $V_c:=R_c-Q_c$ we obtain, after some work we obtain

$$\ln\sum_c p_c\mathbb{E}_{\sigma_c}\left[\int \mathrm{d}\boldsymbol{\eta}\int \mathrm{d}\tau\int_{\mathcal{Y}}\mathrm{d}y\, P_0\left(y|\tau\right)\prod_{a=1}^{s}P_\ell\left(y|\eta^a\right)\,\mathcal{N}\left(\left(\begin{smallmatrix}\tau\\ \boldsymbol{\eta}\end{smallmatrix}\right);\left(\begin{smallmatrix}t_c^0\\ t_c\mathbf{1}_s\end{smallmatrix}\right),\left(\begin{smallmatrix}\sigma_c^2\beta_\star^2 & M_c\mathbf{1}_s^\intercal\\ M_c\mathbf{1}_s & Q_c\mathbf{1}_{s\times s}\end{smallmatrix}\right)\right)\right]$$
$$=s\sum_c p_c\mathbb{E}_{\sigma_c,\zeta}\left[\int_{\mathcal{Y}}\mathrm{d}yZ_0\left(y,t_c^0+\frac{M_c\zeta}{\sqrt{Q_c}},\sigma_c^2\beta_\star^2-\frac{M_c^2}{Q_c}\right)\ln Z_\ell\left(y,t_c+\sqrt{Q_c}\zeta,V_c\right)\right]+o(s),\quad(39)$$

with $\zeta\sim\mathcal{N}(0,1)$ is normally distributed and we have introduced the function

$$Z_\bullet(y,\mu,V)\coloneqq\int\frac{\mathrm{d}\tau P_\bullet(y|\tau)}{\sqrt{2\pi V}}\,\mathrm{e}^{-\frac{(\tau-\mu)^2}{2V}},\qquad \bullet\in\{0,\ell\}.\quad(40)$$

On the other hand, denoted by $\hat{V}_c=\hat{R}_c+\hat{Q}_c$, and introducing $\hat{q}_c\coloneqq\mathbb{E}_{\sigma_c}[\sigma_c^2\hat{Q}_c]$, $\hat{v}_c\coloneqq\mathbb{E}_{\sigma_c}[\sigma_c^2\hat{V}_c]$, and $\hat{m}_c\coloneqq\mathbb{E}_{\sigma_c}[\sigma_c^2\hat{M}_c]$

$$\frac{1}{d}\ln\prod_{a=1}^{s}\left(\int P_w(\boldsymbol{\beta}^a)\mathrm{d}\boldsymbol{\beta}^a\prod_c\mathrm{e}^{-\frac{\hat{v}_c}{2}\|\boldsymbol{\beta}^a\|_2^2+\boldsymbol{\beta}^{a\intercal}(\hat{m}_c\boldsymbol{\beta}_\star+\hat{t}_c\boldsymbol{\mu}_c)}\prod_{b,c}\mathrm{e}^{\frac{1}{2}\hat{q}_c\boldsymbol{\beta}^{a\intercal}\boldsymbol{\beta}^b}\right)=$$
$$=\frac{s}{d}\mathbb{E}_{\boldsymbol{\xi}}\ln\left[\int P_w(\boldsymbol{\beta})\mathrm{d}\boldsymbol{\beta}\prod_c\exp\left(-\frac{\hat{v}_c\|\boldsymbol{\beta}\|_2^2}{2}+\boldsymbol{\beta}^\intercal(\hat{m}_c\boldsymbol{\beta}_\star+\hat{t}_c\boldsymbol{\mu}_c)+\sqrt{\hat{q}_c}\boldsymbol{\xi}^\intercal\boldsymbol{\beta}\right)\right]+o(s).\quad(41)$$

In the expression above we have introduced $\boldsymbol{\xi}\sim\mathcal{N}(\mathbf{0},\boldsymbol{I}_d)$. Therefore, the (replicated) *replica symmetric* free-energy is given by

$$\lim_{s\to0}\frac{\beta}{s}\Phi_{\mathrm{RS}}^{(s)}=\frac{1}{d}\sum_c\hat{t}_ct_c+\sum_c\hat{M}_cM_c+\frac{\sum_c\mathbb{E}_{\sigma_c}\left[\hat{V}_cQ_c-\hat{Q}_cV_c-\hat{V}_cV_c\right]}{2}-\alpha\beta\Psi_\ell(M,Q,V)-\beta\Psi_w(\hat{m},\hat{q},\hat{v})\quad(42)$$

where we have defined two contributions

$$\Psi_\ell(M,Q,V)\coloneqq\frac{1}{\beta}\sum_c p_c\mathbb{E}_{\sigma_c,\zeta}\left[\int_{\mathcal{Y}}\mathrm{d}yZ_0\left(y,t_c^0+\frac{M_c\zeta}{\sqrt{Q_c}},\sigma_c^2\beta_\star^2-\frac{M_c^2}{Q_c}\right)\ln Z_\ell\left(y,t_c+\sqrt{Q_c}\zeta,V_c\right)\right],$$
$$\Psi_w(\hat{m},\hat{q},\hat{v})\coloneqq\frac{1}{\beta d}\mathbb{E}_{\boldsymbol{\xi}}\ln\left[\int P_w(\boldsymbol{\beta})\mathrm{d}\boldsymbol{\beta}\prod_c\exp\left(-\frac{\hat{v}_c\|\boldsymbol{\beta}\|_2^2}{2}+\boldsymbol{\beta}^\intercal(\hat{m}_c\boldsymbol{\beta}_\star+\hat{t}_c\boldsymbol{\mu}_c)+\sqrt{\hat{q}_c}\boldsymbol{\xi}^\intercal\boldsymbol{\beta}\right)\right].\quad(43)$$

Note that we have separated the contribution coming from the chosen loss (the so-called *channel* part Ψ_ℓ) from the contribution depending on the regularisation (the *prior* part Ψ_w). To write down the saddle-point equations in the $\beta\to+\infty$ limit, let us first rescale our order parameters as $\hat{M}_c\mapsto\beta\hat{M}_c$, $\hat{t}_c\mapsto d\beta\hat{t}_c$, $\hat{Q}_c\mapsto\beta^2\hat{Q}_c$, $\hat{V}_c\mapsto\beta\hat{V}_c$ and $V_c\mapsto\beta^{-1}V_c$. Also, for future convenience, let us rescale $Q_c\mapsto\sigma_c^2q_c$, $M_c\mapsto\sigma_c^2m_c$, $V_c\mapsto\sigma_c^2v_c$. For $\beta\to+\infty$ the channel part is

$$\Psi_\ell(m,q,v,t)=$$
$$=-\sum_c p_c\mathbb{E}_{\sigma_c,\zeta}\left[\int_{\mathcal{Y}}\mathrm{d}yZ_0\left(y,t_c^0+\frac{\sigma_cm\zeta}{\sqrt{q_c}},\sigma_c^2\beta_\star^2-\frac{\sigma_c^2m_c^2}{q_c}\right)\left(\frac{(h_c-t_c-\sigma_c\sqrt{q_c}\zeta)^2}{2\sigma_c^2v_c}+\ell(y,h_c)\right)\right].\quad(44)$$

where we have written Ψ_ℓ of a Moreau envelope, i.e., in terms of a proximal

$$h_c\coloneqq\arg\min_u\left[\frac{(u-\omega_c)^2}{2\sigma_c^2v_c}+\ell(y,u)\right]\qquad\text{with }\omega_c=t_c+\sigma_c\sqrt{q_c}\zeta.\quad(45)$$

A similar expression can be obtained for Ψ_w. Introducing the proximal

$$\boldsymbol{g}=\arg\min_{\boldsymbol{\beta}}\left(\frac{\|\boldsymbol{\beta}\|_2^2\sum_c\hat{v}_c}{2}-\boldsymbol{\beta}^\intercal\sum_c\left(\hat{m}_c\boldsymbol{\beta}_\star+d\hat{t}_c\boldsymbol{\mu}_c\right)-\boldsymbol{\xi}^\intercal\boldsymbol{\beta}\sum_c\sqrt{\hat{q}_c}+\lambda r(\boldsymbol{\beta})\right)\in\mathbb{R}^d\quad(46)$$

We can rewrite the prior contribution Ψ_w as

$$\Psi_w(\hat{m},\hat{q},\hat{v},\hat{t})=-\frac{1}{d}\mathbb{E}_{\boldsymbol{\xi}}\left[\frac{\|\boldsymbol{g}\|_2^2}{2}\sum_c\hat{v}_c-\boldsymbol{g}^\intercal\sum_c\left(\hat{m}_c\boldsymbol{\beta}_\star+\hat{t}_c\boldsymbol{\mu}_c\right)-\boldsymbol{\xi}^\intercal\boldsymbol{g}\sum_c\sqrt{\hat{q}_c}+\lambda r(\boldsymbol{g})\right]\quad(47)$$

The parallelism between the two contributions is evident, aside from the different dimensionality of the involved objects. The replica symmetric free energy in the $\beta \to +\infty$ limit is computed by extremising with respect to the introduced order parameters,

$$\begin{aligned} f_{\mathrm{RS}}{=}\mathrm{Extr}\Bigg[&\sum_c \mathbb{E}_{\sigma_c}[\sigma_c^2 \hat{M}_c m_c]{+}\frac{1}{2}\sum_c \mathbb{E}_{\sigma_c}\left[\sigma_c^2\Big(\hat{V}_c q_c{-}\hat{Q}_c v_c\Big)\right]{+}\sum_c t_c \hat{t}_c \\ &-\alpha\Psi_\ell(m,q,v,t) - \Psi_w(\hat{m},\hat{q},\hat{v},\hat{t})\Bigg]. \end{aligned} \tag{48}$$

To do so, we have to write down a set of saddle-point equations and solve them.

Saddle-point equations. The saddle-point equations are derived straightforwardly from the obtained free energy functionally extremising with respect to all parameters. It is easily seen that v_c, q_c and m_c are independent from σ_c, and that it is possible to reduce the number of variables by introducing $\hat{v} = \sum_c \hat{v}_c$, so that we can write

$$v_c = \frac{\mathbb{E}_{\boldsymbol{\xi}}[\boldsymbol{g}^\intercal \boldsymbol{\xi}]}{d\sqrt{\hat{q}_c}}, \tag{49a}$$

$$q = \frac{\mathbb{E}_{\boldsymbol{\xi}}[\|\boldsymbol{g}\|_2^2]}{d}, \tag{49b}$$

$$m = \frac{\mathbb{E}_{\boldsymbol{\xi}}\left[\boldsymbol{g}^\intercal \boldsymbol{\beta}_\star\right]}{d}, \tag{49c}$$

$$t_c = \mathbb{E}_{\boldsymbol{\xi}}\left[\boldsymbol{g}^\intercal \boldsymbol{\mu}_c\right]. \tag{49d}$$

and the remaining equations can be rewritten as

$$\hat{q}_c{=}\alpha p_c\int_{\mathcal{Y}}\mathrm{d}y\mathbb{E}_{\sigma_c,\zeta}\left[\sigma_c^2 Z_0\left(y{,}t_c^0{+}\frac{\sigma_c m}{\sqrt{q}}\zeta{,}\sigma_c^2\beta_\star^2{-}\frac{\sigma_c^2 m^2}{q}\right)f_c^2\right], \tag{50a}$$

$$\hat{v}{=}{-}\alpha\sum_c p_c\int_{\mathcal{Y}}\mathrm{d}y\mathbb{E}_{\sigma_c,\zeta}\left[\sigma_c^2 Z_0\left(y{,}t_c^0{+}\frac{\sigma_c m}{\sqrt{q}}\zeta{,}\sigma_c^2\beta_\star^2{-}\frac{\sigma_c^2 m^2}{q}\right)\partial_\omega f_c\right], \tag{50b}$$

$$\hat{m}_c{=}\alpha p_c\int_{\mathcal{Y}}\mathrm{d}y\mathbb{E}_{\sigma_c,\zeta}\left[\sigma_c^2 \partial_\mu Z_0\left(y{,}t_c^0{+}\frac{\sigma_c m}{\sqrt{q}}\zeta{,}\sigma_c^2\beta_\star^2{-}\frac{\sigma_c^2 m^2}{q}\right)f_c\right], \tag{50c}$$

$$\hat{t}_c{=}\alpha p_c\int_{\mathcal{Y}}\mathrm{d}y\mathbb{E}_{\sigma_c,\zeta}\left[Z_0\left(y{,}t_c^0{+}\frac{\sigma_c m}{\sqrt{q}}\zeta{,}\sigma_c^2\beta_\star^2{-}\frac{\sigma_c^2 m^2}{q}\right)f_c\right] \tag{50d}$$

with

$$\begin{aligned} f_c &\coloneqq \arg\min_u \left[\frac{\sigma_c^2 v_c u^2}{2} + \ell(y,\omega_c + \sigma_c^2 v_c u)\right], \qquad \omega_c = t_c + \sigma_c\sqrt{q}\zeta, \\ \boldsymbol{g} &= \arg\min_{\boldsymbol{\beta}}\left(\frac{\|\boldsymbol{\beta}\|_2^2 \hat{v}}{2} - \boldsymbol{\beta}^\intercal \sum_c \left(\hat{m}_c \boldsymbol{\beta}_\star + d\hat{t}_c \boldsymbol{\mu}_c\right) - \boldsymbol{\xi}^\intercal \boldsymbol{\beta}\sum_c \sqrt{\hat{q}_c} + \lambda r(\boldsymbol{\beta})\right). \end{aligned} \tag{51}$$

To obtain the replica symmetric free energy, therefore, the given set of equations has to be solved, and the result is then plugged in Eq. 48. The obtained saddle-point equations correspond to the ones given in the Result 4.3.

Training and test errors. Let us show how to use the previous result to estimate the training loss and the generalisation error. Let us start from estimating

$$\varepsilon_\ell \coloneqq \lim_{n\to+\infty}\frac{1}{n}\sum_{i=1}^n \ell(y_i, \hat{\boldsymbol{\beta}}_\lambda^\intercal \boldsymbol{x}_i). \tag{52}$$

The best way to proceed is to observe that

$$\varepsilon_\ell{=}{-}\lim_{\beta\to+\infty}\partial_\beta\left(\beta\Psi_\ell\right){=}\sum_c p_c\int_{\mathcal{Y}}\mathrm{d}y\mathbb{E}_{\sigma_c,\zeta}\left[Z_0\left(y{,}t_c^0{+}\frac{\sigma_c m}{\sqrt{q}}\zeta{,}\sigma_c^2\beta_\star^2{-}\frac{\sigma_c^2 m^2}{q}\right)\ell(y{,}h_c)\right]. \tag{53}$$

This concentration result holds for a generic function $\varphi\colon \mathcal{Y}\times\mathbb{R}\to\mathbb{R}$, so that more generally, under Assumption 2.3,

$$\frac{1}{n}\sum_{i=1}^{n}\varphi(y_i,\hat{\boldsymbol{\beta}}_\lambda^\intercal\boldsymbol{x}_i)\xrightarrow[n,d\to+\infty]{\mathrm{P}}\sum_c p_c\int_{\mathcal{Y}}\mathrm{d}y\mathbb{E}_{\sigma_c,\zeta}\left[Z_0\left(y,t_c^0+\frac{\sigma_c m}{\sqrt{q}}\zeta,\sigma_c^2\beta_\star^2-\frac{\sigma_c^2 m^2}{q}\right)\varphi(y,h_c)\right]. \tag{54}$$

The expressions above hold in general, but, as anticipated, important simplifications can occur in the set of saddle-point equations Eq. 50 and Eq. 49 depending on the choice of the loss ℓ and of the regularization function r. The population's expectation (e.g., in the computation of the test error) of a performance function $\varphi\colon \mathcal{Y}\times\mathbb{R}\to\mathbb{R}$ is given instead by

$$\varepsilon_g \coloneqq \lim_{n\to+\infty}\mathbb{E}_{(y,\boldsymbol{x})}\left[\varphi(y,\hat{\boldsymbol{\beta}}_\lambda^\intercal\boldsymbol{x})\right], \tag{55}$$

where the expectation has to be taken on a newly sampled datapoint $(y,\boldsymbol{x})\notin\mathcal{D}$. This expression can be rewritten as

$$\begin{aligned}&\mathbb{E}_{y|\boldsymbol{x}}\left[\int \mathrm{d}\eta\varphi(y,\eta)\mathbb{E}_{\boldsymbol{x}}\left[\delta\left(\eta-\hat{\boldsymbol{\beta}}_\lambda^\intercal\boldsymbol{x}\right)\right]\right]\\ &\quad\xrightarrow[n,d\to+\infty]{\mathrm{P}}\int\mathrm{d}\eta\int\mathrm{d}\tau\int_{\mathcal{Y}}\mathrm{d}y\,P_0\left(y|\tau\right)\varphi(y,\eta)\sum_c p_c\mathbb{E}_{\sigma_c}\left[\mathcal{N}\left(\left(\begin{smallmatrix}\tau\\ \eta\end{smallmatrix}\right);\left(\begin{smallmatrix}t_c^0\\ t_c\end{smallmatrix}\right),\sigma_c^2\left(\begin{smallmatrix}\beta_\star^2 & m\\ m & q\end{smallmatrix}\right)\right)\right].\end{aligned} \tag{56}$$

This can be easily computed numerically once the order parameters are given. Finally, another relevant quantity for our investigations is the estimation mean-square error

$$\varepsilon_{\text{est}} \coloneqq \lim_{d\to+\infty}\frac{1}{d}\mathbb{E}_{\mathcal{D}}\left[\|\hat{\boldsymbol{\beta}}_\lambda-\boldsymbol{\beta}_\star\|_2^2\right]=\beta_\star^2-2m+q. \tag{57}$$

A.2 Bayes-optimal estimator for $K=1$

A derivation similar to the one discussed above can be repeated to obtain information on the statistical properties of the Bayes optimal estimator presented in Result 2.6. Given a dataset $\mathcal{D}$ of observation, we have that

$$P(\boldsymbol{\beta}|\mathcal{D})=\frac{P(\boldsymbol{\beta})P(\mathcal{D}|\boldsymbol{\beta})}{\mathcal{Z}(\mathcal{D})}=\frac{P(\boldsymbol{\beta})}{\mathcal{Z}(\mathcal{D})}\prod_{i=1}^{n}P_0(y_i|\boldsymbol{\beta}^\intercal\boldsymbol{x}_i) \tag{58}$$

where $P(\boldsymbol{\beta})$ is the prior on the teacher that we assume to be Gaussian, $P(\boldsymbol{\beta})=\mathcal{N}(\boldsymbol{\beta};\boldsymbol{0},\beta_\star^2\boldsymbol{I}_d)$, and

$$\mathcal{Z}(\mathcal{D})\coloneqq\int\mathrm{d}\boldsymbol{\beta}P(\boldsymbol{\beta})\prod_{i=1}^{n}P_0(y_i|\boldsymbol{\beta}^\intercal\boldsymbol{x}_i)=\frac{1}{(2\pi)^{n/2}}\int\mathrm{d}\boldsymbol{\beta}\exp\left(-\frac{\|\boldsymbol{\beta}\|_2^2}{2\beta_\star^2}+\sum_{i=1}^{n}\ln P_0(y_i|\boldsymbol{\beta}^\intercal\boldsymbol{x}_i)\right). \tag{59}$$

The calculation of $\mathcal{Z}(\mathcal{D})$ gives access in particular to the free entropy $\phi(\mathcal{D})\coloneqq\lim_n\frac{1}{n}\ln\mathcal{Z}(\mathcal{D})$. The computation of ϕ, which has an information-theoretical interpretation as mutual information, provides us the statistics of $\boldsymbol{\beta}$ according to the true posterior $P(\boldsymbol{\beta}|\mathcal{D})$. By assuming a concentration in the large n limit, the calculation is performed on $\mathbb{E}_{\mathcal{D}}[\ln\mathcal{Z}(\mathcal{D})]$. Using the replica trick as before, the calculation can be repeated almost identically. For the sake of simplicity, we focus on the case in which only one cluster is present, centered in the origin. It is found that the statistics of a Bayes optimal estimator can be characterised therefore by an order parameter q satisfying a self-consistent equation not different from the one presented above (we will use below a different font to stress that we are currently analysing the Bayes optimal setting)

$$\hat{\mathsf{q}}=\alpha\int_{\mathcal{Y}}\mathrm{d}y\,\mathbb{E}_{\sigma,\zeta}\left[\sigma^2 Z_0(y,\mu,V)\left(\partial_\mu\ln Z_0(y,\mu,V)\right)^2\Big|_{\substack{\mu=\sigma\sqrt{\mathsf{q}}\zeta\\ V=\sigma^2(\beta_\star^2-\mathsf{q})}}\right],\quad \mathsf{q}=\frac{\beta_\star^4\hat{\mathsf{q}}}{1+\beta_\star^2\hat{\mathsf{q}}}. \tag{60}$$

with $Z_0(y,\mu,V)\coloneqq\mathbb{E}_z[P_0(y|\mu+\sqrt{V}z)]$ with $z\sim\mathcal{N}(0,1)$. We can then compute the Bayes optimal estimation error for the Bayes optimal estimator $\hat{\boldsymbol{\beta}}_{\text{BO}}=\mathbb{E}_{\boldsymbol{\beta}|\mathcal{D}}[\boldsymbol{\beta}]$ as

$$\varepsilon_{\text{BO}}=\lim_{d\to+\infty}\frac{1}{d}\|\boldsymbol{\beta}_\star-\hat{\boldsymbol{\beta}}_{\text{BO}}\|_2^2=\beta_\star^2-\mathsf{q}. \tag{61}$$

B Asymptotic results for ridge-regularised losses

Let us fix now $r(\boldsymbol{x}) = \frac{1}{2}\|\boldsymbol{x}\|_2^2$. In this case, the computation of Ψ_w can be performed explicitly via a Gaussian integration, and the saddle-point equations can take a more compact form that is particularly suitable for a numerical solution. In particular, the prior proximal is found as

$$\boldsymbol{g} = \frac{\sum_c \left(\hat{m}_c \boldsymbol{\beta}_\star + d\hat{t}_c \boldsymbol{\mu}_c\right) + \sum_c \sqrt{\hat{q}_c}\boldsymbol{\xi}}{\lambda + \hat{v}} \tag{62}$$

so that the prior saddle-point equations obtained from Ψ_w become

$$\begin{aligned} q&=\tfrac{1}{d}\left(\textstyle\sum_c \frac{\hat{m}_c\boldsymbol{\beta}_\star + d\hat{t}\boldsymbol{\mu}_c}{\lambda+\hat{v}}\right)^2 + \left(\textstyle\sum_c \frac{\sqrt{q_c}}{\lambda+\hat{v}}\right)^2 & \hat{q}_c&=\alpha p_c \textstyle\int_{\mathcal{Y}} \mathrm{d}y \mathbb{E}_{\sigma_c,\zeta}\left[\sigma_c^2 Z_0 f_c^2\right], \\ m&=\frac{\sum_c\left(\beta_\star^2\hat{m}_c + t_c^0\hat{t}_c\right)}{\lambda+\hat{v}} & \hat{v}&=-\alpha\textstyle\sum_c p_c \int_{\mathcal{Y}} \mathrm{d}y\mathbb{E}_{\sigma_c,\zeta}\left[\sigma_c^2 Z_0 \partial_\omega f_c\right], \\ v_c&=\tfrac{1}{\lambda+\hat{v}}\textstyle\sum_{c'}\sqrt{\frac{\hat{q}_{c'}}{\hat{q}_c}} & \hat{m}_c&=\alpha p_c\textstyle\int_{\mathcal{Y}}\mathrm{d}y\mathbb{E}_{\sigma_c,\zeta}\left[\sigma_c^2\partial_\mu Z_0 f_c\right], \\ t_c&=\frac{\sum_{c'}\left(\hat{t}_{c'}\mu_{c'c} + t_c^0\hat{m}_{c'}\right)}{\lambda+\hat{v}}, & \hat{t}_c&=\alpha p_c\textstyle\int_{\mathcal{Y}}\mathrm{d}y\mathbb{E}_{\sigma_c,\zeta}[Z_0 f_c] \end{aligned} \tag{63}$$

We have used the shorthand notation $Z_0 \equiv Z_0\left(y, t_c^0 + \frac{\sigma_c m}{\sqrt{q}}\zeta, \sigma_c^2\beta_\star^2 - \frac{\sigma_c^2 m^2}{q}\right)$ and $\mu_{cc'} := d\boldsymbol{\mu}_{c'}^\intercal\boldsymbol{\mu}_c$.

Regression on one cloud: consistency In the special case in which $|\mathcal{C}| = 1$ and the cloud is centered in the origin, we have $t_1 = \hat{t}_1 = 0$ and, dropping the subscript referring to the cluster,

$$\begin{aligned} q &= \frac{\beta_\star^2\hat{m}^2 + \hat{q}}{(\lambda+\hat{v})^2} & \hat{q} &= \alpha\textstyle\int_{\mathcal{Y}}\mathrm{d}y\,\mathbb{E}_{\sigma,\zeta}\left[\sigma^2 Z_0\left(y, \frac{\sigma m}{\sqrt{q}}\zeta, \sigma^2\beta_\star^2 - \frac{\sigma^2 m^2}{q}\right)f^2\right], \\ m &= \frac{\beta_\star^2\hat{m}}{\lambda+\hat{v}} & \hat{v} &= -\alpha\textstyle\int_{\mathcal{Y}}\mathrm{d}y\,\mathbb{E}_{\sigma,\zeta}\left[\sigma^2\, Z_0\left(y, \frac{\sigma m}{\sqrt{q}}\zeta, \sigma^2\beta_\star^2 - \frac{\sigma^2 m^2}{q}\right)\partial_\omega f\right], \\ v &= \frac{1}{\lambda+\hat{v}}, & \hat{m} &= \alpha\textstyle\int_{\mathcal{Y}}\mathrm{d}y\,\mathbb{E}_{\sigma,\zeta}\left[\sigma^2\,\partial_\mu Z_0\left(y, \frac{\sigma m}{\sqrt{q}}\zeta, \sigma^2\beta_\star^2 - \frac{\sigma^2 m^2}{q}\right)f\right], \end{aligned} \tag{64}$$

where as usual $f := \arg\min_u\left[\frac{\sigma^2 v u^2}{2} + \ell(y, \omega + \sigma^2 v u)\right]$ and $\omega = \sigma\sqrt{q}\zeta$. Let us now perform the rescaling $v \mapsto \alpha v$, $\hat{q} \mapsto \alpha\hat{q}$, $\hat{m} \mapsto \alpha\hat{m}$, and $\hat{v} \mapsto \alpha\hat{v}$, where $v = O(1)$, $\hat{v} = O(1)$, $\hat{m} = O(1)$, $\hat{q} = O(1)$. Then, under these assumptions, in the $\alpha \to +\infty$ limit

$$\begin{aligned} q &= \frac{\beta_\star^2\hat{m}^2}{\hat{v}^2} & \hat{q} &= \textstyle\int_{\mathcal{Y}}\mathrm{d}y\,\mathbb{E}_{\sigma,\zeta}\left[\sigma^2 Z_0\left(y, \frac{\sigma m}{\sqrt{q}}\zeta, \sigma^2\beta_\star^2 - \frac{\sigma^2 m^2}{q}\right)f^2\right], \\ m &= \frac{\beta_\star^2\hat{m}}{\hat{v}} & \hat{v} &= -\textstyle\int_{\mathcal{Y}}\mathrm{d}y\,\mathbb{E}_{\sigma,\zeta}\left[\sigma^2\, Z_0\left(y, \frac{\sigma m}{\sqrt{q}}\zeta, \sigma^2\beta_\star^2 - \frac{\sigma^2 m^2}{q}\right)\partial_\omega f\right], \\ v &= \frac{1}{\hat{v}}, & \hat{m} &= \textstyle\int_{\mathcal{Y}}\mathrm{d}y\,\mathbb{E}_{\sigma,\zeta}\left[\sigma^2\,\partial_\mu Z_0\left(y, \frac{\sigma m}{\sqrt{q}}\zeta, \sigma^2\beta_\star^2 - \frac{\sigma^2 m^2}{q}\right)f\right], \end{aligned} \tag{65}$$

Moreover, in the large α limit, $f = -\partial_\eta\ell(y,\eta)|_{\eta=\sigma\sqrt{q}\zeta}$. It follows that, independently from the adopted loss, the angle $\pi^{-1}\arccos\frac{m}{\beta_\star\sqrt{q}}$ between the estimator $\hat{\boldsymbol{\beta}}_\lambda$ and the teacher $\boldsymbol{\beta}_\star$ goes to zero as $\alpha \to +\infty$. In this limit, it is easily found that $\varepsilon_{\text{est}} \to \beta_\star^{-2}(m - \beta_\star^2)^2$, hence the estimator is consistent if $m \to \beta_\star^2$.

Uncorrelated teachers: universality To study the universality properties in the ridge setting, let us introduce two possible new assumptions.

Assumption B.1. *For all $c \in [K]$, $\lim_{d\to+\infty} t_c^0 = 0$.*

This assumption holds, for example, by assuming the centroids $\boldsymbol{\mu}_c \sim \mathcal{N}(\boldsymbol{0}, {}^1\!/\!_d\boldsymbol{I}_d)$. It expresses in general the fact that the teacher $\boldsymbol{\beta}_\star$ is completely uncorrelated from the different centroids $\boldsymbol{\mu}_c$.

Assumption B.2. *The following symmetry properties hold*

$$P_0(y|\tau) = P_0(-y|-\tau), \qquad \ell(y,\eta) = \ell(-y,-\eta). \tag{66}$$

Under Assumption B.1 and Assumption B.2, $t_c = \hat{t}_c = 0\ \forall c$ is a saddle-point solution of the equations 63. Indeed, if $\hat{t}_c = 0$ the prior equation implies $t_c = 0$. On the other hand, if $t_c = 0$, $\hat{t}_c = 0$ for parity reason (Pesce et al., 2023). We recover therefore in our setting the *mean universality* discussed by Pesce et al. (2023) in the Gaussian setting: the learning task is mean-independent and equivalent to one on c clouds all centered in the origin, i.e., a problem obtained assuming $\boldsymbol{x} \sim \sum_c p_c \mathbb{E}_{\sigma_c}[\mathcal{N}(\mathbf{0}, {}^{\sigma_c}/{}_d \boldsymbol{I}_d)]$. Note that in the Gaussian setting (i.e., assuming $\rho_c(\sigma) = \delta(\sigma - \bar{\sigma}_c)$ for some fixed $\bar{\sigma}_c$ for each $c \in [C]$) Pesce et al. (2023) also observed that in the limit $\lambda \to 0^+$, *covariance universality* holds in the Gaussian case: ε_g and ε_ℓ are independent on the covariance of the clouds. This fact does not extend to the case in which the distribution of σ_c is not atomic (not even in the case in which σ_c are all identically distributed), as it has been verified by Adomaityte et al. (2023).

B.1 Square loss

If we consider a square loss $\ell(y,\eta) = \frac{1}{2}(y-\eta)^2$, then an explicit formula for the proximal can be found, namely

$$f_c = \frac{y - \omega_c}{1 + v_c \sigma_c^2}, \qquad \omega_c = t_c + \sigma_c \sqrt{q} \zeta, \tag{67}$$

so that the second set of saddle-point equations Eq. 50 can be made even more explicit. Let us assume, that labels are generated according to the linear model in Eq. 1a, where the noise term η_i has $\mathbb{E}[\eta_i] = 0$ and $\hat{\sigma}_0^2 \coloneqq \mathbb{E}[\eta_i^2] < +\infty$. In this setting, the channel equations can be written as

$$\hat{q}_c = \alpha p_c \hat{\sigma}_0^2 \mathbb{E}_{\sigma_c}\left[\frac{\sigma_c^2}{(1+v_c\sigma_c^2)^2}\right] + \alpha p_c \mathbb{E}_{\sigma_c}\left[\left(\frac{\sigma_c^2}{1+v_c\sigma_c^2}\right)^2\right](\beta_\star^2 - 2m + q + (t_c^0 - t_c)^2), \tag{68a}$$

$$\hat{v} = \alpha \sum_c p_c \mathbb{E}_{\sigma_c}\left[\frac{\sigma_c^2}{1+v_c\sigma_c^2}\right], \tag{68b}$$

$$\hat{m}_c = \alpha p_c \mathbb{E}_{\sigma_c}\left[\frac{\sigma_c^2}{1+v_c\sigma_c^2}\right] \tag{68c}$$

$$\hat{t}_c = \alpha p_c (t_c^0 - t_c) \mathbb{E}_{\sigma_c}\left[\frac{1}{1+v_c\sigma_c^2}\right]. \tag{68d}$$

In this setting the generalisation error becomes

$$\varepsilon_g \coloneqq \lim_{d\to+\infty} \mathbb{E}\left[\left(y - \hat{\boldsymbol{\beta}}_\lambda^\intercal \boldsymbol{x}\right)^2\right] = \hat{\sigma}_0^2 + \sum_c p_c (t_c^0 - t_c)^2 + (\beta_\star^2 - 2m + q)\sum_c p_c \mathbb{E}_{\sigma_c}[\sigma_c^2], \tag{69}$$

which is finite if and only if $\mathbb{E}_{\sigma_c}[\sigma_c^2] < +\infty$ for all c. Note that the dependence on $\hat{\varrho}$ is through its second moment only. Observe that the possible power-law behavior of the noise on the label *does not* influence the generalisation performances, that only depends on the noise variance $\hat{\sigma}_0^2$. The training loss, on the other hand, is

$$\varepsilon_\ell \coloneqq \lim_{d\to+\infty} \frac{1}{2n}\sum_{i=1}^n \left(y_i - \hat{\boldsymbol{\beta}}_\lambda^\intercal \boldsymbol{x}_i\right)^2 \xrightarrow{d\to+\infty} \frac{\hat{\sigma}_0^2}{2}\sum_c p_c \mathbb{E}_{\sigma_c}\left[\frac{1}{(1+v_c\sigma_c^2)^2}\right] + \\ + \sum_c \frac{p_c}{2}\left[\mathbb{E}_{\sigma_c}\left[\frac{1}{(1+v_c\sigma_c^2)^2}\right](t_c^0 - t_c)^2 + \mathbb{E}_{\sigma_c}\left[\frac{\sigma_c^2}{(1+v_c\sigma_c^2)^2}\right](\beta_\star^2 - 2m + q)\right]. \tag{70}$$

Strong universality of ε_ℓ for $\lambda \to 0^+$ We will show now that, under the Assumption B.1 (Assumption B.2 is satisfied in the setting under consideration), the strong universality of the training loss observed by Pesce et al. (2023) is preserved. Let us put ourselves in the case of a single cluster centered in the origin (an assumption that is not restrictive, as shown above). In this case, let us introduce

$$S_v \coloneqq \mathbb{E}_\sigma\left[\frac{1}{1+v\sigma^2}\right] \tag{71}$$

which can be interpreted in terms of the Stieljes transform of the random variable σ^2. The saddle-point equations are

$$\begin{aligned} &\hat{q}{=}{-}\alpha\hat{\sigma}_0^2\partial_v S_v{+}\alpha(1{-}S_v{+}v\partial_v S_v)\frac{\beta_\star^2{-}2m{+}q}{v^2}, && q{=}\frac{\hat{m}^2\beta_\star^2{+}\hat{q}}{(\lambda{+}\hat{v})^2}\\ &\hat{v}{=}\alpha\frac{1{-}S_v}{v}, && m{=}\frac{\beta_\star^2\hat{m}}{\lambda{+}\hat{v}}\\ &\hat{m}{=}\alpha\frac{1{-}S_v}{v}, && v{=}\frac{1}{\lambda{+}\hat{v}}. \end{aligned} \tag{72}$$

The training loss can be written as

$$\varepsilon_\ell = -\frac{\hat{\sigma}_0^2\partial_v S_v}{2} - \frac{(\beta_\star^2 - 2m + q)\partial_v S_v}{2}. \tag{73}$$

In the limit $\lambda \to 0$,

$$x \coloneqq \frac{\beta_\star^2 - 2m + q}{v} = \frac{(1 - S_v + v\partial_v S_v)x - v\partial_v S_v\hat{\sigma}_0^2}{1 - S_v} \Rightarrow x = \hat{\sigma}_0^2 \tag{74}$$

so that $\varepsilon_\ell = \frac{1}{2}S_v\hat{\sigma}_0^2$. The quantity S_v can be extracted from the equation for v, as it has to satisfy, in the zero regularisation limit, $\alpha(1 - S_v) = 1 \Rightarrow S_v = 1 - \frac{1}{\alpha}$ which is a valid solution for $\alpha > 1$ only. As a result, we obtain a *universal* formula for the training loss

$$\varepsilon_\ell = \frac{\hat{\sigma}_0^2}{2}\left(1 - \frac{1}{\alpha}\right)_+, \qquad \text{where} \quad (x)_+ \coloneqq x\theta(x). \tag{75}$$

Note that the formula above is valid for *any* distribution of σ, including distributions with no second moment.

Generalisation error rate — We conclude this section by extracting the generalisation error rate for $n \to +\infty$ and large but fixed d, i.e., for $\alpha \to +\infty$. For simplicity, let us focus, once again, on the case $K = 1$ and $\boldsymbol{\mu}_1 = \mathbf{0}$, corresponding to the fixed-point equations given in Eq. 72. Let us assume that $\hat{\sigma}_0^2 < +\infty$ and that $\sigma_0^2 \coloneqq \mathbb{E}[\sigma^2] < +\infty$. From Eq. 72 v satisfies the equation $\alpha(1 - S_v) = 1 - \lambda v$: as $S_v \in [0, 1]$ and $v > 0$, then for $\alpha \to +\infty$ we must have $S_v \to 1$ and $v \to 0$, so that for $\alpha \to +\infty$, $S_v = 1 - \frac{1}{\alpha} + O(\alpha^{-1})$. In this limit, therefore, by direct inspection of the fixed-point equations, $q \to \beta_\star^2$ and $m \to \beta_\star^2$ so that $\varepsilon_{\text{est}} \to 0$ and the estimator $\hat{\boldsymbol{\beta}}_\lambda$ is unbiased.

In the hypothesis that σ_0^2 is finite (i.e., $\varrho(\sigma) \sim \sigma^{-2a-1}$ with $a > 1$ for $\sigma \gg 1$), then, for small v, as $S_v \simeq 1 - v\sigma_0^2 + o(v)$, it is found that

$$q = \beta_\star^2 + \frac{\hat{\sigma}_0^2 - 2\beta_\star^2\lambda}{\sigma_0^2}\frac{1}{\alpha} + o\left(\frac{1}{\alpha}\right), \qquad m = \beta_\star^2 - \frac{\lambda\beta_\star^2}{\sigma_0^2}\frac{1}{\alpha} + o\left(\frac{1}{\alpha}\right), \qquad v = \frac{1}{\sigma_0^2\alpha} + o\left(\frac{1}{\alpha}\right), \tag{76}$$

which, together with our general formulas for ε_{est}, imply $\varepsilon_{\text{est}} \sim \alpha^{-1}$ for large α.

On the other hand, let us consider the case in which $\varrho(\sigma) \sim \sigma^{-2a-1}$ for $\sigma \gg 1$, with $0 < a < 1$. In this case, $\sigma_0^2 = +\infty$ and S_v has an expansion in the form $S_v = 1 - \tilde{\sigma}_0^2 v^a + O(v)$ for some finite positive quantity $\tilde{\sigma}_0^2$. Such asymptotic implies that $v \simeq (\tilde{\sigma}_0^2\alpha)^{-1/a}$ for $\alpha \gg 1$. By replacing this in the fixed point equations, it is found that $m = \beta_\star^2 - \lambda\beta_\star^2(\tilde{\sigma}_0^2\alpha)^{-1/a} + o(1/\alpha)$ and $q = \beta_\star^2 + [\hat{\sigma}_0^2 - 2\lambda\beta_\star^2](\tilde{\sigma}_0^2\alpha)^{-1/a} + o(1/\alpha)$, so that in the end $\varepsilon_{\text{est}} \sim \alpha^{-1/a}$.

The $a = 1$ case is marginal, as $S_v \simeq 1 + \tilde{\sigma}_0^2 v \log v$ for $\alpha \gg 1$ for some positive constant $\tilde{\sigma}_0^2$. Therefore $v = (\tilde{\sigma}_0^2\alpha\ln\alpha)^{-1}$. By consequence, for $\alpha \gg 1$ $m \simeq \beta_\star^2 - \lambda\beta_\star^2(\tilde{\sigma}_0^2\alpha\ln\alpha)^{-1}$ and $q \simeq \beta_\star^2 + \left[\hat{\sigma}_0^2 - 2\lambda\beta_\star^2\right](\tilde{\sigma}_0^2\alpha\ln\alpha)^{-1}$, so that $\varepsilon_{\text{est}} \sim (\alpha\ln\alpha)^{-1}$.

B.2 Huber loss and robust regression in the presence of fat tails

B.2.1 A model for the study of robustness

A toy model for the study of robustness has been introduced recently by Vilucchio et al. (2023). Here we will consider a more general setting to include the possibility of having fat tails. We consider the case of one cloud only, $K = 1$, so that $P(\boldsymbol{x}) = \mathbb{E}_\sigma[\mathcal{N}(\boldsymbol{x}; \mathbf{0}, \sigma^2/d \boldsymbol{I}_d)]$, and $P_0(y|\tau) = \mathbb{E}_{\hat\sigma}[\mathcal{N}(y; \hat\eta\tau, \hat\sigma^2)]$, where the expectation is taken over the joint distribution $\hat\varrho$ for the pair $(\hat\eta, \hat\sigma)$ of (possibly correlated) random variables. Vilucchio et al. (2023) adopted, in particular, the distribution $\hat\varrho(\hat\eta, \hat\sigma) = \epsilon\delta_{\hat\eta,\hat\eta_{\mathrm{out}}}\delta_{\hat\sigma,\hat\sigma_{\mathrm{out}}} + (1-\epsilon)\delta_{\eta,1}\delta_{\hat\sigma,\hat\sigma_{\mathrm{in}}}$ for $\epsilon \in [0, 1]$, with $(\hat\eta_{\mathrm{out}}, \hat\sigma_{\mathrm{out}})$ referring to "outlier labels", and $(1, \hat\sigma_{\mathrm{in}})$ referring to "inlier labels". The general fixed-point equations can be adapted to this case quite easily. We assume, once again, a ridge regularisation. Here we comment on the fact that in this case, it can be interesting to consider, beyond the ERM estimator $\hat{\boldsymbol\beta}_\lambda$ and the Bayes-optimal estimator $\hat{\boldsymbol\beta}_{\mathrm{BO}}$, the estimator that minimises the (posterior-averaged) mean-square test error

$$\hat{\boldsymbol\beta}_{g,\mathrm{BO}} = \arg\min_{\boldsymbol\beta} \mathbb{E}_{\hat{\boldsymbol\beta}|\mathcal{D}}\left[\mathbb{E}_{(y,\boldsymbol{x})|\hat{\boldsymbol\beta}}\left[\left(y - \boldsymbol\beta^\intercal \boldsymbol{x}\right)^2\right]\right] = \mathbb{E}[\hat\eta]\hat{\boldsymbol\beta}_{\mathrm{BO}}. \tag{77}$$

In the expression above, $\mathbb{E}_{(y,\boldsymbol{x})|\hat{\boldsymbol\beta}}$ expresses the fact that the pair $(y, \boldsymbol{x})$ has been generated with a teacher vector $\hat{\boldsymbol\beta}$, sampled by the posterior. Using the results on the Bayes optimal estimator, it is simple to derive the errors obtained by using $\hat{\boldsymbol\beta}_{g,\mathrm{BO}}$. Under the assumptions that $\sigma_0^2 \coloneqq \mathbb{E}[\sigma^2] < +\infty$ and $\hat\sigma_0^2 \coloneqq \mathbb{E}[\hat\sigma^2]$,

$$\varepsilon_{g,\mathrm{BO}} \coloneqq \mathbb{E}_{(y,\boldsymbol{x})}\left[\left(y - \boldsymbol{x}^\intercal \boldsymbol\beta_{g,\mathrm{BO}}\right)^2\right] = \sigma_0^2\left(\beta_\star^2\mathbb{E}[\hat\eta^2] - \mathbb{E}[\hat\eta]^2\mathsf{q}\right) + \hat\sigma_0^2,$$

where q is provided by Eq. 60. As in the pure Gaussian case, by imposing the ansatz $\mathsf{q} = \beta_\star^2 - \frac{1}{\alpha}q_0 + \Theta(\alpha^{-2})$, and consequently $\hat{\mathsf{q}} = \alpha\hat q_0 + \Theta(1)$ for large α, we can obtain

$$\frac{1}{q_0} = \hat q_0 = \int_y \mathrm{d}y\, \mathbb{E}_{\sigma,\zeta}\left[\sigma^2 P_0(y|\omega)\left(\partial_\omega \ln P_0(y|\omega)\right)^2\Big|_{\omega=\sigma\beta_\star\zeta}\right]. \tag{78}$$

In the $\alpha \to +\infty$ limit, then, $\mathsf{q} \to \rho$ and $\varepsilon_{\mathrm{est,BO}} \coloneqq \lim_{d\to+\infty}\frac{1}{d}\mathbb{E}_{\mathcal{D}}[\|\hat{\boldsymbol\beta}_{g,\mathrm{BO}} - \boldsymbol\beta\|_2^2] = \frac{q_0}{\alpha} + \Theta(\alpha^{-2}) \to 0$. On the other hand, $\varepsilon_{g,\mathrm{BO}} = \hat\sigma_0^2 + \sigma_0^2\beta_\star^2\mathrm{Var}[\hat\eta] - \frac{\sigma_0^2 q_0}{\alpha} + \Theta(\alpha^{-2})$.

B.2.2 Huber loss and its application

The Huber loss is a strongly convex loss depending on a tunable parameter $\delta \geq 0$ and is defined as

$$\ell_\delta(y, \eta) = \begin{cases} \frac{(y-\eta)^2}{2} & \text{if } |y - \eta| < \delta \\ \delta|y - \eta| - \frac{\delta^2}{2} & \text{otherwise.} \end{cases} \tag{79}$$

This loss is widely adopted in robust regression as it is less sensitive to outliers than the most commonly adopted square loss, and is associated with the following expression for the proximal

$$h_c = \omega_c + \frac{(y-\omega_c)v_c\sigma_c^2}{\max(\delta^{-1}|y-\omega_c|, 1+v_c\sigma_c^2)} \Leftrightarrow f_c = \frac{y-\omega_c}{\max(\delta^{-1}|y-\omega_c|, 1+v_c\sigma_c^2)}, \quad \omega_c = t_c + \sigma_c\sqrt{q}\zeta. \tag{80}$$

The prior equations are therefore the usual in Eq. 64. The channel equations are instead

$$\hat m = \alpha\mathbb{E}\left[\frac{\sigma^2\hat\eta \operatorname{erf}\chi}{1 + v\sigma^2}\right] \tag{81a}$$

$$\hat q = \alpha\mathbb{E}\left[\frac{\sigma^2\psi \operatorname{erf}\chi}{(1 + v\sigma^2)^2} + \sigma^2\delta^2(1 - \operatorname{erf}\chi) - \sqrt{\frac{2\psi}{\pi}}\frac{\sigma^2\delta\, \mathrm{e}^{-\chi^2}}{1 + v\sigma^2}\right] \tag{81b}$$

$$\hat v = \alpha\mathbb{E}\left[\frac{\sigma^2 \operatorname{erf}\chi}{1 + v\sigma^2}\right]. \tag{81c}$$

where the expectation is over all random variables involved in the expressions (namely, σ, $\hat{\sigma}$, and $\hat{\eta}$) and we used the short-hand notation

$$\psi \coloneqq \hat{\sigma}^2 + \sigma^2(\hat{\eta}^2\beta_\star^2 - 2\hat{\eta}m + q), \qquad \chi \coloneqq \frac{\delta(1+v\sigma^2)}{\sqrt{2\psi}} \tag{82}$$

Note that we recover the expressions obtained for the square loss for $\delta \to +\infty$.

With the usual notation convention $\hat{\sigma}_0^2 \coloneqq \mathbb{E}[\hat{\sigma}^2]$ and $\sigma_0^2 \coloneqq \mathbb{E}[\sigma^2]$, the estimation error is given by the general formula in Eq. 57, whereas the generalisation error is

$$\varepsilon_g \coloneqq \mathbb{E}\left[(y - \hat{\boldsymbol{\beta}}_\lambda^\intercal \boldsymbol{x})^2\right] = \hat{\sigma}_0^2 + (\beta_\star^2\mathbb{E}[\hat{\eta}^2] - 2\mathbb{E}[\hat{\eta}]m + q)\sigma_0^2, \tag{83}$$

ε_g being finite if $\sigma^2 < +\infty$, $\hat{\sigma}_0^2 < +\infty$ and $\mathbb{E}[\hat{\eta}] < +\infty$. We aim now at extrapolating the large-α behavior of such errors and at studying the consistency of $\hat{\boldsymbol{\beta}}_\lambda$ with respect to the Bayes optimal estimators discussed in Section A.2. To do so, we rescale $\hat{m} \mapsto \alpha\hat{m}$, $\hat{v} \mapsto \alpha\hat{v}$, $v \mapsto \alpha^{-1}v$ and $\hat{q} \mapsto \alpha\hat{q}$. We also assume that $\lambda \mapsto \lambda + \alpha\lambda'$ (the role of $\lambda' \neq 0$ will be clear in the following). The set of fixed point equations become, for $\alpha \to +\infty$

$$\begin{aligned}
&\hat{m} = \mathbb{E}\left[\sigma^2\hat{\eta}\,\mathrm{erf}\,\bar{\chi}\right] && q = \frac{\beta_\star^2\hat{m}^2}{(\lambda'+\hat{v})^2} \\
&\hat{q} = \mathbb{E}\left[\sigma^2\psi\,\mathrm{erf}\,\bar{\chi} + \sigma^2\delta^2(1-\mathrm{erf}\,\bar{\chi}) - \sqrt{\frac{2\psi}{\pi}}\sigma^2\delta\,\mathrm{e}^{-\bar{\chi}^2}\right] && m = \frac{\beta_\star^2\hat{m}}{\lambda'+\hat{v}} \qquad , \qquad \bar{\chi} \coloneqq \frac{\delta}{\sqrt{2\psi}}. \\
&\hat{v} = \mathbb{E}\left[\sigma^2\,\mathrm{erf}\,\bar{\chi}\right]. && v = \frac{1}{\lambda'+\hat{v}}
\end{aligned} \tag{84}$$

In this limit, as $\beta_\star^2 q = m^2$, $\psi = \hat{\sigma}_0^2 + \frac{\sigma_0^2}{\beta_\star^2}(m - \beta_\star^2\hat{\eta})^2$, so that

$$\varepsilon_{\mathrm{est}} = \frac{(m-\beta_\star^2)^2}{\beta_\star^2}, \qquad \varepsilon_g = \hat{\sigma}_0^2 + \sigma^2\frac{\mathbb{E}[(m-\hat{\eta}\beta_\star^2)^2]}{\beta_\star^2}. \tag{85}$$

It is possible to choose λ' so that $\lim_{\alpha\to+\infty}\varepsilon_g = \lim_{\alpha\to+\infty}\varepsilon_g^{\mathrm{BO}}$, i.e.,

$$\hat{\sigma}_0^2 + \sigma_0^2\frac{\mathbb{E}[(m-\hat{\eta}\beta_\star^2)^2]}{\beta_\star^2} = \hat{\sigma}_0^2 + \sigma_0^2\beta_\star^2\mathrm{Var}[\hat{\eta}] \Rightarrow m = \beta_\star^2\mathbb{E}[\hat{\eta}]. \tag{86}$$

We can try to satisfy this condition by tuning properly λ_1, under the constraint that $\lambda' \geq 0$. We can write in particular

$$\lambda' = \frac{\hat{m}}{\mathbb{E}[\hat{\eta}]} - \hat{v} = \frac{\mathbb{E}[\sigma^2\hat{\eta}\,\mathrm{erf}\,\bar{\chi}] - \mathbb{E}[\hat{\eta}]\mathbb{E}[\sigma^2\,\mathrm{erf}\,\bar{\chi}]}{\mathbb{E}[\hat{\eta}]} \geq 0 \Rightarrow \mathbb{E}[\sigma^2(\hat{\eta} - \mathbb{E}[\hat{\eta}])\,\mathrm{erf}\,\bar{\chi}] \geq 0 \tag{87}$$

to be computed with

$$\bar{\chi} \equiv \frac{\delta}{\sqrt{2\psi_g}}, \quad \psi_g = \hat{\sigma}_0^2 + \sigma_0^2\beta_\star^2(\mathbb{E}[\hat{\eta}] - \hat{\eta})^2. \tag{88}$$

Note that the condition is always satisfied in the case of the square loss (i.e., for $\delta \to +\infty \Leftrightarrow \bar{\chi} \to 1$).

Consistency of the estimator. The consistency of the estimator can be imposed by properly tuning λ, by requiring that $\lim_{\alpha\to+\infty}\varepsilon_{\mathrm{est}} = 0$, i.e., $m = \beta_\star^2$ in this limit. In the same spirit as above, this implies a condition on λ' given by

$$\lambda' = \hat{m} - \hat{v} = \mathbb{E}[\sigma^2\hat{\eta}\,\mathrm{erf}\,\bar{\chi}] - \mathbb{E}[\sigma^2\,\mathrm{erf}\,\bar{\chi}] \geq 0 \Rightarrow \mathbb{E}[\sigma^2(\hat{\eta}-1)\,\mathrm{erf}\,\bar{\chi}] \geq 0 \tag{89}$$

to be computed with

$$\bar{\chi} \equiv \frac{\delta}{\sqrt{2\psi_{\mathrm{est}}}}, \quad \psi_{\mathrm{est}} = \hat{\sigma}_0^2 + \sigma_0^2\beta_\star^2(1-\hat{\eta})^2. \tag{90}$$

When imposing the equality, the conditions above provide the values of δ (if any) for a consistent estimator if $\lambda = \Theta(1)$ in the $\alpha \to +\infty$ limit.

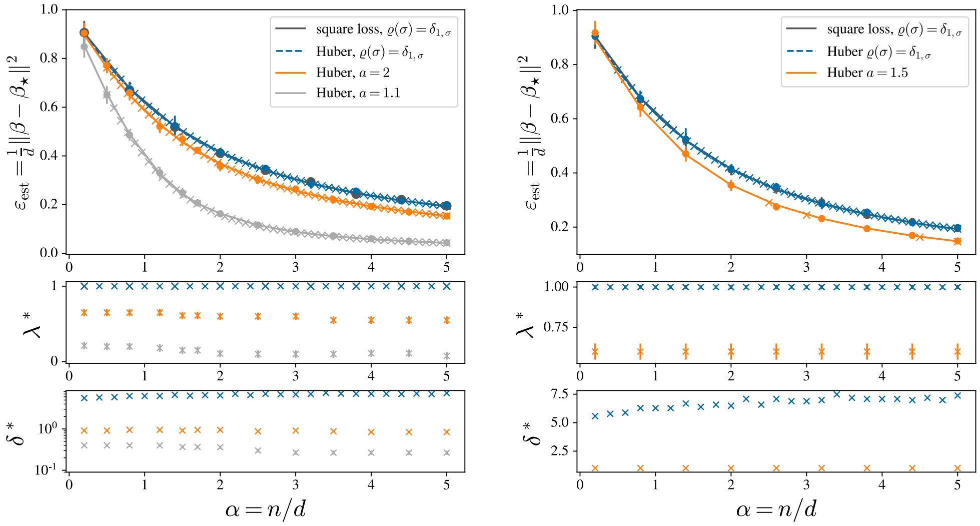

Figure 5: Gaussian covariates fully-contaminated by fat-tailed noise with distribution $p_\eta(\eta) = \mathbb{E}_\sigma[\mathcal{N}(x; 0, \sigma^2)]$, where parameters of various $\varrho(\sigma)$ are varied: inverse-gamma (see Table. 1, **left**) and Pareto (**right**). (**Top**) Estimation error ε_{est} as a function of the sample complexity $\alpha = {}^n/_d$ for optimally regularised ridge regression (black), Huber with optimal location parameter and optimal regularisation (orange) and Bayes-optimal performance (crosses). (**Center.**) Value of the optimal regularisation parameter $\lambda^\star$ for the Huber loss. (**Bottom.**) Value of the optimal location parameter $\delta^\star$ for the Huber loss. Both optimal values are displayed by varying the scale parameter a controlling the tails of the noise distribution. Dots indicate numerical simulations averaged over 20 seeds with $d = 10^3$.

C Further numerical results

C.1 Further results for the case of fat-tailed noise

In this Appendix, we add some details about the case of ϵ-contamination in the labels, as in Eq. 5, for different ϱ_0 generating the contaminating noise. Figure 5 compares the performance of various losses for different fully-contaminated ($\epsilon_{\rm n} = 1$) label noise distributions, obtained picking for $\varrho_0(\sigma)$ taken to be inverse-Gamma as in Table 1 with $a = b + 1 > 1$ (left) or Pareto (right). In all cases, the chosen parametrisations enforce unit variance for the noise, $\mathbb{E}[\eta^2] = 1$. Taking the limit $a \to \infty$ in the inverse-Gamma or in the Pareto distributions results in recovering the Gaussian distribution for label noise. In our experiments, we observe the same phenomenology as in Fig. 2 (bottom) for all these densities which generate different noise label distributions. As long as the label noise variance is kept the same, optimally tuned regularised Huber loss performs the task better with heavier tails, in terms of the estimation Eq. 11. As in the cases discussed in the main text, optimally regularised Huber achieves Bayes-optimal performance.