Asymptotic Characterisation of the Performance of Robust Linear Regression in the Presence of Outliers

Matteo Vilucchio
IdePHICS laboratory,
EPFL, 1015, Lausanne,
Switzerland

Emanuele Troiani
SPOC laboratory,
EPFL, 1015, Lausanne,
Switzerland

Vittorio Erba
SPOC laboratory,
EPFL, 1015, Lausanne,
Switzerland

Florent Krzakala
IdePHICS laboratory,
EPFL, 1015, Lausanne,
Switzerland

Abstract

We study robust linear regression in high-dimension, when both the dimension d and the number of data points n diverge with a fixed ratio $\alpha = n/d$, and study a data model that includes outliers. We provide exact asymptotics for the performances of the empirical risk minimisation (ERM) using ℓ_2-regularised ℓ_2, ℓ_1, and Huber losses, which are the standard approach to such problems. We focus on two metrics for the performance: the generalisation error to similar datasets with outliers, and the estimation error of the original, unpolluted function. Our results are compared with the information theoretic Bayes-optimal estimation bound. For the generalization error, we find that optimally-regularised ERM is asymptotically consistent in the large sample complexity limit if one perform a simple calibration, and compute the rates of convergence. For the estimation error however, we show that due to a norm calibration mismatch, the consistency of the estimator requires an oracle estimate of the optimal norm, or the presence of a cross-validation set not corrupted by the outliers. We examine in detail how performance depends on the loss function and on the degree of outlier corruption in the training set and identify a region of parameters where the optimal performance of the Huber loss is identical to that of the ℓ_2 loss, offering insights into the use cases of different loss functions.

---

1 INTRODUCTION

The problem of robust regression — *i.e.* linear regression with spurious outliers in the training set — has been extensively studied in the classical low-dimensional setting [Huber, 2011]. More recently, the high-dimensional setting has also been considered mainly concerning the characterisation of the asymptotic performance of robust estimators — see Section 2 for a review of related works. In this paper, we provide new theoretical results on the problem in the high-dimensional regime where the dimensionality of the problem d is comparable to the number of training samples n available ($n, d \to \infty$ with fixed ratio $\alpha = n/d$), investigating in particular the phenomenology arising due to the presence of outliers.

We build upon the recent progresses in high-dimensional statistics that characterise the asymptotic performance of the Bayes Optimal (BO) and the Empirical Risk Minimisation (ERM) estimators for general data models learned with generalised linear models. In particular, we consider a probabilistic model of training datasets — first introduced in [Box and Tiao, 1968] — that features a tunable presence of outliers and we investigate the high-dimensional setting performances of some standard regression estimators. We focus our attention on empirical risk minimisers under the ℓ_2-regularised ℓ_2, ℓ_1 and Huber losses [Huber, 1964], and study how they asymptotically perform in various outlier settings (e.g. small/large outlier percentage, small/large outlier variance). We focus on two metrics for the performance, the *generalisation error* to similar datasets with outliers, and the *estimation error* of the original, noiseless function, and we compare to information theoretic Bayes-optimal bounds.

For the probabilistic model of outliers we consider, eq. (1), we provide closed-form expressions for the high-dimensional limit of the generalisation error of ERM for the ℓ_2, ℓ_1 and Huber losses, and for the BO estimator (for which we compute exact rates at large

sample complexity α). **Our main results are:**

• **Optimally-regularised ERM is not always consistent.** Based on the value of the parameters characterising the outliers' distribution, optimally-regularised ERM estimation can fail to achieve BO performance for large sample complexity α, i.e. fail to achieve consistency. This happens for both the generalisation error and for the estimation error, in different regions of the parameter space. Notably, for the generalisation error there is a regime in which optimally-regularised ERM with ℓ_1 loss or the Huber loss (with un-optimised scale parameter) is not consistent, while optimally-regularised ERM with ℓ_2 and optimally-scaled Huber loss is consistent. This demonstrates the superiority of optimally-scaled Huber loss over the other losses.

• **Calibration of ERM minimiser norm restores consistency.** In the aforementioned cases in which ERM is not consistent, the inconsistency is caused exclusively by a mismatch between the norm of the ERM minimiser and that of the ground-truth. We show that consistency for the generalization error, and optimal rates, can be recovered by cross-validating on a test set to find the optimal norm.

• **Non-consistency of the estimation error.** Even when the cross-validated estimator is consistent and achieves optimal rates for the generalization error, it may remain inconsistent for the estimation error on the original, unpolluted function, which is arguably a more interesting objective. While this, too, can be calibrated, such a calibration requires either having access to an oracle knowledge of the BO norm, or to a clean, unpolluted dataset to cross-validate for the optimal norm. In the absence of such an ideal hold-out dataset, the Huber and ℓ_1 methods do *not* lead to consistent estimates of the objective functions, and robust estimation is biased.

• **Dependency of the generalisation error on the parameters.** We study the dependency of the generalisation error on the amount and strength of the outliers in the training dataset. We find that there is an extended region of parameters where the performance of the Huber loss is the same as the simpler ℓ_2 function, offering insights into the use cases of different loss functions.

In Section 2 we discuss related works. In Section 3 we introduce the data model we consider, and the noise model accounting for outliers. In Section 4 we provide the analytical characterisation for the estimation and generalisation errors for the ERM and BO estimators. In Section 5 we discuss the large sample complexity limit, and prove our results concerning the consistency of ERM estimators. In Section 6 we discuss the dependency of the generalisation error on the parameters characterising the outliers' distribution. We provide

more details and proofs figures in the Supplemental Materials. The code to reproduce the figures can be found at github.com/IdePHICS/RobustRegression.

2 RELATED WORKS

The problem of regression in the presence of outliers has been studied extensively, starting with Huber's seminal works [Huber, 1964, Huber, 1965] where he first introduced M-estimators as a robust alternative to classical estimators and secondly extended the concept of robustness to hypothesis testing. A comprehensive framework for understanding and defining robustness in statistical procedures was then given by [Hampel, 1968, Hampel, 1971].

Several algorithms have been proposed to approach problems with outliers — see the review [Huber, 2011]. In most cases, these are variants of standard linear regression based, for example, on custom loss functions [Huber, 1964], on hard-thresholding [Bhatia et al., 2017], on some sparsity assumption [Suggala et al., 2019] or on the sum of squares method [Raghavendra and Yau, 2020].

The choice of the outlier model greatly affects both the theoretical analyses and practical algorithms. The double Gaussian noise model that we consider has been first introduced in [Box and Tiao, 1968]. A similar model that considers outliers as a two-way process is studied in [Berger and Berliner, 1986]. Recently this model has been reconsidered by [Bellec, 2020], where they provide bounds on the generalisation error in the case of Huber-loss regression. Other models have also been proposed; see, for example, [Bean et al., 2013, Tsakonas et al., 2014, Yuen and Ortiz, 2017, Dalalyan and Thompson, 2019, Gagnon et al., 2020].

There have been numerous works analysing the generalisation error for M-estimation and robust regression in high dimensions. In [Karoui et al., 2013, Bean et al., 2013, Donoho and Montanari, 2016], they consider high-dimensional M-estimation for a model of data with exponentially-distributed outlier noise. They provide results on exact asymptotics and study the trade-off between using ℓ_2 and ℓ_1 loss. In [Hastie et al., 2022], they prove bounds for the precise asymptotics of the generalisation error of ℓ_2 regularised ℓ_2 loss, while in [Thrampoulidis et al., 2018], they prove asymptotics for M-estimation using Gordon min-max techniques. In [Liu et al., 2019], they consider a sparse model under an almost-high-dimensional limit $n \sim \log(d)$. In [Huang, 2020], a comparison of the precise asymptotics of ℓ_1-regularised ℓ_2, ℓ_1 and Huber loss is presented, with a focus on the interplay between the sparsity and the sample complexity of the

Matteo Vilucchio, Emanuele Troiani, Vittorio Erba, Florent Krzakala

problem. In [Bellec et al., 2022], they present results on the asymptotic normality of M-estimators.

The characterisation of the generalisation error for ERM that we use and specialise to our specific data model eq. (1) has been proven in [Aubin et al., 2020] and more recently in greater generality in [Loureiro et al., 2022]. The characterisation of the Bayes optimal performance has been rigorously established in [Barbier et al., 2019]. We are using these results as the basis of our investigations in this paper. We stress that all these works do not model outliers, thus providing no direct insight on robust regression.

3 DATA AND NOISE MODEL

We will consider a linear regression task in $\mathbb{R}^d$, *i.e.* we will fit a linear model — the "student" — $\hat{y}(\boldsymbol{x}; \boldsymbol{w}) = \boldsymbol{x} \cdot \boldsymbol{w}/\sqrt{d}$ to a dataset $\mathcal{D} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ made of n pairs of data-points $\boldsymbol{x}_i \in \mathbb{R}^d$ and labels $y_i \in \mathbb{R}$ sampled independently from a joint probability distribution $P(\boldsymbol{x}, y)$.

We will consider a probabilistic generative model $P(\boldsymbol{x}, y) = P_{\text{out}}(y|\boldsymbol{x} \cdot \boldsymbol{w}^\star/\sqrt{d})P_{\text{input}}(\boldsymbol{x})$. In this model, the dataset labels are generated by corrupting the output of a "teacher" linear model with weights $\boldsymbol{w}^*$ through an output channel P_{out}. We will consider the simple Gaussian case where the teacher vector is distributed as $\boldsymbol{w}^\star \sim \mathcal{N}(0, \mathbb{1}_d)$, and the samples $\boldsymbol{x}$ have Gaussian distribution $P_{\text{input}} = \mathcal{N}(0, \mathbb{1}_d)$. Assuming Gaussian data may seem limiting, but a long line of work has proven to various degrees that real data often behave as Gaussian data due to an underlying mechanism called "Gaussian universality", see for example [Hu and Lu, 2023]. Notice also that correlated feature may be addressed under the setting of [Loureiro et al., 2022]. We do not do that, as we found an interesting phenomenology already for uncorrelated features.

In order to model datasets with outliers, we consider the following generative process for the labels

$$y(x) = \begin{cases} y^\star + z\sqrt{\Delta_{\text{IN}}} & \text{with prob. } 1 - \epsilon \\ \beta y^\star + z\sqrt{\Delta_{\text{OUT}}} & \text{with prob. } \epsilon \end{cases} \quad (1)$$

where $y^\star = \boldsymbol{x} \cdot \boldsymbol{w}^\star/\sqrt{d}$ is the clean teacher-generated label, z is a standard Gaussian noise, Δ_{IN} is the variance of the non-outlying training points, Δ_{OUT} is the variance of the outlying training points, ϵ is the fraction of outliers and $\beta \geq 0$ is a parameter controlling the outliers' norm. In the following, we will consider often the special case of $\beta = 0$, but all of our results can be extended for any value $\beta \in \mathbb{R}_+$ (see Appendix A.2).

We will consider two different metrics to analyse the performance of the estimators. The first metric is the

generalisation error of a model with weights $\boldsymbol{w}$, defined as

$$E_{\text{gen}}(\boldsymbol{w}) = \mathbb{E}_{(\boldsymbol{x}_{\text{new}}, y_{\text{new}})} \left[\left(y_{\text{new}} - \frac{\boldsymbol{x}_{\text{new}} \cdot \boldsymbol{w}}{\sqrt{d}} \right)^2 \right], \quad (2)$$

where the average is over a sample-label pair $(\boldsymbol{x}_{\text{new}}, y_{\text{new}})$ generated with the same distribution as the training set. This metric is relevant for assessing the predictive power of the fitted model, as it computes the average error that the model achieves on an unseen input $\boldsymbol{x}_{\text{new}}$, extracted from the same distribution of the training dataset. The second metric we will consider is the estimation error, which is the mean-square error between the teacher and student weights, defined as

$$E_{\text{estim}}(\boldsymbol{w}) = \frac{1}{d}\|\boldsymbol{w}^\star - \boldsymbol{w}\|_2^2, \quad (3)$$

This metric is relevant for parameter estimation, where one is interested in recovering the value of the parameters used to generate the distribution. We remark that, as the sample distribution $P_{\text{input}}(\boldsymbol{x})$ is Gaussian, the estimation error also coincides with the generalisation error evaluated on a clean dataset

$$E_{\text{estim}}(\boldsymbol{w}) = \mathbb{E}_{\boldsymbol{x}} \left[\left(y^\star(\boldsymbol{x}) - \frac{\boldsymbol{x} \cdot \boldsymbol{w}}{\sqrt{d}} \right)^2 \right], \quad (4)$$

where the average is over the sample distribution $P_{\text{input}}(\boldsymbol{x})$, and $y^\star = \boldsymbol{x} \cdot \boldsymbol{w}^\star/\sqrt{d}$ is the noiseless teacher-generated label. Thus, in this setting, cross-validation on clean datasets is equivalent to cross-validation on the mean-square error with the ground truth. We will be interested in characterising the behaviour of the generalisation error and the estimation error when both the dimension d and the number of training samples n are large, $d, n \to \infty$ with finite sample complexity $\alpha = n/d$.

4 CHARACTERISATION OF THE PERFORMANCE OF BO AND ERM ESTIMATORS

Bayes Optimal estimator. The BO estimator is defined as the estimator achieving, on average over the data distribution, the lowest value of a given performance metric. Define the posterior distribution as

$$P(\boldsymbol{w} \mid \mathcal{D}) = \frac{1}{\mathcal{Z}_{\mathcal{D}}} P(\boldsymbol{w}) P(\mathcal{D} \mid \boldsymbol{w}), \quad (5)$$

where $\mathcal{Z}_{\mathcal{D}}$ is a normalisation factor, $\mathcal{D}$ is the training dataset, and in our case $P(\boldsymbol{w}) = \mathcal{N}(0, \mathbb{1}_d)$. It is well known [Cover and Thomas, 1991] that the optimal estimator error, as defined in eq. (3), is achieved from the posterior average

$$\hat{\boldsymbol{w}}_{\text{estim}}^{\text{BO}}(\mathcal{D}) = \mathbb{E}_{\boldsymbol{w}|\mathcal{D}}[\boldsymbol{w}] \quad (6)$$

Instead, the optimal estimator for the generalisation error, as defined in eq.(2) is given by (see Appendix A.1)

$$\hat{\boldsymbol{w}}_{\text{gen}}^{\text{BO}}(\mathcal{D}) = \mathbb{E}_{\boldsymbol{w}|\mathcal{D}}\left[\mathbb{E}_{\bar{\boldsymbol{x}},\bar{y}|\boldsymbol{w}}[\bar{y}\,\bar{\boldsymbol{x}}]\right], \qquad (7)$$

where $\mathbb{E}_{\boldsymbol{w}|\mathcal{D}}[\cdot]$ is the posterior average, and $\mathbb{E}_{\bar{\boldsymbol{x}},\bar{y}|\boldsymbol{w}}[\cdot]$ is the average over a sample-label pair $(\bar{\boldsymbol{x}}, \bar{y})$ generated with the data model eq. (1) using $\boldsymbol{w}$ as a teacher vector.

In the high-dimensional limit, one can characterise analytically the behaviour of both the BO generalisation error and the BO estimation error. This was first rigorously proven in [Barbier et al., 2019], whose results we adapt to our situation to characterise the information theoretic Bayesian bound. In [Barbier et al., 2019] the authors showed that in generalized linear models, the value of the estimation error concentrates on

$$E_{\text{estim}}^{\text{BO}} = 1 - q_{\text{b}}, \qquad (8)$$

while the value of the generalisation error concentrates on (here we specify their equations to the particular model of data used in the present paper)

$$E_{\text{gen}}^{\text{BO}} = 1 + \epsilon(\beta^2 - 1) - q_{\text{b}}(1 + \epsilon(\beta - 1))^2 + \Delta_{\text{eff}}, \quad (9)$$

where $\Delta_{\text{eff}} = (1 - \epsilon)\Delta_{\text{IN}} + \epsilon\Delta_{\text{OUT}}$ and the order parameter q_{b} — which can be interpreted as the normalised overlap between the posterior average and the teacher vector — can be found by solving the system of two equations

$$q_{\text{b}} = \frac{\hat{q}_{\text{b}}}{1 + \hat{q}_{\text{b}}},$$

$$\hat{q}_{\text{b}} = \alpha \int \mathbb{E}_{\xi}\left[\mathcal{Z}_{\text{out}}\left(\cdot\right)f_{\text{out}}\left(y, q_{\text{b}}^{1/2}\xi, \rho_{\text{w}^{\star}} - q_{\text{b}}\right)^2\right] dy, \tag{10}$$

with f_{out} and $\mathcal{Z}_{\text{out}}$ defined as

$$\begin{aligned} \mathcal{Z}_{\text{out}}(y, \omega, V) &= \mathbb{E}_{z \sim \mathcal{N}(\omega, V)}\left[P_{\text{out}}(y|z)\right], \\ f_{\text{out}}(y, \omega, V) &= \partial_{\omega} \log(\mathcal{Z}_{\text{out}}(y, \omega, V)), \end{aligned} \tag{11}$$

and evaluated both in the same parameters.

As $\alpha \to \infty$, $q_b \to 1$, showing that for large sample complexity the BO estimation goes to zero, while the BO generalisation error does not, as a non-zero excess error remains. For this reason, in the following we will often work with the *excess generalisation error*, defined as

$$\begin{aligned} E_{\text{gen}}^{\text{excess}}(\boldsymbol{w}) &= E_{\text{gen}}(\boldsymbol{w}) - E_{\text{gen}}^{\text{BO}}(\alpha = +\infty) \\ &= E_{\text{gen}}(\boldsymbol{w}) - \epsilon(1 - \epsilon)(1 - \beta)^2 - \Delta_{\text{eff}}. \end{aligned} \tag{12}$$

Notice that, in practical scenarios, computing the posterior average, or more generally providing an algorithm that efficiently reaches Bayes optimal performance — for example, in polynomial time in n, d — is challenging. In this specific case, an efficient algorithm to compute the BO estimator accurately in the considered limit on large n, d and fixed ratio n/d exists. It is the generalised approximate message passing (GAMP) algorithm [Kabashima and Uda, 2004, Rangan, 2011, Barbier et al., 2019], which we will use to plot numerical simulations for BO errors in the following (see Appendix B.3 for details).

We will use the BO estimator as a baseline to compare the performance of ERM with the selected losses to the optimal one in the presence of outliers. Specifically, eq. (8) and eq. (9) provide the information-theoretical lower bounds for the estimation error and generalisation error, respectively.

Empirical risk minimisation. Define the risk function

$$\mathcal{R}(\boldsymbol{w}) = \sum_{i=1}^{n} \mathcal{L}(y_i, \hat{y}_i(\boldsymbol{w})) + \lambda\, r(\boldsymbol{w}), \qquad (13)$$

where $\hat{y}_i(\boldsymbol{w}) = \boldsymbol{x}_i \cdot \boldsymbol{w}/\sqrt{d}$ are the labels predicted by the model on each training sample, y_i are the noisy labels of the training dataset, $\mathcal{L}$ is a convex function of $\hat{y}$, r a convex function of $\boldsymbol{w}$ and $\lambda \geq 0$. Minimising the risk function $\mathcal{R}(\boldsymbol{w})$ provides an estimate $\hat{\boldsymbol{w}}$ of the teacher vector $\boldsymbol{w}^{\star}$, and this estimate is uniquely defined thanks to convexity of the risk.

The risk function is a sum of a data-dependent term $\mathcal{L}(y, \hat{y})$ and of a regularisation function $r(\boldsymbol{w})$. For the data-dependent loss function $\mathcal{L}$, we will consider three convex losses: ℓ_2, ℓ_1 and Huber loss. The ℓ_2 loss $\ell_2(y, \hat{y}) = \frac{1}{2}(y - \hat{y})^2$ is the most common choice in regression and will provide a good baseline for more outlier-robust losses. The ℓ_1 loss $\ell_1(y, \hat{y}) = |y - \hat{y}|$ is the simplest choice for outlier-robust regression, as it penalises less poor interpolation of outliers w.r.t. the ℓ_2 loss. The Huber loss

$$\ell_a^{\text{Huber}}(y, y') = \begin{cases} \frac{1}{2}(y - y')^2 & \text{if } |y - y'| < a \\ a|y - y'| - \frac{a^2}{2} & \text{if } |y - y'| \geq a \end{cases} \tag{14}$$

has been proposed as a middle-ground between ℓ_2 and ℓ_1 [Huber, 1964]. It is more resilient to outliers than the ℓ_2 loss, while remaining smooth and convex as the ℓ_2 loss. Proper tuning of the scale parameter a effectively produces an ℓ_2 loss landscape for the non-outlying data points and an ℓ_1-like one for the outliers. For $a \to +\infty$ the Huber loss reduces to the ℓ_2 loss, while for $a \to 0$ there is a straightforward mapping to a ℓ_1 loss (one needs to rescale the total loss by a factor a to take this limit, obtaining an effective regularisation λa). While our analysis could be carried on for any choice of convex regulariser $r(\boldsymbol{w})$, we will only study the ℓ_2 regularisation $r(\boldsymbol{w}) = \frac{1}{2}\|\boldsymbol{w}\|_2^2$. Notice that due to our choices of scaling, for λ to have an effect on the

norm of the learned weights in the limit $\alpha \to \infty$, one needs the scaling $\lambda = \mathcal{O}(\alpha)$.

We will mostly consider optimally-tuned risk minimisation, meaning that we will set λ, and a in the case of the Huber loss, such as to minimise either the generalisation or the estimation error. In practice, the former is equivalent to cross-validation on a dataset statistically identical to the training set, while the latter to cross-validation on a clean dataset, without any noise or outliers. We will perform numerical simulations for the errors of ERM estimators using, for the ℓ_2 loss, the explicit solution for its minimiser. For Huber loss, we use the L-BFGS method [Liu and Nocedal, 1989] initialised on a random Gaussian weight. For the ℓ_1 loss we use its mapping onto Huber with small scale parameter, setting $a = 10^{-3}$.

Our first main result is the analytic characterisation of the generalisation and estimation errors of the ERM estimator.

Theorem 4.1. *For the ERM estimator of the risk function eq. (13) with ℓ_2 regularisation and $\lambda \geq 0$, under the data model defined in eq. (1) with $\beta = 0$ and in the high dimensional limit $n, d \to \infty$ with $n/d = \alpha$ fixed, we have that the excess generalisation error and the estimation error concentrate on*

$$
\begin{aligned}
E_{\text{gen}}^{\text{excess}} &= 1 - \epsilon + q - 2m(1 - \epsilon)\,, \\
E_{\text{estim}} &= 1 + q - 2m\,,
\end{aligned}
\tag{15}
$$

where the values of q and m are the solutions of a system of six self-consistent equations for the unknowns $(m, q, \Sigma, \hat{m}, \hat{q}, \hat{\Sigma})$. Three equations are loss-independent

$$
m = \frac{\hat{m}}{\lambda + \hat{\Sigma}}\,, \quad q = \frac{\hat{m}^2 + \hat{q}}{(\lambda + \hat{\Sigma})^2}\,, \quad \Sigma = \frac{1}{\lambda + \hat{\Sigma}}\,, \tag{16}
$$

while the other three are loss-dependent. For the ℓ_2 loss we have that

$$
\begin{aligned}
\hat{m} &= \frac{\alpha}{1 + \Sigma}(1 - \epsilon)\,, \quad \hat{\Sigma} = \frac{\alpha}{1 + \Sigma}\,, \\
\hat{q} &= \frac{\alpha(1 + q + (1 - \epsilon)\Delta_{IN} + \epsilon\Delta_{OUT} - \epsilon - 2m(1 - \epsilon))}{(1 + \Sigma)^2}\,.
\end{aligned}
\tag{17}
$$

For the ℓ_1 loss and for the Huber loss we have that:

$$
\begin{aligned}
\hat{m} &= \frac{\alpha}{\nu}[(1 - \epsilon)\operatorname{erf}(\chi_{IN})]\,, \\
\hat{\Sigma} &= \frac{\alpha}{\nu}[(1 - \epsilon)\operatorname{erf}(\chi_{IN}) + \epsilon\operatorname{erf}(\chi_{OUT})]\,, \\
\hat{q} &= \frac{\alpha}{\nu^2}\Bigg[(1 - \epsilon)\left(\zeta_{IN} - \mu^2\right)\operatorname{erf}(\chi_{IN}) \\
&\quad + \epsilon\left(\zeta_{OUT} - \mu^2\right)\operatorname{erf}(\chi_{OUT}) + \mu^2 \\
&\quad - \mu\sqrt{\frac{2}{\pi}}\left((1 - \epsilon)\sqrt{\zeta_{IN}}\,e^{-\chi_{IN}^2} + \epsilon\sqrt{\zeta_{OUT}}\,e^{-\chi_{OUT}^2}\right)\Bigg]\,,
\end{aligned}
\tag{18}
$$

where $\zeta_{IN} = \Delta_{IN} - 2m + q + 1$, $\zeta_{OUT} = \Delta_{OUT} + q$, while the other constants depend on the loss. For ℓ_1 loss

$$
\begin{aligned}
\mu^{\ell_1} &= \nu^{\ell_1} = \Sigma\,, \\
\chi_{IN}^{\ell_1} &= \Sigma/\sqrt{2\zeta_{IN}}\,, \quad \chi_{OUT}^{\ell_1} = \Sigma/\sqrt{2\zeta_{OUT}}\,,
\end{aligned}
\tag{19}
$$

and for Huber loss

$$
\begin{aligned}
\nu^{Huber} &= \Sigma + 1\,, \quad \mu^{Huber} = a\nu^{Huber}\,, \\
\chi_{IN}^{Huber} &= \mu^{Huber}/\sqrt{2\zeta_{IN}}\,, \\
\chi_{OUT}^{Huber} &= \mu^{Huber}/\sqrt{2\zeta_{OUT}}\,.
\end{aligned}
\tag{20}
$$

The previous Theorem can be proven following the proof of Theorem 1 in [Loureiro et al., 2022], through a reduction to our case and a computation of loss-specific quantities (namely their proximal operators) that is provided in detail in Appendix A.2. We also provide a similar result for $\beta > 0$. The parameters m and q appearing in the previous statement have a simple interpretation as the values around which the teacher-student and student-student overlaps concentrate in high dimension

$$
m = \frac{1}{d}\boldsymbol{w}^\star \cdot \hat{\boldsymbol{w}}\,, \quad q = \frac{1}{d}\|\hat{\boldsymbol{w}}\|_2^2\,. \tag{21}
$$

In the case of Ridge regression we have that the equations can be solved explicitly and get to an explicit form for the generalisation error. We show this in Appendix A.2.1.

5 CONSISTENCY (AND LACK THEREOF) OF THE GENERALISATION & ESTIMATION ERROR

We now study, for both generalisation and estimation error and at large sample complexity α, whether optimally-regularised ERM estimators achieve BO performance or not, *i.e.* whether ERM is a consistent estimator.

We start by remarking that the large sample complexity asymptotics of the BO generalisation error and of the BO estimation error is

$$
\begin{aligned}
E_{\text{estim}}^{\text{BO}} &= 0 + \mathcal{O}\left(\alpha^{-1}\right)\,, \\
E_{\text{gen}}^{\text{BO}} &= \epsilon(1 - \epsilon)(1 - \beta)^2 + \Delta_{\text{eff}} + \mathcal{O}\left(\alpha^{-1}\right)
\end{aligned}
\tag{22}
$$

which can be found by expanding around $q_b = 1 + c/\alpha + \mathcal{O}\left(\alpha^{-2}\right)$ in eq. (10). We prove the (fast) rate $1/\alpha$ for the errors in Appendix A.3. This rate is a typical feature of BO estimators, see for example [Aubin et al., 2020, Haussler et al., 1991].
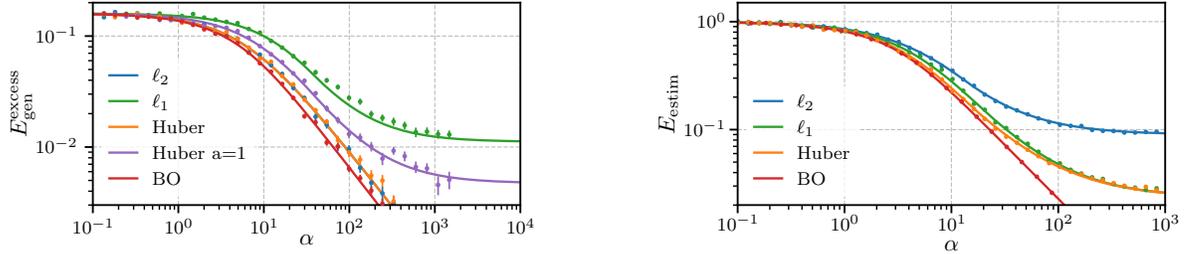
Figure 1: Generalisation and estimation errors as a function of the sample complexity α for $\beta = 0$, in two regimes where ERM do not always achieve BO performance for large α. **(Left)** Here we plot the excess generalisation error for $\epsilon = 0.6$, $\Delta_{\text{IN}} = 1$ and $\Delta_{\text{OUT}} = 0.5$, and for all losses we use the optimal value of λ. We plot two versions of the Huber loss, one with optimal scale parameter a and the other with fixed $a = 1$. We see that only the error of the ℓ_2 and Huber loss with optimal scale a (superimposed in the plot) vanishes at large sample complexity, reaching BO performance. For fixed-scale Huber loss, and ℓ_1 loss, the error converges to a finite value. The parameters used for the simulations (dots) of BO are $d = 1000$ averaged over 50 samples, while for the ERM $d = 200$ averaged over 1000 samples. **(Right)** Here we plot the estimation error for $\epsilon = 0.3$, $\Delta_{\text{IN}} = 1.0$ and $\Delta_{\text{OUT}} = 5.0$, and for all losses we use optimal (λ, a). We see that all losses converge to a finite value of the estimation error, while the BO error goes to zero for large sample complexity. The parameters used for BO are $d = 4000$ averaged over 10 samples, while for the ERM $d = 200$ aveaged over 1000 samples.

We then consider ERM. We find that ERM always estimates correctly the direction of the teacher weights. One could then expect that by adjusting the ℓ_2 regularisation parameter, ERM estimators could also achieve the BO norm for the metric under consideration (generalisation/estimation error). We show that this is not always possible if we restrict $\lambda \geq 0$ — a very natural constraint given that for $\lambda < 0$ the risk function for ℓ_1 and Huber losses becomes non-convex and unbounded from below. This is due to the fact that for some values of the parameters characterising the outliers' distribution, the norm of ERM estimators is shorter than the BO one, requiring *a priori* a negative regularisation to push it to larger norms.

Theorem 5.1. *Under the same conditions of Theorem 4.1, as $\alpha \to \infty$ (taken after the $n, d \to \infty$ limit), we have that*

- *the normalised scalar product between ERM estimators (for the three losses considered) with the teacher weights converges to one, regardless of the value of the regulariser λ and of the scale parameter a of the Huber loss.*
- *ℓ_2 and optimally-scaled Huber ERM estimators achieve consistency with respect to the generalisation error for all fixed values of $\lambda > 0$ with the BO rate $1/\alpha$ for all values of the parameters of the outliers' distribution.*
- *Optimally-regularised ℓ_1 and Huber ERM estimators (with the scale parameter of the Huber loss left fixed, un-optimised) achieve consistency with respect to the generalisation error with the BO rate $1/\alpha$ if and only if $\Delta_{OUT} - \Delta_{IN} \geq (1 - \beta)^2 (2\epsilon - 1)$ for $0 < \beta < 1$, and $\Delta_{OUT} - \Delta_{IN} \leq (1 - \beta)^2 (2\epsilon - 1)$ for $\beta > 1$.*
- *Optimally-regularised ERM estimators (for the three*

losses considered) achieve consistency with respect to the estimation error with the BO rate $1/\alpha$ if and only if $\beta > 1$.

We prove Theorem 5.1 in Appendix A.4 by expanding at large α the self-consistent equations presented in Theorem 4.1. Thus, we proved that there exists some regions of parameters in which, for both generalisation and estimation error, ℓ_2 regularisation is not sufficient to reach BO performance. In those regimes, consistency for the generalisation error can always be recovered by cross-validating on a test set to find the best norm, given that the angle between teacher and student vanishes. Due to homogeneity properties of the three losses considered, the rescaling can also be performed before training on the labels, as discussed in Appendix A.4.6. The estimation error can be calibrated similarly, but we stress that such a calibration requires either having access to an oracle knowledge of the BO norm, or to a clean, unpolluted dataset to cross-validate for the optimal norm. In the absence of such an ideal hold-out dataset, the Huber and ℓ_1 methods do *not* lead to consistent estimates of the objective functions, and robust estimation is biased.

In Figure 1 we show, for both the generalisation and estimation error, the dependency of the error on the sample complexity α in two regimes were some of the losses do not achieve consistency. It is worth noticing that consistency could be achieved by carefully allowing for negative ℓ_2 regularisation. In the case of the ℓ_2 loss, a negative regularisation does not immediately break the convexity of the risk function. Indeed, the risk function is convex up to $\lambda = -(1 - \sqrt{\alpha})^2$ (see Appendix A.5). Thus, for ℓ_2 loss consistency for

Matteo Vilucchio, Emanuele Troiani, Vittorio Erba, Florent Krzakala
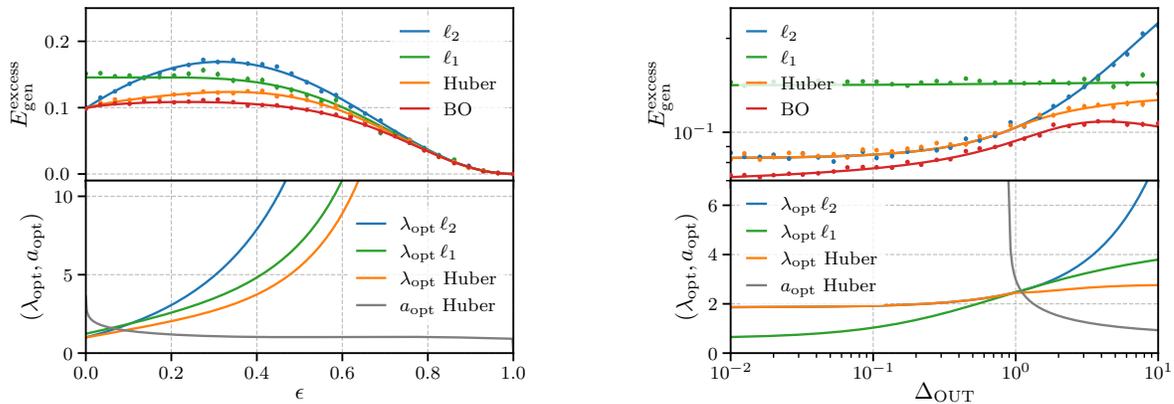
Figure 2: Excess generalisation error (top panels) as a function of the percentage of outliers' ϵ (left) and of the outliers' variance Δ_{OUT} (right), along with the respective optimal hyper-parameters (λ, a) (bottom panels). In both plots $\alpha = 10$, $\beta = 0$, $\Delta_{\text{IN}} = 1$, and for all numerical simulations (dots) have $d = 200$, and are averaged over 1000 samples. **(Left)** In this plot $\Delta_{\text{OUT}} = 5$. We observe that for small ϵ both the ℓ_2 and the Huber loss converge to BO performance, while ℓ_1 does not. **(Right)** In this plot $\epsilon = 0.3$. For $\Delta_{\text{OUT}} < \Delta_{\text{IN}}$ we notice that the Huber scale parameter diverges, so the Huber and ℓ_2 loss become identical. Furthermore, the BO estimator has a non-monotonic behaviour as a function of Δ_{OUT}, while ERM estimators are monotonic.

the estimation error can be achieved without using an oracle norm up to (in the large sample complexity limit) $\beta > 1 - \epsilon^{-1}$, by allowing for negative lambda and retaining convexity of the risk function. On the other hand, for ℓ_1 and Huber loss, even an arbitrarily small negative regularisation implies a non-convex and unbounded-from-below risk function. It is reasonable to think though that an optimisation procedure restricted to small enough norms would still result in a well-defined ERM procedure, allowing for negative regularisation also in these case. We leave a more thorough exploration of this effect for future work. We also mention that negative regularisation has been considered in a series of recent works [Wu and Xu, 2020, Kobak et al., 2022, Hastie et al., 2022].

6 DEPENDENCY OF THE GENERALISATION ON THE PARAMETERS

We now investigate the dependence of on the parameters characterising the amount of corruption caused by outliers. As in the previous section, we plot the excess generalisation error, and we always consider optimally tuned hyperparameters λ and a. Again, we discuss the phenomenology arising for a particular value for the parameters of the model, but we stress that the phenomenology is qualitatively the same for other values of the parameters, see Appendix D for additional experiments, as well as Appendix C for a more realistic scenario.

Dependency of the generalisation error on the amount of outliers. We investigate the influence of the percentage of outliers ϵ on the generalisation error in Figure 2 left panel. We identify two qualitatively different regimes where the behaviour is different: the regime with small ϵ where outliers are rare and vice-versa the regime with ϵ close to 1, where the samples that would be considered outliers in the $\epsilon \to 0$ limit dominate the dataset.

We observe that in the small ϵ regime, for all losses, the ERM excess generalisation error increases as the percentage of outliers increases. In the limit of small ϵ, the Huber loss performs similarly to ℓ_2, which is known to be Bayes optimal at $\epsilon = 0$ if optimally regularised [Cover and Thomas, 1991]. In contrast, the ℓ_1 loss converges to a higher value for the generalisation error not reaching BO performances. For all finite values ϵ, ERMs fails to reach BO performance for all three choices of the loss we studied. By expanding the explicit asymptotic form for the generalisation error of ERM with ℓ_2 loss presented in Corollary A.1, we are able to provide an analytical characterisation of the small ϵ behaviour of ℓ_2 ERM.

Theorem 6.1. *As $\epsilon \to 0$ (taken after the $n, d \to \infty$ limit), we have that the generalisation error of optimally-regularised ℓ_2 loss and the optimal regularisation satisfy*

$$E_{\text{gen}}^{\ell_2} = E_{\text{gen}}^{\text{BO}} + \kappa \epsilon + \mathcal{O}(\epsilon^2), \quad \lambda = \Delta_{IN} + \lambda_1 \epsilon + \mathcal{O}(\epsilon^2) \tag{23}$$

where $E_{\text{gen}}^{\text{BO}}$ is the Bayes optimal generalisation error, and κ and λ_1 are functions of the model parameters that we provide in Appendix A.6.
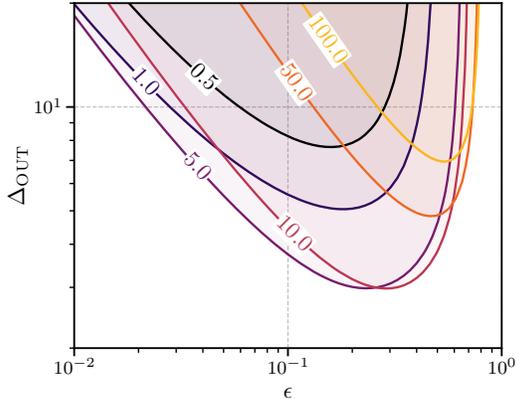
Figure 3: Difference between the generalisation error of optimally-regularised ℓ_2 and Huber loss losses as a function of ϵ and Δ_{OUT} for $\Delta_{\mathrm{IN}} = 1$, $\beta = 0$. Each shaded region denotes where the two losses have a non-zero difference, *i.e.* where it is better to use the Huber loss, for different values of α as indicated by the labels. Focusing on the region of small ϵ and large Δ_{OUT}, *i.e.* when dealing with just a few extremely noisy outliers, the plot suggests that having many samples (large α) makes using the Huber loss not advantageous.

In the $\epsilon \gtrsim 0.5$ regime, where outliers are dominant, we observe a decrease in the excess generalisation error, which eventually vanishes as ϵ approaches one. In this limit the majority of the dataset is composed of data generated from a β-rescaled teacher, so that ERM will be biased to smaller ($0 < \beta < 1$) or larger ($\beta > 1$) norms than if trained on a dataset with no outliers. As discussed previously, the optimal ℓ_2 regularisation will counterbalance as much as possible this effect, aiming at the BO norm. In the extreme case where $\beta = 0$, outlier samples provide no meaningful information on the ground truth and the BO estimator equals the null vector $\boldsymbol{w} = 0$. This is compatible with what we observe in the figure, where λ_{opt} diverges as ϵ approaches one.

Dependency of the generalisation error on the variance of outliers. In Figure 2, right panel, we consider the dependency of the generalisation error as a function of the variance of the outliers Δ_{OUT}. We start by noticing that all ERM estimators have a generalisation error that grows monotonically with Δ_{OUT}, while the BO generalisation error has a non-monotonic behaviour, with a local maximum. This behaviour is also present for $\beta > 0$. Moreover, we notice that for large outliers' variance the ℓ_2 loss performs worse than the more resilient ℓ_1 and Huber losses, as expected.

Performance difference between Huber and ℓ_2 loss We observe an interesting behaviour in compar-

ing the Huber loss and the ℓ_2 loss. Observing Figure 2, right panel, we observe that as Δ_{OUT} decreases the scale parameter value a_{opt} of the Huber loss diverges to infinity at a finite value of Δ_{OUT}. When $a \to +\infty$ the Huber loss converges to the ℓ_2 loss, as confirmed by the fact that their predicted generalisation errors coincides (Figure 2 right). It is remarkable that optimally-regularised Huber loss is not always strictly better than the simpler ℓ_2 loss in a whole region of model parameters.

We explore further this phase transition by plotting the difference between the generalisation error of the optimal ℓ_2 and Huber losses as a function of both ϵ and Δ_{OUT} in Figure 3. We remark that the two plots in Figure 2 correspond respectively to horizontal and vertical slices in this plot. Also, we see that the difference in performance between the two losses is relevant just in the region of large Δ_{OUT} and intermediate ϵ. Moreover, focusing on the region of small ϵ, large Δ_{OUT} and large α, *i.e.* for really rare and noisy outliers, Figure 3 suggests that the more samples one has (larger α), the less convenient it becomes to use the Huber loss, at fixed value of all other parameters.

7 CONCLUDING REMARKS

We studied robust ERM in a model of data with outliers, focusing on three commonly used losses (ℓ_2-regularised ℓ_2, ℓ_1, and Huber loss). The percentage of outliers, variance, and typical norm can be varied, allowing for a thorough exploration of the effect of outliers on the regression task. We derived analytically exact asymptotics for the generalisation and estimation errors of ERM estimators trained on this data model in the high-dimensional limit, where the number of samples n and the dimension d are large with a fixed finite ratio, and we compared them with BO estimators.

For the generalisation error, we found that optimally regularised ERM is asymptotically consistent in the large sample complexity limit at the condition of performing a norm calibration. For the estimation error, we showed that due to a norm calibration mismatch, the consistency of the estimator requires an oracle estimate of the optimal norm or the presence of a cross-validation set not corrupted by the outliers. We also observed an interesting phase transition as a function of the outlier parameters. In a full region of the parameter space ($\epsilon, \Delta_{\mathrm{OUT}}$), the optimal scale parameter of the Huber loss a_{out} diverges, and the error of the ERM with Huber and ℓ_2 losses coincides.

Our work opens the way to the study of more complicated settings, for instance random features regression [Rahimi and Recht, 2007] or deep learning in the lazy/neural tangent kernel regime [Chizat et al., 2019],

where one could study effects that are not directly explorable in the vanilla generalised linear regression setting, such as over-parametrization [Belkin et al., 2020], benign overfitting [Bartlett et al., 2020], implicit regularisation [Gunasekar et al., 2017], and their interplay with outliers.

Acknowledgements

References

[Aubin et al., 2020] Aubin, B., Krzakala, F., Lu, Y., and Zdeborová, L. (2020). Generalization error in high-dimensional perceptrons: Approaching bayes error with convex optimization. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H., editors, *Advances in Neural Information Processing Systems*, volume 33, pages 12199–12210. Curran Associates, Inc.

[Barbier et al., 2019] Barbier, J., Krzakala, F., Macris, N., Miolane, L., and Zdeborová, L. (2019). Optimal errors and phase transitions in high-dimensional generalized linear models. *Proceedings of the National Academy of Sciences*, 116(12):5451–5460.

[Bartlett et al., 2020] Bartlett, P. L., Long, P. M., Lugosi, G., and Tsigler, A. (2020). Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, 117(48):30063–30070.

[Bean et al., 2013] Bean, D., Bickel, P. J., Karoui, N. E., and Yu, B. (2013). Optimal m-estimation in high-dimensional regression. *Proceedings of the National Academy of Sciences of the United States of America*, 110(36):14563–14568.

[Belkin et al., 2020] Belkin, M., Hsu, D., and Xu, J. (2020). Two models of double descent for weak features. *SIAM Journal on Mathematics of Data Science*, 2(4):1167–1180.

[Bellec, 2020] Bellec, P. C. (2020). Out-of-sample error estimate for robust m-estimators with convex penalty.

[Bellec et al., 2022] Bellec, P. C., Shen, Y., and Zhang, C.-H. (2022). Asymptotic normality of robust M-estimators with convex penalty. *Electronic Journal of Statistics*, 16(2):5591 – 5622.

[Berger and Berliner, 1986] Berger, J. and Berliner, L. M. (1986). Robust bayes and empirical bayes analysis with ϵ-contaminated priors. *The Annals of Statistics*, 14(2):461–486.

[Bhatia et al., 2017] Bhatia, K., Jain, P., Kamalaruban, P., and Kar, P. (2017). Consistent robust regression. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

[Box and Tiao, 1968] Box, G. E. P. and Tiao, G. C. (1968). A bayesian approach to some outlier problems. *Biometrika*, 55(1):119–129.

[Chizat et al., 2019] Chizat, L., Oyallon, E., and Bach, F. (2019). On lazy training in differentiable programming. *Advances in neural information processing systems*, 32.

[Cover and Thomas, 1991] Cover, T. M. and Thomas, J. A. (1991). Information theory and statistics. *Elements of Information Theory*, 1(1):279–335.

[Dalalyan and Thompson, 2019] Dalalyan, A. S. and Thompson, P. (2019). Outlier-robust estimation of a sparse linear model using l1-penalized huber's m-estimator. In *Neural Information Processing Systems*.

[Donoho and Montanari, 2016] Donoho, D. and Montanari, A. (2016). High dimensional robust m-estimation: asymptotic variance via approximate message passing. *Probability Theory and Related Fields*, 166(3):935–969.

[Gagnon et al., 2020] Gagnon, P., Desgagné, A., and Bédard, M. (2020). A New Bayesian Approach to Robustness Against Outliers in Linear Regression. *Bayesian Analysis*, 15(2):389 – 414.

[Gunasekar et al., 2017] Gunasekar, S., Woodworth, B. E., Bhojanapalli, S., Neyshabur, B., and Srebro, N. (2017). Implicit regularization in matrix factorization. *Advances in Neural Information Processing Systems*, 30.

[Hampel, 1968] Hampel, F. (1968). Contributions to the theory of robust estimation [ph. d. thesis]. *University of California, Berkeley, Calif, USA*.

[Hampel, 1971] Hampel, F. R. (1971). A general qualitative definition of robustness. *The Annals of mathematical statistics the official journal of the Institute of Mathematical Statistics*, 42(6).

[Hastie et al., 2022] Hastie, T., Montanari, A., Rosset, S., and Tibshirani, R. J. (2022). Surprises in high-dimensional ridgeless least squares interpolation. *The Annals of Statistics*, 50(2):949 – 986.

[Haussler et al., 1991] Haussler, D., Kearns, M., Opper, M., and Schapire, R. (1991). Estimating average-case learning curves using bayesian, statistical physics and vc dimension methods. *Advances in Neural Information Processing Systems*, 4.

[Hu and Lu, 2023] Hu, H. and Lu, Y. M. (2023). Universality laws for high-dimensional learning with random features. *IEEE Transactions on Information Theory*, 69(3):1932–1964.

[Huang, 2020] Huang, H. (2020). Asymptotic risk and phase transition of l_1-penalized robust estimator. *The Annals of Statistics*, 48(5):3090 – 3111.

[Huber, 1964] Huber, P. J. (1964). Robust Estimation of a Location Parameter. *The Annals of Mathematical Statistics*, 35(1):73 – 101.

[Huber, 1965] Huber, P. J. (1965). A Robust Version of the Probability Ratio Test. *The Annals of Mathematical Statistics*, 36(6):1753 – 1758.

[Huber, 2011] Huber, P. J. (2011). *Robust Statistics*, pages 1248–1251. Springer Berlin Heidelberg, Berlin, Heidelberg.

[Kabashima and Uda, 2004] Kabashima, Y. and Uda, S. (2004). A bp-based algorithm for performing bayesian inference in large perceptron-type networks. In Ben-David, S., Case, J., and Maruoka, A., editors, *Algorithmic Learning Theory*, pages 479–493, Berlin, Heidelberg. Springer Berlin Heidelberg.

[Karoui et al., 2013] Karoui, N. E., Bean, D., Bickel, P. J., Lim, C., and Yu, B. (2013). On robust regression with high-dimensional predictors. *Proceedings of the National Academy of Sciences*, 110(36):14557–14562.

[Kobak et al., 2022] Kobak, D., Lomond, J., and Sanchez, B. (2022). The optimal ridge penalty for real-world high-dimensional data can be zero or negative due to the implicit ridge regularization. *J. Mach. Learn. Res.*, 21(1).

[Lecun et al., 1998] Lecun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324.

[Liu and Nocedal, 1989] Liu, D. C. and Nocedal, J. (1989). On the limited memory bfgs method for large scale optimization. *Mathematical Programming*, 45(1):503–528.

[Liu et al., 2019] Liu, L., Li, T., and Caramanis, C. (2019). High dimensional robust m-estimation: Arbitrary corruption and heavy tails.

[Loureiro et al., 2022] Loureiro, B., Gerbelot, C., Cui, H., Goldt, S., Krzakala, F., Mézard, M., and Zdeborová, L. (2022). Learning curves of generic features maps for realistic datasets with a teacher-student model*. *Journal of Statistical Mechanics: Theory and Experiment*, 2022(11):114001.

[Pedregosa et al., 2011] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

[Raghavendra and Yau, 2020] Raghavendra, P. and Yau, M. (2020). List decodable learning via sum of squares. In *Proceedings of the 2020 ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 161–180.

[Rahimi and Recht, 2007] Rahimi, A. and Recht, B. (2007). Random features for large-scale kernel machines. *Advances in neural information processing systems*, 20.

[Rangan, 2011] Rangan, S. (2011). Generalized approximate message passing for estimation with random linear mixing. In *2011 IEEE International Symposium on Information Theory Proceedings*, pages 2168–2172. IEEE.

[Suggala et al., 2019] Suggala, A. S., Bhatia, K., Ravikumar, P., and Jain, P. (2019). Adaptive hard thresholding for near-optimal consistent robust regression. *CoRR*, abs/1903.08192.

[Thrampoulidis et al., 2018] Thrampoulidis, C., Abbasi, E., and Hassibi, B. (2018). Precise error analysis of regularized m-estimators in high dimensions. *IEEE Transactions on Information Theory*, 64(8):5592–5628.

[Tsakonas et al., 2014] Tsakonas, E., Jaldén, J., Sidiropoulos, N. D., and Ottersten, B. (2014). Convergence of the huber regression m-estimate in the presence of dense outliers. *IEEE Signal Processing Letters*, 21(10):1211–1214.

[Wu and Xu, 2020] Wu, D. and Xu, J. (2020). On the optimal weighted ℓ_2 regularization in overparameterized linear regression. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H., editors, *Advances in Neural Information Processing Systems*, volume 33, pages 10112–10123. Curran Associates, Inc.

[Yuen and Ortiz, 2017] Yuen, K.-V. and Ortiz, G. A. (2017). Outlier detection and robust regression for correlated data. *Computer Methods in Applied Mechanics and Engineering*, 313:632–646.

Checklist

1. For all models and algorithms presented, check if you include:

   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]

   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Not Applicable]

   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes]

2. For any theoretical claim, check if you include:

   (a) Statements of the full set of assumptions of all theoretical results. [Yes]

   (b) Complete proofs of all theoretical results. [Yes]

   (c) Clear explanations of any assumptions. [Yes]

3. For all figures and tables that present empirical results, check if you include:

   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes]

   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes]

   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes]

   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Not Applicable]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:

   (a) Citations of the creator If your work uses existing assets. [Yes]

   (b) The license information of the assets, if applicable. [Yes]

   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]

   (d) Information about consent from data providers/curators. [Not Applicable]

   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:

   (a) The full text of instructions given to participants and screenshots. [Not Applicable]

   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]

   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

A   Mathematical details

A.1   Claim: Bayes optimal estimator for the generalisation error

Define the generalisation error as in eq. (2)

$$E_{\text{gen}}(\boldsymbol{w}) = \mathbb{E}_{(\bar{\boldsymbol{x}}, \bar{y}|\boldsymbol{w}^\star)}\left[\left(\bar{y} - \frac{\bar{\boldsymbol{x}} \cdot \boldsymbol{w}}{\sqrt{d}}\right)^2\right], \tag{24}$$

where the average is over a sample-label pair $(\bar{\boldsymbol{x}}, \bar{y})$ generated with the same distribution as the training set, namely with the same teacher vector $\boldsymbol{w}^\star$. The posterior-averaged generalisation error is

$$\mathcal{R}(\boldsymbol{w}) = \mathbb{E}_{\boldsymbol{w}^\star|\mathcal{D}}\left[\mathbb{E}_{(\bar{\boldsymbol{x}}, \bar{y})|\boldsymbol{w}^\star}\left[\left(\bar{y} - \frac{\bar{\boldsymbol{x}} \cdot \boldsymbol{w}}{\sqrt{d}}\right)^2\right]\right] \tag{25}$$

where $\mathcal{D}$ is the observed training dataset, $\mathbb{E}_{\boldsymbol{w}^\star|\mathcal{D}}[\cdot]$ is the posterior average and $(\bar{\boldsymbol{x}}, \bar{y})$ is a sample label pair. Differentiating and setting the derivative to zero (in order to minimise the posterior-averaged generalisation error) gives that the minimiser $\hat{\boldsymbol{w}}$, *i.e.* the BO estimator, satisfies

$$\hat{\boldsymbol{w}}_{\text{gen}}^{\text{BO}}(\mathcal{D}) = \mathbb{E}_{\boldsymbol{w}|\mathcal{D}}\left[\mathbb{E}_{\bar{\boldsymbol{x}}, \bar{y}|\boldsymbol{w}}[\bar{y}\,\bar{\boldsymbol{x}}]\right], \tag{26}$$

which for our specific choice of data model reduces to

$$\hat{\boldsymbol{w}}_{\text{gen}}^{\text{BO}}(\mathcal{D}) = (1 - \epsilon + \beta\epsilon)\mathbb{E}_{\boldsymbol{w}|\mathcal{D}}[\boldsymbol{w}]. \tag{27}$$

A.2   Theorem 4.1: self-consistent equations for the errors of ERM

The more general proof of the form of the self-consistent equations can be found in [Loureiro et al., 2022]. Our case corresponds to taking $\boldsymbol{\varphi}_s(\boldsymbol{x}) = \boldsymbol{\varphi}_t(\boldsymbol{x}) = \boldsymbol{x}$ which is equivalent to the setting presented in [Aubin et al., 2020], of which we will follow the notation.

In the following we will refer to the loss function as $\mathcal{L}(\cdot, \cdot)$ and to the regularisation function as $r(\cdot)$. We then define the general quantities that will appear in the fixed point equations, they are:

$$
\begin{aligned}
\mathcal{Z}_{\text{out}^\star}(y, \omega, V) &= \mathbb{E}_{z \sim \mathcal{N}(\omega, V)}\left[P_{\text{out}}(y|z)\right], \\
f_{\text{out}^\star}(y, \omega, V) &= \partial_\omega \log(\mathcal{Z}_{\text{out}^\star}(y, \omega, V)), \\
f_{\text{out}}(y, \omega, V) &= (\mathcal{P}_V[\mathcal{L}(y, \cdot)](\omega) - \omega)/V, \\
\mathcal{Z}_{\text{w}^\star}(\gamma, \Lambda) &= \mathbb{E}_{\omega \sim P_{\text{w}^\star}}\left[e^{-\frac{1}{2}\Lambda\omega^2 + \gamma\omega}\right], \\
f_{\text{w}^\star}(\gamma, \Lambda) &= \partial_\gamma \log \mathcal{Z}_{\text{w}^\star}(\gamma, \Lambda), \\
f_{\text{w}}(\gamma, \Lambda) &= \mathcal{P}_{\Lambda^{-1}}[r(.)]\left(\Lambda^{-1}\gamma\right),
\end{aligned}
\tag{28}
$$

where ω in the definition of $\mathcal{Z}_{\text{w}^\star}(\gamma, \Lambda)$ is sampled with the teacher weight distribution and $\mathcal{P}_V[f](x)$ is the proximal operator defined as follows

$$\mathcal{P}_V[f(y, \cdot)](x) = \arg\min_z\left[f(y, z) + \frac{1}{2V}(z - x)^2\right]. \tag{29}$$

The self-consistent equations in general form then read

$$
\begin{aligned}
m &= \mathbb{E}_\xi \left[\mathcal{Z}_{\mathrm{w}^\star}(\sqrt{\hat{\eta}}\xi, \hat{\eta}) f_{\mathrm{w}^\star}(\sqrt{\hat{\eta}}\xi, \hat{\eta}) f_{\mathrm{w}}\left(\hat{q}^{1/2}\xi, \hat{\Sigma}\right) \right], \\
q &= \mathbb{E}_\xi \left[\mathcal{Z}_{\mathrm{w}^\star}(\sqrt{\hat{\eta}}\xi, \hat{\eta}) f_{\mathrm{w}}\left(\hat{q}^{1/2}\xi, \hat{\Sigma}\right)^2 \right], \\
\Sigma &= \mathbb{E}_\xi \left[\mathcal{Z}_{\mathrm{w}^\star}(\sqrt{\hat{\eta}}\xi, \hat{\eta}) \partial_\gamma f_{\mathrm{w}}\left(\hat{q}^{1/2}\xi, \hat{\Sigma}\right) \right], \\
\hat{m} &= \alpha \int \mathbb{E}_\xi \left[\mathcal{Z}_{\mathrm{out}^\star}(y, \sqrt{\eta}\xi, (1-\eta)) \cdot f_{\mathrm{out}^\star}(y, \sqrt{\eta}\xi, (1-\eta)) f_{\mathrm{out}}\left(y, q^{1/2}\xi, \Sigma\right) \right] dy, \\
\hat{q} &= \alpha \int \mathbb{E}_\xi \left[\mathcal{Z}_{\mathrm{out}^\star}(y, \sqrt{\eta}\xi, (1-\eta)) f_{\mathrm{out}}\left(y, q^{1/2}\xi, \Sigma\right)^2 \right] dy, \\
\hat{\Sigma} &= -\alpha \int \mathbb{E}_\xi \left[\mathcal{Z}_{\mathrm{out}^\star}(y, \sqrt{\eta}\xi, (1-\eta)) \partial_\omega f_{\mathrm{out}}\left(y, q^{1/2}\xi, \Sigma\right) \right] dy.
\end{aligned}
\tag{30}
$$

where we have denoted with $\mathbb{E}_\xi$ the expectation value over a standard Gaussian variable $\xi \sim \mathcal{N}(0,1)$, and we defined $\eta = m^2/q$ and $\hat{\eta} = \hat{m}^2/\hat{q}$.

In the main text we fixed the regularisation function to the ℓ_2 regularisation, *i.e.* $r(\mathbf{w}) = \frac{1}{2}\|\mathbf{w}\|_2^2$ and the teacher weights to be extracted from the distribution $\boldsymbol{w}^\star \sim \mathcal{N}(0, \mathbb{1}_d)$. The quantities $\mathcal{Z}_{\mathrm{w}^\star}$ and $f_{\mathrm{w}^\star}$ can be written as

$$
\mathcal{Z}_{\mathrm{w}^*}(\gamma, \Lambda) = \frac{e^{\frac{\gamma^2}{2(\Lambda+1)}}}{\sqrt{\Lambda+1}}, \quad f_{\mathrm{w}^*}(\gamma, \Lambda) = \frac{\gamma}{\Lambda+1}.
\tag{31}
$$

Also we have that

$$
\mathcal{Z}_{\mathrm{w}}^{\ell_2, \lambda}(\gamma, \Lambda) = \frac{e^{\frac{\gamma^2}{2(\lambda+\Lambda)}}}{\sqrt{\lambda+\Lambda}}, \quad f_{\mathrm{w}}^{\ell_2, \lambda}(\gamma, \Lambda) = \frac{\gamma}{\lambda+\Lambda}.
\tag{32}
$$

This allows to write the first three self consistent equations as:

$$
m = \frac{\hat{m}}{\lambda+\hat{\Sigma}}, \quad q = \frac{\hat{m}^2+\hat{q}}{(\lambda+\hat{\Sigma})^2}, \quad \Sigma = \frac{1}{\lambda+\hat{\Sigma}}.
\tag{33}
$$

As explained in [Aubin et al., 2020] these first three equations only depend on the regularisation term and they will be thus fixed during our analysis.

Thus, to compute the performance of ERM with a given loss, we just need to compute $\mathcal{Z}_{\mathrm{out}^\star}, f_{\mathrm{out}^\star}$ and f_{out}, and solve the corresponding system of self-consistent equations. For our noise model P_{out} we have that

$$
\mathcal{Z}_{\mathrm{out}^\star}(y, \omega, V) = \mathcal{N}_y(\omega, V + \Delta_{\mathrm{IN}}) + \mathcal{N}_y(\beta\omega, V + \beta^2 \Delta_{\mathrm{OUT}}),
\tag{34}
$$

and $f_{\mathrm{out}^\star}$ is given from the definition in eq. (28).

In the following subsections we specialise these equations to the ℓ_2, ℓ_1 and Huber losses by computing f_{out}, which is the only loss-dependent part in the equations.

ℓ_2 **loss case.** In this case, the f_{out} in eq. (28) reads

$$
f_{\mathrm{out}}^{\ell_2}(y, \omega, V) = \frac{y-\omega}{1+V}.
\tag{35}
$$

Then, the remaining integrals can be solved explicitly as they are all expectation of Gaussian variables, giving

$$
\begin{aligned}
m &= \frac{\hat{m}}{\lambda+\hat{\Sigma}}, \quad q = \frac{\hat{m}^2+\hat{q}}{(\lambda+\hat{\Sigma})^2}, \quad \Sigma = \frac{1}{\lambda+\hat{\Sigma}}, \\
\hat{m} &= \frac{\alpha}{1+\Sigma}(1+\epsilon(\beta-1)), \quad \hat{\Sigma} = \frac{\alpha}{1+\Sigma}, \\
\hat{q} &= \frac{\alpha}{(1+\Sigma)^2}\left[1+q+\Delta_{\mathrm{eff}}+\epsilon(\beta^2-1)-2m(1+\epsilon(\beta-1))\right].
\end{aligned}
\tag{36}
$$

For this specific case we have that one can find an explicit form for the order parameters which is presented in Appendix A.2.1.

Asymptotic Characterisation of the Performance of Robust Linear Regression in the Presence of Outliers

Useful Integrals. To perform the double integration over (y, ξ) for the last three equations in eq. (30) we will use a linear change of variables. The relevant change of variables is $(u, v) = (y - \beta\sqrt{\eta}\xi, y - \sqrt{q}\xi)$. The associated Jacobian is constant and equal to $\left|\beta\sqrt{\eta} - \sqrt{q}\right|$. With this change of variable, we obtain explicit expressions for the following integrals, that will be used below. The first integral integral is

$$\mathcal{I}_1(\Delta, \beta, \kappa, \Lambda, a, m, q) = \int_{\mathbb{R}^2} \mathrm{d}\xi\, \mathrm{d}y\, e^{-\frac{1}{2}\xi^2 - \frac{1}{2}\frac{(\beta\sqrt{\eta}\xi - y)^2}{\Delta + \beta^2(1-\eta)}} \left(\beta\sqrt{\eta}\xi - y\right) \begin{cases} a & \text{if } y - \sqrt{q}\xi > \kappa \\ \frac{y - \sqrt{q}\xi}{\Lambda} & \text{if } \left|y - \sqrt{q}\xi\right| < \kappa \\ -a & \text{if } y - \sqrt{q}\xi < -\kappa \end{cases}$$

$$= \frac{2\pi}{\Lambda}\left((1-\eta)\beta^2 + \Delta\right)^{\frac{3}{2}} \left[\frac{\sqrt{2}(\kappa - a\Lambda)e^{-\frac{\kappa^2}{2(\beta^2 + \Delta - 2\beta m + q)}}}{\sqrt{\pi(q - 2m\beta + \beta^2 + \Delta)}} - \mathrm{erf}\left(\kappa/\sqrt{2(\beta^2 + \Delta - 2\beta m + q)}\right)\right]. \tag{37}$$

The second integral is:

$$\mathcal{I}_2(\Delta, \beta, \kappa, \Lambda, a, m, q) = \int_{\mathbb{R}^2} \mathrm{d}\xi\, \mathrm{d}y\, e^{-\frac{1}{2}\xi^2 - \frac{1}{2}\frac{(\beta\sqrt{\eta}\xi - y)^2}{\Delta + \beta^2(1-\eta)}} \begin{cases} \left(\frac{y - \sqrt{q}\xi}{\Lambda}\right)^2 & \text{if } \left|y - \sqrt{q}\xi\right| < \kappa \\ a^2 & \text{otherwise} \end{cases}$$

$$= \frac{2\pi}{\Lambda^2}\sqrt{(1-\eta)\beta^2 + \Delta}\left[a^2 - \kappa\sqrt{2(q - 2m\beta + \beta^2 + \Delta)/\pi}\, e^{-\frac{\kappa^2}{2(\beta^2 + \Delta - 2\beta m + q)}}\right. \tag{38}$$

$$\left. + \left(q - 2m\beta + \beta^2 + \Delta - a^2\right)\mathrm{erf}\left(\kappa/\sqrt{2(\beta^2 + \Delta - 2\beta m + q)}\right)\right].$$

The third integral is

$$\mathcal{I}_3(\Delta, \beta, \kappa, \Lambda, a, m, q) = \int_{\mathbb{R}^2} \mathrm{d}\xi\, \mathrm{d}y\, e^{-\frac{1}{2}\xi^2 - \frac{1}{2}\frac{(y - \beta\sqrt{\eta}\xi)^2}{\Delta + \beta^2(1-\eta)}} \mathbb{I}_{\left|y - \sqrt{q}\xi\right| < \kappa}$$

$$= 2\pi\sqrt{(1-\eta)\beta^2 + \Delta}\, \mathrm{erf}\left(\kappa/\sqrt{2(\beta^2 + \Delta - 2\beta m + q)}\right)/\Lambda, \tag{39}$$

Using these integrals we can write explicitly the self-consistent equations, leading to the equations presented in Theorem 4.1.

ℓ_1 **loss case.** For the ℓ_1 loss, the f_{out} in eq. (28) reads

$$f_{\mathrm{out}}^{\ell_1}(y, \omega, V) = \begin{cases} \frac{1}{V}(y - \omega) & \text{if } |\omega - y| \leq V \\ \mathrm{sign}(\omega - y) & \text{otherwise} \end{cases}. \tag{40}$$

With the integration technique explained above we have that the explicit form of the self-consistent equations is

$$\hat{m} = \frac{\alpha}{\Sigma}\left[(1-\epsilon)\,\mathrm{erf}\left(\chi_{\mathrm{IN}}\right) + \beta\epsilon\,\mathrm{erf}\left(\chi_{\mathrm{OUT}}\right)\right],$$

$$\hat{q} = \frac{\alpha}{\Sigma^2}\left[(1-\epsilon)\left(\zeta_{\mathrm{IN}} - \mu^2\right)\mathrm{erf}\left(\chi_{\mathrm{IN}}\right) + \epsilon\left(\zeta_{\mathrm{OUT}} - \mu^2\right)\mathrm{erf}\left(\chi_{\mathrm{OUT}}\right)\right.$$

$$\left. - \mu\sqrt{\frac{2}{\pi}}\left((1-\epsilon)\sqrt{\zeta_{\mathrm{IN}}}\, e^{-\chi_{\mathrm{IN}}^2} + \epsilon\sqrt{\zeta_{\mathrm{OUT}}}\, e^{-\chi_{\mathrm{OUT}}^2}\right) + \mu^2\right], \tag{41}$$

$$\hat{\Sigma} = \frac{\alpha}{\Sigma}\left[(1-\epsilon)\,\mathrm{erf}\left(\chi_{\mathrm{IN}}\right) + \epsilon\,\mathrm{erf}\left(\chi_{\mathrm{OUT}}\right)\right],$$

where, as in the main text, we defined $\zeta_{\mathrm{IN}} = \Delta_{\mathrm{IN}} - 2m + q + 1$, $\zeta_{\mathrm{OUT}} = \Delta_{\mathrm{OUT}} + \beta^2 + q - 2\beta m$ and

$$\mu^{\ell_1} = \nu^{\ell_1} = \Sigma, \quad \chi_{\mathrm{IN}}^{\ell_1} = \Sigma/\sqrt{2\zeta_{\mathrm{IN}}}, \quad \chi_{\mathrm{OUT}}^{\ell_1} = \Sigma/\sqrt{2\zeta_{\mathrm{OUT}}}, \tag{42}$$

Huber loss case. For the Huber loss, the f_{out} in eq. (28) reads

$$f_{\text{out}}^{\text{Huber}}(y, \omega, V) = \begin{cases} -a & \text{if } y - \omega > a(V+1) \\ \frac{y-\omega}{V+1} & \text{if } |y - \omega| < a(V+1) \\ a & \text{otherwise} \end{cases} . \tag{43}$$

we have that the integrals in eq. (30) become

$$\hat{m} = \frac{\alpha}{\Sigma + 1}[(1 - \epsilon)\operatorname{erf}(\chi_{\text{IN}}) + \beta\epsilon\operatorname{erf}(\chi_{\text{OUT}})],$$

$$\hat{q} = \frac{\alpha}{(\Sigma + 1)^2}\left[(1 - \epsilon)\left(\zeta_{\text{IN}} - \mu^2\right)\operatorname{erf}(\chi_{\text{IN}}) + \epsilon\left(\zeta_{\text{OUT}} - \mu^2\right)\operatorname{erf}(\chi_{\text{OUT}})\right.$$
$$\left. - \mu\sqrt{\frac{2}{\pi}}\left((1 - \epsilon)\sqrt{\zeta_{\text{IN}}}\,e^{-\chi_{\text{IN}}^2} + \epsilon\sqrt{\zeta_{\text{OUT}}}\,e^{-\chi_{\text{OUT}}^2}\right) + \mu^2\right], \tag{44}$$

$$\hat{\Sigma} = \frac{\alpha}{\Sigma + 1}[(1 - \epsilon)\operatorname{erf}(\chi_{\text{IN}}) + \epsilon\operatorname{erf}(\chi_{\text{OUT}})]$$

where $\zeta_{\text{IN}} = \Delta_{\text{IN}} - 2m + q + 1$, $\zeta_{\text{OUT}} = \Delta_{\text{OUT}} + \beta^2 + q - 2\beta m$ and

$$\nu^{\text{Huber}} = \Sigma + 1, \quad \mu^{\text{Huber}} = a\nu^{\text{Huber}}, \quad \chi_{\text{IN}}^{\text{Huber}} = \mu^{\text{Huber}}/\sqrt{2\zeta_{\text{IN}}}, \quad \chi_{\text{OUT}}^{\text{Huber}} = \mu^{\text{Huber}}/\sqrt{2\zeta_{\text{OUT}}}. \tag{45}$$

A.2.1 Corollary: explicit form of the error for ℓ_2

Corollary A.1. *In the Ridge regression case, for a generic value of β, with fixed regularisation $\lambda \geq 0$ the generalisation error has the explicit form*

$$E_{\text{gen}} = 1 + (\beta^2 - 1)\epsilon + q - 2m(1 + (\beta - 1)\epsilon),$$

$$m = \frac{2\alpha\Gamma}{p + t + 1}, \tag{46}$$

$$q = \frac{4\alpha\left[\alpha\Gamma^2(p + t - 3) + \Lambda(p + t + 1)\right]}{(p + t + 1)[p^2 - 2c + t^2 + 2t(p + 1) + 1]},$$

where $\Delta_{eff} = (1 - \epsilon)\Delta_{IN} + \epsilon\Delta_{OUT}$, $p = \alpha + \lambda$, $c = \alpha - \lambda$, $t = \sqrt{(p-1)^2 + 4\lambda}$, $\Gamma = 1 + \epsilon(\beta - 1)$ and $\Lambda = 1 + \Delta_{eff} + \epsilon(\beta^2 - 1)$.

Proof. The form for the excess generalisation error is the same as in Theorem 4.1, so we look for a solution for the self-consistent equations, eqs. (36). The equations can be simplified if one defines $\Gamma = 1 + \epsilon(\beta - 1)$ and $\Lambda = 1 + \Delta_{\text{eff}} + \epsilon(\beta^2 - 1)$ with $\Delta_{\text{eff}} = (1 - \epsilon)\Delta_{\text{IN}} + \epsilon\Delta_{\text{OUT}}$. To solve them we first notice that the equations for Σ and $\hat{\Sigma}$ only depend on themselves and thus can be solved, getting:

$$\Sigma = \frac{1 - \alpha - \lambda + t}{2\lambda}, \quad \hat{\Sigma} = \frac{-1 + \alpha - \lambda + t}{2}, \tag{47}$$

where here and in the following we have that $t = \sqrt{4\lambda + (-1 + \alpha + \lambda)^2}$. Plugging these values in the equations for m and $\hat{m}$ gives

$$m = \frac{2\alpha\Gamma}{\alpha + \lambda + t + 1}, \quad \hat{m} = -\frac{2\alpha\Gamma\lambda}{-\alpha + \lambda + t + 1} \tag{48}$$

Finally, plugging everything in the equations for q and $\hat{q}$, we obtain

$$q = \frac{4\alpha\left(\alpha\Gamma^2(p + t - 3) + \Lambda(p + t + 1)\right)}{(p + t + 1)\left(-2c + (p + t)^2 + 2t + 1\right)}$$

$$\hat{q} = \frac{4\alpha\lambda^2}{(2\lambda - p + t + 1)^2}\left(\frac{4\alpha\left(\alpha\Gamma^2(p + t - 3) + \Lambda(p + t + 1)\right)}{(p + t + 1)\left(-2c + (p + t)^2 + 2t + 1\right)} + \Lambda - \frac{4\alpha\Gamma^2}{p + t + 1}\right) \tag{49}$$

where we defined $p = \alpha + \lambda$ and $c = \alpha - \lambda$. Upon rearranging of terms we have the same form for q as in the Corollary. $\square$

A.3 The rate of BO is $1/\alpha$

The equations for the BO are presented in the main text eq. (10). If we suppose that $q_{\mathrm{b}} = 1 - c/\alpha + \mathcal{O}(\alpha^{-2})$ and $\hat{q}_{\mathrm{b}} = \hat{c}\alpha + \mathcal{O}(1)$ we have that

$$q_{\mathrm{b}} = \frac{\hat{q}_{\mathrm{b}}}{1 + \hat{q}_{\mathrm{b}}} = 1 - \frac{1}{\hat{c}\alpha} \tag{50}$$

thus the relation between the expansion coefficients which is $\hat{c} = 1/c$. The coefficient $\hat{c}$ is found from the first order expansion of the self-consistent equations and it is equal to

$$\hat{c} = \int \mathrm{d}\xi \, \mathrm{d}y \, e^{-\frac{1}{2}\xi^2} \frac{\left(\frac{\epsilon(\beta\xi - y)e^{-\frac{(y-\beta\xi)^2}{2\Delta_{\mathrm{OUT}}}}}{\sqrt{2\pi}\Delta_{\mathrm{OUT}}} + \frac{(\epsilon-1)(y-\xi)e^{-\frac{(y-\xi)^2}{2\Delta_{\mathrm{IN}}}}}{\sqrt{2\pi}\Delta_{\mathrm{IN}}} \right)^2}{\frac{\epsilon e^{-\frac{(y-\beta\xi)^2}{2\Delta_{\mathrm{OUT}}}}}{\sqrt{2\pi}\sqrt{\Delta_{\mathrm{OUT}}}} - \frac{(\epsilon-1)e^{-\frac{(y-\xi)^2}{2\Delta_{\mathrm{IN}}}}}{\sqrt{2\pi}\sqrt{\Delta_{\mathrm{IN}}}}} \tag{51}$$

Note that even if we supposed a faster decay for the second term of q_{b} we would find anyway that $\hat{q}_{\mathrm{b}} \propto \alpha$.

A.4 Theorem 5.1: large sample size limit for the self-consistent equations for ERM

In this Appendix we derive the asymptotic value of the generalisation and estimation errors of ERM for $\alpha \to \infty$. We will first consider the case for the ℓ_2 loss where the explicit solution of the order parameters gives a more direct approach to study the above mentioned limit. After we will consider the ℓ_1 and Huber loss with a framework together.

ℓ_2 **loss.** In this case we can leverage Corollary A.1 to get the large α limit of the order parameters and thus of the generalisation error. We notice that the solution presented in Corollary A.1 are formal solution that do not suppose anything on the scalings in α of the other parameters, *i.e.* λ. We will consider the case where λ is constant in α and see that this suffices to have BO rates for the generalisation error. From eq. (46) we have by expanding

$$q = \Gamma^2 + \frac{\Lambda - \Gamma^2(2\lambda + 1)}{\alpha} + \frac{\Gamma^2\left(3\lambda^2 - 1\right) - 2\lambda\Lambda + \Lambda}{\alpha^2} + \mathcal{O}\left(\frac{1}{\alpha^3}\right), \tag{52}$$

and

$$m = \Gamma - \frac{\Gamma\lambda}{\alpha} + \frac{\Gamma(\lambda - 1)\lambda}{\alpha^2} + \mathcal{O}\left(\frac{1}{\alpha^3}\right), \tag{53}$$

with the same definition of $\Gamma = 1 + \epsilon(\beta - 1)$ and $\Lambda = 1 + \Delta_{\mathrm{eff}} + \epsilon(\beta^2 - 1)$. By the knowledge of the order parameters and the formula for the generalisation and estimation error one has that

$$
\begin{aligned}
E_{\mathrm{gen}} &= \frac{\Delta_{\mathrm{eff}} - (\beta - 1)^2(\epsilon - 1)\epsilon}{\alpha} \\
&+ \frac{2(\beta-1)^2\lambda(\epsilon-1)\epsilon + ((\beta-1)\lambda\epsilon + \lambda)^2 - ((\beta-1)^2(\epsilon-1)\epsilon) - 2\Delta_{\mathrm{eff}}\lambda + \Delta_{\mathrm{eff}}}{\alpha^2} + \mathcal{O}(\alpha^{-3}), \\
E_{\mathrm{estim}} &= (\beta - 1)^2\epsilon^2 + \frac{(\beta-1)\epsilon((2\lambda+1)(\epsilon-1) - \beta(2\lambda\epsilon + \epsilon - 1)) + \Delta_{\mathrm{eff}}}{\alpha} \\
&+ \frac{\lambda((\beta-1)\epsilon(\lambda(3(\beta-1)\epsilon + 4) - 2\beta) + \lambda) - ((\beta-1)^2(\epsilon-1)\epsilon) - 2\Delta_{\mathrm{eff}}\lambda + \Delta_{\mathrm{eff}}}{\alpha^2} + \mathcal{O}(\alpha^{-3}).
\end{aligned}
\tag{54}
$$

Notice that in the estimation error the value of the plateu depends only on the noise parameters ϵ and β but not on the noise variances.

ℓ_1 **and Huber loss** We make the following ansatz for the leading order form of the following parameters in α

$$m = m_0, \qquad q = q_0, \qquad \Sigma = \frac{\Sigma_0}{\alpha}, \qquad \hat{m} = \hat{m}_0\alpha, \qquad \hat{q} = \hat{q}_0\alpha, \qquad \hat{\Sigma} = \hat{\Sigma}_0\alpha, \tag{55}$$

where $m_0, q_0, \Sigma_0, \hat{m}_0, \hat{q}_0$ and $\hat{\Sigma}_0$ are constants not depending on α in the limit. We use the following ansatz for λ as a function of α

$$\lambda = \lambda_0 + \lambda_1\alpha, \tag{56}$$

again with λ_1 and λ_2 finite in the large α limit. We now show that this ansatz leads to a system of self-consistent equations involving only $m_0, q_0, \Sigma_0, \hat{m}_0, \hat{q}_0$ and λ_1, λ_0 and additionaly find the conditions for which the first order term in λ vanishes. This system characterises at leading order the behaviour of the errors of ERM.

From the loss-independent self-consistent equations presented in Theorem 4.1

$$m = \frac{\hat{m}}{\lambda + \hat{\Sigma}}, \quad q = \frac{\hat{m}^2 + \hat{q}}{(\lambda + \hat{\Sigma})^2}, \quad \Sigma = \frac{1}{\lambda + \hat{\Sigma}} \tag{57}$$

we obtain at large α the equations

$$m_0 = \frac{\hat{m}_0}{\lambda_1 + \hat{\Sigma}_0}, \quad q_0 = \frac{\hat{m}_0^2}{(\lambda_1 + \hat{\Sigma}_0)^2} = m_0^2, \quad \Sigma_0 = \frac{1}{\lambda_1 + \hat{\Sigma}_0} \tag{58}$$

Notice that the expressions are independent of λ_0, which means that for λ to have an effect on the generalisation error it needs to have a non-vanishing coefficient for the term linear in α, *i.e.* $\lambda_1 \neq 0$. At leading order, the loss-independent constants defined in Theorem 4.1 read

$$\zeta_{\text{IN},0} = \Delta_{\text{IN}} + (m_0 - 1)^2, \qquad \zeta_{\text{OUT},0} = \Delta_{\text{OUT}} + (m_0 - \beta)^2. \tag{59}$$

ℓ_1 **loss.** For the ℓ_1 loss we have at leading order

$$\chi_{\text{IN}} = \chi_{\text{OUT}} = \mu = \nu = \mathcal{O}(1/\alpha), \tag{60}$$

giving for the three loss-dependent self-consistent equations

$$
\begin{aligned}
\hat{m}_0 &= \sqrt{\frac{2}{\pi}} \left(\frac{1 - \epsilon}{\sqrt{\zeta_{\text{IN},0}}} + \frac{\beta \epsilon}{\sqrt{\zeta_{\text{OUT},0}}} \right), \\
\hat{\Sigma}_0 &= \sqrt{\frac{2}{\pi}} \left(\frac{1 - \epsilon}{\sqrt{\zeta_{\text{IN},0}}} + \frac{\epsilon}{\sqrt{\zeta_{\text{OUT},0}}} \right), \\
\hat{q}_0 &= 1.
\end{aligned}
\tag{61}
$$

Huber loss. For the Huber loss we will consider the scale parameter a to be α independent. Thus we have at leading order

$$\chi_{\text{IN}} = a/\sqrt{2\zeta_{\text{IN},0}}, \quad \chi_{\text{OUT}} = a/\sqrt{2\zeta_{\text{OUT},0}}, \quad \mu = a, \quad \nu = 1, \tag{62}$$

giving for the three loss-dependent self-consistent equations

$$
\begin{aligned}
\hat{m}_0 &= (1 - \epsilon) \operatorname{erf}\left(\frac{a}{\sqrt{2\zeta_{\text{IN},0}}} \right) + \beta \epsilon \operatorname{erf}\left(\frac{a}{\sqrt{2\zeta_{\text{OUT},0}}} \right), \\
\hat{\Sigma}_0 &= (1 - \epsilon) \operatorname{erf}\left(\frac{a}{\sqrt{2\zeta_{\text{IN},0}}} \right) + \epsilon \operatorname{erf}\left(\frac{a}{\sqrt{2\zeta_{\text{OUT},0}}} \right), \\
\hat{q}_0 &= (1 - \epsilon)\left(\zeta_{\text{IN},0} - a^2 \right) \operatorname{erf}\left(\frac{a}{\sqrt{2\zeta_{\text{IN},0}}} \right) + \epsilon \left(\zeta_{\text{OUT},0} - a^2 \right) \operatorname{erf}\left(\frac{a}{\sqrt{2\zeta_{\text{OUT},0}}} \right).
\end{aligned}
\tag{63}
$$

At leading order, the generalisation and estimation error read

$$
\begin{aligned}
E_{\text{gen}} &= 1 + \epsilon(\beta^2 - 1) + m_0^2 - 2m_0(1 + \epsilon(\beta - 1)) + \Delta_{\text{eff}} = \\
&= (m_0 - 1)^2 - 2m_0\epsilon(\beta - 1) + \epsilon(\beta^2 - 1) + \Delta_{\text{eff}}, \\
E_{\text{estim}} &= 1 + q_0 - 2m_0 = (m_0 - 1)^2.
\end{aligned}
\tag{64}
$$

A.4.1 Rates of ERM

Before deriving any conclusion from the set of self-consistent equations that we just derived in the large α limit, it is worth considering what is the scaling of the first subleading term for m and q, as it determines the scaling of the first subleading term for the generalisation and estimation errors.

It is easy to see that a general expansion in integer powers of alpha for the order parameters (so more general of the leading order ansatz of eq. (55)) still satisfies the self-consistent equations, as the order parameters enter the equations only through $\mathcal{C}^\infty$ functions. We just need to check that our functions are indeed smooth. This is contingent on the fact that $\Sigma_0 \neq 0$ and $\hat{\Sigma}_0 + \lambda_1 \neq 0$. The first condition can just assumed to be true because of eq. (58). Indeed, for Σ_0 to be zero we need $\hat{\Sigma}_0$ to diverge, which can't be since $\zeta_{\text{IN},0} > 0$, $\zeta_{\text{OUT},0} > 0$ as long as $\Delta_{\text{IN}} > 0$, $\Delta_{\text{OUT}} > 0$ (equation (59)). For the second one we can simply notice that as long as $\epsilon \leq 1$, $\hat{\Sigma}_0 > 0$.

Thus, the subleading order of m and q, and thus of the errors, is at least of order $\mathcal{O}(\alpha^{-1})$.

On the other hand, at the values of the parameters for which we have that the error of ERM converges to the BO performance, we have that the subleading order of the errors, is at most of order $\mathcal{O}(\alpha^{-1})$, which is the BO scaling.

Thus, whenever ERM is consistent, it also achieves BO rates. A more detailed computation would be needed to asses whether ERM is identical to BO at order $\mathcal{O}(\alpha^{-1})$, but we believe that this will not be the case in general.

A.4.2 Angle between ERM minimiser and teacher goes to zero

The angle between the student and the teacher weights is given in concentrates in high-dimension onto [Aubin et al., 2020]

$$\theta_{\boldsymbol{w}^\star, \hat{\boldsymbol{w}}} = \frac{1}{\pi} \arccos\left(\frac{m}{\sqrt{q}}\right) \tag{65}$$

but from the loss-independent equations eq. (58) we have $q \to m^2$, so that $\theta \to 0$ for large α.

A.4.3 Consistency of generalisation error

In order for the generalisation error to be consistent, we need to impose that the excess generalisation error vanishes, *i.e.* we require that

$$E_{\text{gen}} = (m_0 - 1)^2 - 2m_0\epsilon(\beta - 1) + \epsilon(\beta^2 - 1) + \Delta_{\text{eff}} \overset{!}{=} \epsilon(1 - \epsilon)(1 - \beta)^2 + \Delta_{\text{eff}} \tag{66}$$

Which results in $m_0 = \Gamma = 1 + \epsilon(\beta - 1)$. This imposes a condition on λ_1 because of eq. (58) which reads

$$\lambda_1 = \frac{\hat{m}_0}{\Gamma} - \hat{\Sigma}_0 \tag{67}$$

This needs to be compatible with $\lambda \geq 0$, which at large α implies $\lambda_1 \geq 0$. Firstly we will consider the ℓ_2 loss. In this case we have $\hat{m}_0 = \Gamma$ and $\hat{\Sigma}_0 = 1$, so $\lambda_1 = 0$ meaning that the value of λ doesn't need to diverge with α to have consistency of E_{gen}. This result is compatible to the one obtained by looking at the explicit solution of the order parameters as in Corollary A.1, where we have found that even with a finite value of λ in the limit $\alpha \to \infty$ the generalisation error has optimal rate.

For the other two losses the inequality in eq. (67) is equivalent to

$$(1 - \beta)(1 - \epsilon)\epsilon \left[\text{erf}\left(\frac{a}{\sqrt{2\zeta_{\text{IN},0}}}\right) - \text{erf}\left(\frac{a}{\sqrt{2\zeta_{\text{OUT},0}}}\right) \right] \geq 0 \tag{68}$$

For $\beta < 1$, considering that $\epsilon \in (0, 1)$, remembering the definitions in eq. (59), this condition is equivalent to

$$\Delta_{\text{OUT}} - \Delta_{\text{IN}} \geq (1 - \beta)^2(2\epsilon - 1). \tag{69}$$

If $\beta > 1$ the condition in eq. (69) flips the order of the condition.

We remark that in the limit $a \to \infty$ the equations is always satisfied. This tells us that whenever L2 is better then Huber, the optimisation over a of Huber will lead to $a \to \infty$, *i.e.* the Huber loss becomes the ℓ_2 loss.

A.4.4 Consistency of estimation error

In order for the estimation error to be consistent, we need to impose that the estimation error vanishes for $\alpha \to \infty$. This is the condition

$$E_{\text{estim}} = (m_0 - 1)^2 \stackrel{!}{=} 0 \,, \tag{70}$$

which is equivalent to $m_0 = 1$. As in Appendix A.4.3 we use eq. (58) and obtain the value of λ_1 as

$$\lambda_1 = (\beta - 1)\epsilon \operatorname{erf}\left(\frac{a}{\sqrt{2\zeta_{\text{OUT},0}}}\right) \,, \tag{71}$$

for the Huber loss. The other two losses can be obtained by setting $a \to \infty$ for the ℓ_2, and $a \to 0$ with the appropriate rescaling of λ_1 for ℓ_1. Both limits do not alter the overall sign of λ_1, which just depends on $\beta - 1$. This implies that the estimator is can only be consistent in estimation for $\beta > 1$ since we have the constraint $\lambda \geq 0$.

A.4.5 Consistency of estimation error for ℓ_2 with negative regularisation

In Appendix A.5 we show that the Ridge regression risk can endure negative regularisation without losing strict convexity for all $\lambda > -(1 - \sqrt{\alpha})^2$.

Using eq. (71), in the limit $a \to \infty$ to recover the ℓ_2 loss, we see that the condition above for lambda implies, at leading order in α and for the ℓ_2 loss that

$$(\beta - 1)\epsilon > -1 \,. \tag{72}$$

A.4.6 Oracle norm/label rescaling restores consistency

We notice that since the angle between the teacher and the student goes to zero as shown in Appendix A.4.2, it suffices to rescale the student optimally to achieve consistency. This can be achieved either by knowing the norm, or by cross-validation on the appropriate validation set to find the best norm.

Notice that all losses satisfy some kind of homogeneity property, *i.e.*

$$\ell_a^{\text{Huber}}\left(cy, \frac{\boldsymbol{x} \cdot \boldsymbol{w}}{\sqrt{d}}\right) = c^2 \, \ell_{ca}^{\text{Huber}}\left(y, \frac{\boldsymbol{x} \cdot \boldsymbol{w}}{c\sqrt{d}}\right) \,, \tag{73}$$

$$\ell_1\left(cy, \frac{\boldsymbol{x} \cdot \boldsymbol{w}}{\sqrt{d}}\right) = c \, \ell_1\left(y, \frac{\boldsymbol{x} \cdot \boldsymbol{w}}{c\sqrt{d}}\right) \,, \tag{74}$$

$$\ell_2\left(cy, \frac{\boldsymbol{x} \cdot \boldsymbol{w}}{\sqrt{d}}\right) = c^2 \, \ell_2\left(y, \frac{\boldsymbol{x} \cdot \boldsymbol{w}}{c\sqrt{d}}\right) \,. \tag{75}$$

For this reason, the rescaling of the ERM estimator can be produced equivalently by rescaling the labels of the training set, and the hyperparamters, appropriately.

A.5 Claim: Ridge regression risk is still convex for not-too-large negative regularisation

Consider the Ridge regression risk

$$\mathcal{R}_{\text{ridge}}(\boldsymbol{w}) = \frac{1}{2} \sum_{i=1}^{n} \left\| y_i - \frac{\boldsymbol{x}_i \cdot \boldsymbol{w}}{\sqrt{d}} \right\|_2^2 + \frac{\lambda}{2} \|\boldsymbol{w}\|_2^2 \,. \tag{76}$$

This is a quadratic form in the d-dimensional vector $\boldsymbol{w}$, with quadratic part

$$\mathcal{R}_{\text{ridge}}^{\text{quad.}}(\boldsymbol{w}) = \frac{1}{2} \sum_{a,b=1}^{d} w^a w^b \left(\frac{\alpha}{n} \sum_{i=1}^{n} x_i^a x_i^b + \lambda \delta_{ab}\right) \stackrel{\text{def}}{=} \frac{1}{2} \sum_{a,b=1}^{d} w^a w^b H_{a,b} \,, \tag{77}$$

where the Hessian of the quadratic form is the matrix $H = X + \lambda \operatorname{Id}$. The matrix X is a $d \times d$ Marchenko-Pastur random matrix. Supposing $\alpha = n/d > 1$, this matrix has minimal eigenvalue concentrating on $(1 - \sqrt{1/\alpha})^2$, giving minimal Hessian eigenvalue equal to

$$\text{minimal eigenvalue} = (1 - \sqrt{\alpha})^2 + \lambda \,. \tag{78}$$

Thus, the Ridge regression risk is convex as long as the Hessian is positive definite, meaning

$$\forall \lambda : \lambda > -(1 - \sqrt{\alpha})^2. \tag{79}$$

A.6 Theorem 6.1: small outliers' percentage ϵ limit of the error of ERM with ℓ_2 loss

We start from the knowledge of the explicit form of the order parameters in the Ridge regression case, presented in Corollary A.1. If one expands m and q up to second order in ϵ one has that:

$$m = m_0 + m_1\epsilon + m_2\epsilon^2 + \mathcal{O}(\epsilon^3), \; q = q_0 + q_1\epsilon + q_2\epsilon^2 + \mathcal{O}(\epsilon^3), \; \lambda = \lambda_0 + \lambda_1\epsilon + \lambda_2\epsilon^2 + \mathcal{O}(\epsilon^3), \tag{80}$$

with coefficients that are constants in ϵ. The generalisation error up to second order in ϵ is

$$
\begin{aligned}
E_{\text{gen}} &= E_{\text{gen}0} + E_{\text{gen}1}\epsilon + E_{\text{gen}2}\epsilon^2 + \mathcal{O}(\epsilon^3) \\
&= 1 + q + (\beta^2 - 1)\epsilon + \Delta_{\text{eff}} - 2m((\beta - 1)\epsilon + 1) \\
&= (1 - 2m_0 + q_0 + \Delta_{\text{IN}}) + ((\beta - 1)(\beta - 2m_0 + 1) + \Delta_{\text{OUT}} - \Delta_{\text{IN}} - 2m_1 + q_1)\epsilon \\
&\quad + (1 - 2(\beta - 1)m_1 - 2m_2)\epsilon^2 + \mathcal{O}(\epsilon^3)
\end{aligned}
\tag{81}
$$

This shows that the generalisation error has the correct expansion in powers of ϵ. We now need to find the optimal regularisation λ, and plug it back in in the generalisation error to complete the proof.

To start, we notice that at $\epsilon = 0$ the BO generalisation error can be found explicitly, as the output channel become a single Gaussian, giving

$$
\begin{aligned}
E_{\text{gen}0}^{\text{BO}} &= \frac{1}{2}\left(\sqrt{(\alpha + \Delta_{\text{IN}} + 1)^2 - 4\alpha} - \alpha + \Delta_{\text{IN}} + 1\right) \\
E_{\text{gen}0}^{\ell_2} &= \frac{1}{2}\left(\frac{\lambda_0(\alpha + \Delta_{\text{IN}}) + \alpha(\alpha + \Delta_{\text{IN}} - 2) + \Delta_{\text{IN}} + \lambda_0 + 1}{\sqrt{(\alpha + \lambda_0 - 1)^2 + 4\lambda_0}} - \alpha + \Delta_{\text{IN}} + 1\right)
\end{aligned}
\tag{82}
$$

where in the second line we compare with the zero-th order term of ℓ_2 ERM. The two generalisation errors match for $\lambda_0 = \Delta_{\text{IN}}$, giving us the zero-th order value of the expansion of the optimal regularisation.

To find the next orders for the optimal regularisation λ we compute order-by-order the derivative of the generalisation error with respect to λ. At first order we have

$$\frac{\partial E_{\text{gen}1}}{\partial \lambda_1} = \frac{2\alpha(\lambda_0 - \Delta_{\text{IN}})}{(\alpha^2 + 2\alpha(\lambda_0 - 1) + (\lambda_0 + 1)^2)^{3/2}} = 0 \tag{83}$$

which vanishes as soon as $\lambda_0 = \Delta_{\text{IN}}$ as computed previously. Looking at the derivative at second order in ϵ allows to fix λ_1 as

$$\lambda_1 = \frac{(\alpha^2 + 2\alpha(\lambda_0 - 1) + (\lambda_0 + 1)^2)(\beta^2 - 2\beta(\lambda_0 + 1) - \Delta_{\text{IN}} + \Delta_{\text{OUT}} + 2\lambda_0 + 1)}{\alpha^2 + \alpha(3\Delta_{\text{IN}} - \lambda_0 - 2) + (\lambda_0 + 1)(3\Delta_{\text{IN}} - 2\lambda_0 + 1)}. \tag{84}$$

Thus the optimal order parameter can be written as $\lambda_{\text{opt}} = \Delta_{\text{IN}} + \lambda_1\epsilon + \mathcal{O}(\epsilon^2)$ and the value of the error is

$$
\begin{aligned}
E_{\text{gen}}^{\ell_2} = E_{\text{gen}0}^{\text{BO}} &+ \left[\frac{1}{2\kappa}\left(2\alpha^2(\beta - 1) + \alpha(\beta(\beta + 2\Delta_{\text{IN}} - 6) - 3\Delta_{\text{IN}} + \Delta_{\text{OUT}} + 5)\right.\right. \\
&+ (\Delta_{\text{IN}} + 1)(\beta^2 - \Delta_{\text{IN}} + \Delta_{\text{OUT}} - 1)\Big) + \frac{1}{2}(-2\alpha(\beta - 1) + \beta^2 + \Delta_{\text{IN}} - \Delta_{\text{OUT}} - 1)\bigg]\epsilon \\
&+ \mathcal{O}(\epsilon^2),
\end{aligned}
\tag{85}
$$

with the definitions of eq. (82) and

$$\kappa = \sqrt{\alpha^2 + 2\alpha(\Delta_{\text{IN}} - 1) + (\Delta_{\text{IN}} + 1)^2}. \tag{86}$$

B Simulations details

B.1 Solving the self-consistent equations for ERM and BO

The self-consistent equations from Proposition 4.1 are written in a way amenable to be solved via fixed-point iteration. From a random initialisation, we iterate through both the hat and non-hat variable equations until the maximum absolute difference between the order parameters in two successive iterations falls below a tolerance of 10^{-9}.

To speed-up convergence we use a damping scheme, updating each order parameter at iteration i, designated as x_i, using $x_i := x_i \mu + x_{i-1}(1 - \mu)$, with μ as the damping parameter ($\mu \in [0.7, 0.9]$ was found most effective).

The Bayes-optimal (BO) fixed-point equations involve integrals which we could not express explicitly, necessitating a numerical integrator. The integration region must be subdivided into smaller regions to improve accuracy. In the iteration of the self-consistent equations, the above-mentioned damping trick is employed.

Once convergence is achieved for fixed λ and possibly a, hyper-parameters are optimised using the Nelder-Mead optimiser, a gradient-free numerical minimisation procedure.

B.2 Numerical simulations for ERM

We obtain numerical simulations for the Empirical Risk Minimisation (ERM) by sampling synthetic datasets using eq. (1). The ERM problem is solved as follows:

• for ℓ_2 ERM is performed using the closed form solution. If we define the data matrix $\Phi \in \mathbb{R}^{n,d}$ as $\Phi_{ij} = x_i^j$, then the ℓ_2 ERM problem is solved by:

$$\hat{\boldsymbol{w}} = \left(\frac{\Phi^\top \Phi}{d} + \lambda \operatorname{Id} \right)^{-1} \frac{\Phi^\top \boldsymbol{y}}{\sqrt{d}} \tag{87}$$

• For the Huber loss, ERM is performed using the routine of the Scikit-learn library [Pedregosa et al., 2011], which uses a L-BFGS optimiser [Liu and Nocedal, 1989].
• For the ℓ_1 loss we use its mapping onto Huber with small scale parameter, setting $a = 10^{-3}$. To obtain the same result as the Huber regression with regularisation parameter λ oen has to consider the ℓ_1 regression with regularisation parameter $a\lambda$.

For the comparison with synthetic data in Figure 1 we first find the optimal parameters (λ, a) for the generalisation error from the explicit solution of the fixed point equations. These values are then plugged into the ERM procedures described above.

To evaluate the estimation error it is pretty straightforward from the definition in eq. (3). To evaluate the excess generalisation error from the definition of eq. (2), we generated n_t sample dataset from the same teacher and, given an estimate for the weights $\hat{\boldsymbol{w}}$ we evaluated the following

$$\frac{1}{n_t} \sum_{i=1}^{n_t} \left(y_i - \frac{\boldsymbol{x} \cdot \hat{\boldsymbol{w}}}{\sqrt{d}} \right)^2 - \frac{1}{n_t} \sum_{i=1}^{n_t} \left(y_i - (1 - \epsilon + \beta\epsilon) \frac{\boldsymbol{x} \cdot \boldsymbol{w}^\star}{\sqrt{d}} \right)^2, \tag{88}$$

where the rescaling $(1 - \epsilon + \epsilon\beta)$ is the one with minimal generalisation error.

The error bars presented in all the Figures are the standard deviation on the mean over all the realisations for the specific cases.

B.3 Numerical simulations for BO: the Generalised Approximate Message Passing (GAMP) algorithm

We summarise here the Generalised Approximate Message Passing (GAMP) algorithm, firstly introduced in [Rangan, 2011]. We will follow the notation in [Aubin et al., 2020]. Given loss and regularisation functions f_{out} and f_w, computed as in Appendix (A.2), we can estimate the student weight $\hat{\boldsymbol{w}}$ of the associated ERM problem
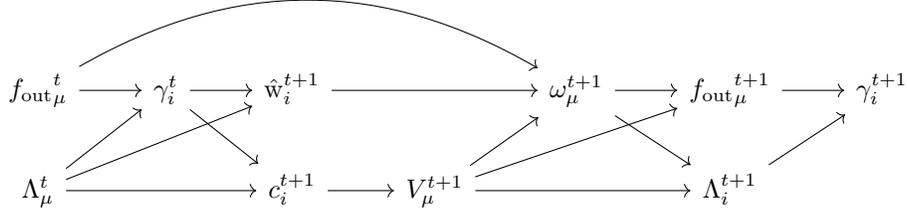
Figure 4: Sketch of the dependence of the various quantities in GAMP algorithm.

by iterating the following fixed point system

$$\Lambda_i^t = -\frac{1}{d}\sum_{\mu=1}^{n} X_{\mu i}^2 \partial_\omega f_{\text{out}\,\mu}^t$$

$$f_{\text{out}\,\mu}^t = f_{\text{out}}(y_\mu, \omega_\mu^t, V_\mu^t)$$

$$\gamma_i^t = \frac{1}{\sqrt{d}}\sum_{\mu=1}^{n} X_{\mu i} f_{\text{out}\,\mu}^t + \Lambda_i^t \hat{\boldsymbol{w}}_i^t$$

$$\hat{\boldsymbol{w}}_i^{t+1} = f_{\text{w}}(\gamma_i^t, \Lambda_i^t)$$

$$c_i^{t+1} = \partial_\gamma f_{\text{w}}(\gamma_i^t, \Lambda_i^t)$$

$$V_\mu^t = \frac{1}{d}\sum_{i=1}^{d} X_{\mu i}^2 \hat{c}_i^t$$

$$\omega_\mu^t = \frac{1}{\sqrt{d}}\sum_{\mu=1}^{n} X_{\mu i}\hat{\boldsymbol{w}}_i^t - V_\mu^t f_{\text{out}\,\mu}^{t-1}$$

(89)

It has been shown [Barbier et al., 2019] that if $f_{\text{w}} = f_{\text{w}^\star}$ and $f_{\text{out}} = f_{\text{out}^\star}$ then GAMP provides a Bayes Optimal estimation $\hat{\boldsymbol{w}} = \boldsymbol{w}_{\text{BO}}$ as per Section 4. We sketch in Figure 4 the dependence of the parameters of GAMP during one iteration step going from iteration time t to iteration time $t+1$.

C   Experiments with non-Gaussian data

We provide here some numerical experiments for a model of data for which the samples x are not Gaussian and non-i.i.d. In particular, we consider a regression problem on the MNIST dataset [Lecun et al., 1998], preprocessed by extracting 300 random Fourier data features. The experiments are presented in Figure 5. We find results which are qualitatively compatible with what we presented in the main text. We observe (Figure 5, left) that the estimation error is not consistent for all the three losses we consider. We also observe (Figure 5, center and right) the sharp transition behaviour where Huber collapses to the ℓ_2 loss, with a_{opt} diverging (the fact that the parameter does not diverge is to be expected, as for finite sample complexity α, there will be a value $a = a_{\text{opt}}$ above which the Huber loss behaves as the ℓ_2 loss at all points in the training dataset, and increasing a above such a threshold will not affect the learning problem anymore.). In particular, Figure 5, center, shows the performance of the ℓ_2 and Huber losses, and Figure 5, right, the corresponding optimal value of the Huber scale parameter. We can see that the phase transition in Δ_{OUT} is conserved even with real data.

The data is generated as follows. We select n MNIST images, each considered as vector $X_i \in \mathbb{R}^{d_0}$ for $i = 1, \ldots, n$, and we compute for each of them p random Fourier features by generating a random real-valued Gaussian matrix $\Omega \in \mathbb{R}^{p \times d_0}$, and computing $\hat{X}_i = \exp(-i\Omega X_i) \in \mathbb{C}^p$, where the exponential is taken component-wise. This provides p complex features for each sample, which, when unpacked into their real and imaginary parts, translate into $2p$ random real Fourier features for each sample. We collect the features into the data matrix $\hat{X} \in \mathbb{R}^{2p \times n}$, whose aspect ratio $\alpha = n/2p$ defines the sample complexity of the linear regression problem. The labels are generated using the model of outliers considered in the manuscript (1) with a random $\boldsymbol{w}^\star \in \mathbb{R}^{2p}$. The regression problem is solved, as detailed in Appendix B.2, for the three choices of losses considered in the submission. In this case, we fix $p = 300$ and choose n depending on the value of α.

The optimal values of the regularisation λ and the scale parameter a are computed by cross-validation on the estimation error.
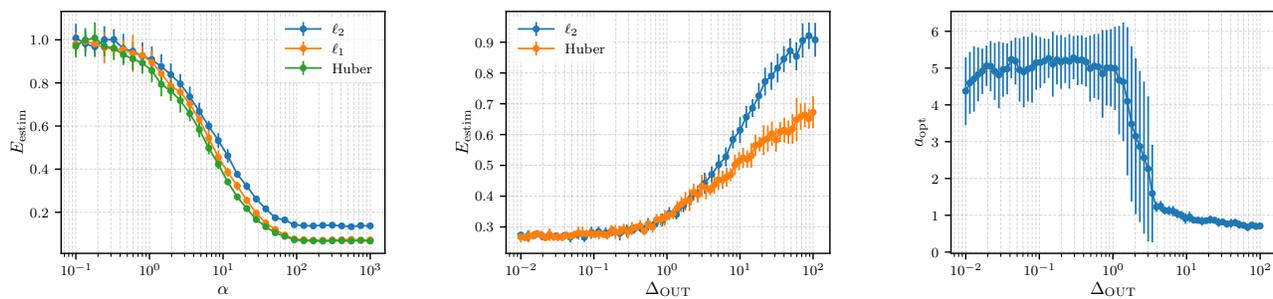
Figure 5: The points represent the mean and error of the mean of 16 instances of the simulations with data as described before. The simulations are performed with 30% of outliers ($\epsilon = 0.3$), with an additive noise variance of the inliers $\Delta_{\mathrm{IN}} = 1$ and correlation coefficient of the outliers $\beta = 0$. The ℓ_2 regularisation parameter is chosen to be the one that minimises the estimation error in all three Figures. **(Left)** This plot is the counterpart to Figure 1 (right) of the main text. Here we vary the sample complexity α and fix the noise variance of the outliers $\Delta_{\mathrm{OUT}} = 5$. Additionally for the Huber loss we fixed the scale parameter $a = 1$. **(Center) & (Right)** These two plots are the counterparts of Figure 2 (right) of the main text. In these plots we are varying Δ_{OUT} with a fixed sample complexity $\alpha = 10$.

D   Exploration of parameters' phenomenology

In this appendix we complement the figures in the main text by offering a more comprehensive exploration of the dependence of parameters α, ϵ, Δ_{IN}, Δ_{OUT} and β. In most of the figures we plot both the error (generalisation or estimation) in the upper panel and the value of the optimal hyper parameters in the bottom panel. Hyperparameters are plotted with colors matching the associated upper panel.

Gray lines correspond to the optimally tuned value of the scale parameter a for Huber.

In Figure 6, 7, 8, 9, 10 and 11 we explore the dependence on α:

- in Figures 6 and 7 we plot the optinally-tuned excess generalisation error as a function of α for two choices of β. In figure 6 we displayed the dependence on Δ_{OUT}. We notice that the phenomenology is qualitatively the same for the two different values of β ($\beta = 0$ on the left column and $\beta = 0.2$ on the right column). As we expect from Theorem 5.1. for large enough values of Δ_{OUT} we achieve consistency with all losses, in particular with a diverging λ_{opt}. On the other hand if the value of Δ_{OUT} is smaller than the threshold in 5.1 no choice of $\lambda \geq 0$ restores consistency in the large α limit. In Figure 7 we show also that the same phenomenology holds when changing ϵ. As in Figure 6 we see that for certain choices of the parameters, at large enough ϵ the estimator is not consistent.

- In Figures 8 and 9, we consider the optimally-tuned estimation error. In the figure 8 we focus on $\beta < 1$ and see that estimation cannot be found for positive values of λ. In figure 9 we have instead $\beta > 1$. Here consistent estimation can be achieved for an optimally tuned λ, but we stress that this requires a clean dataset.

- In Figure 10 we plotted the teacher-student angle as per eq. (65). We see that the figures are consistent with the claims of Section A.4.2 since even for non-optimised values of the parameters, the angle between the teacher and student vector is going to zero with α.

- In Figure 11 we plotted both the generalisation error and the estimation error after the rescaling procedure explained in Section 5. We see that by tuning the norm in this way, without considering optimising over the ℓ_2 regularisation parameter we regin consistency.

Figure 12 and Figure 13 explore the dependency of the generalisation error on ϵ and Δ_{OUT}. We have qualitatively the same behaviour as in Figure 2 of the main text. Interestingly, the excess generalisation error is not zero in the limit $\epsilon \to 1$ for $\beta \neq 0$ for the parameters chosen in figures. The outlier regression problem in this limit is equivalent to regression without outlier, but with variance Δ_{OUT}.

Figure 14 shows quantitatively how the Huber loss performs better than the ℓ_2 loss in the regression task at hand when optimally tuned. This complements Figure 3 of the main text, where we simply delimited the region where the two losses are identical. Interestingly, the advantage of Huber over ℓ_2 grows the further away one is from the borders of this region.

Matteo Vilucchio, Emanuele Troiani, Vittorio Erba, Florent Krzakala
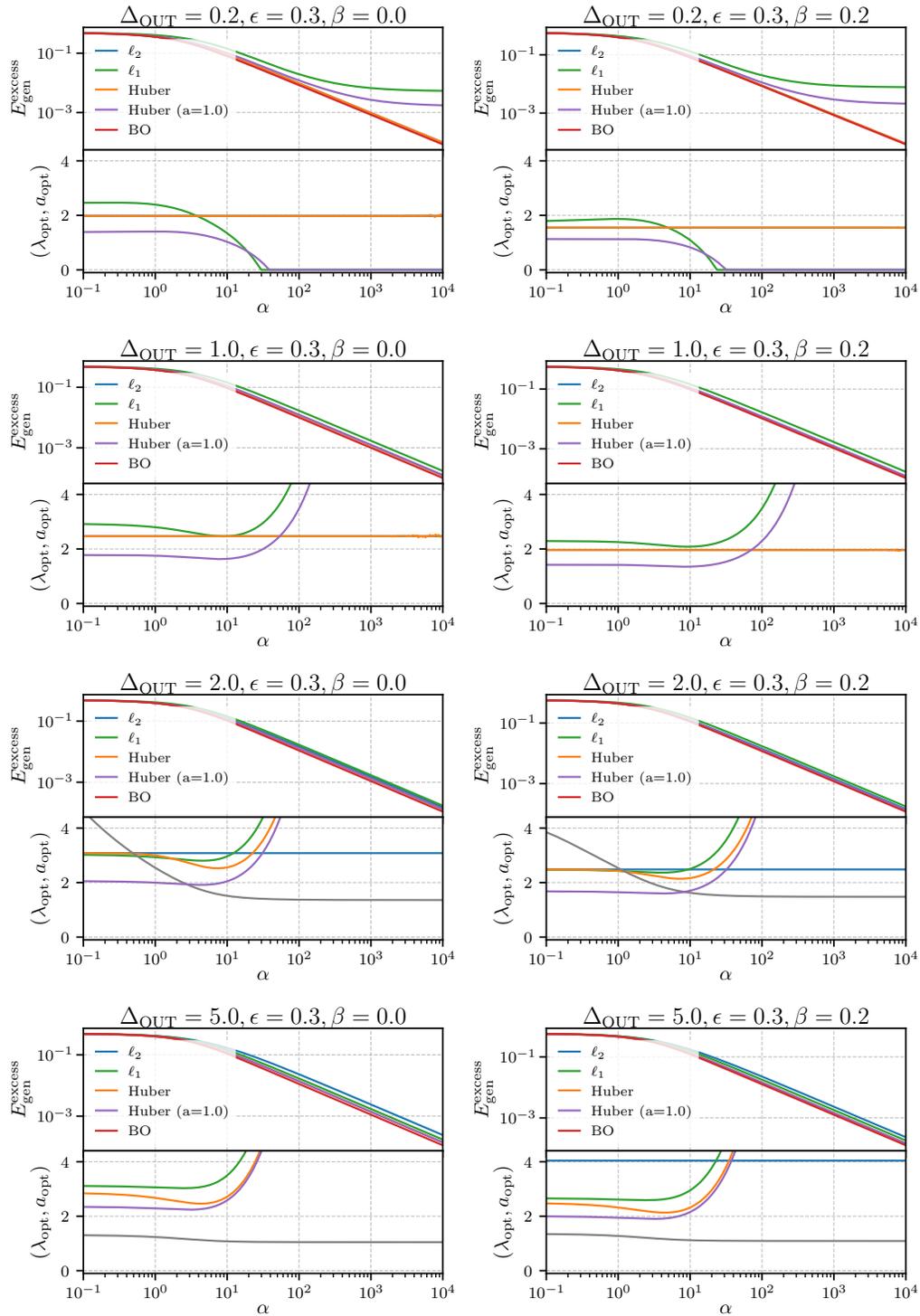
Figure 6: Excess generalisation error as a function of sample complexity for different value of the noise models' parameters. The hyper parameters, λ and optionally a, for the ERM procedures are optimised to obtain the best generalisation error.
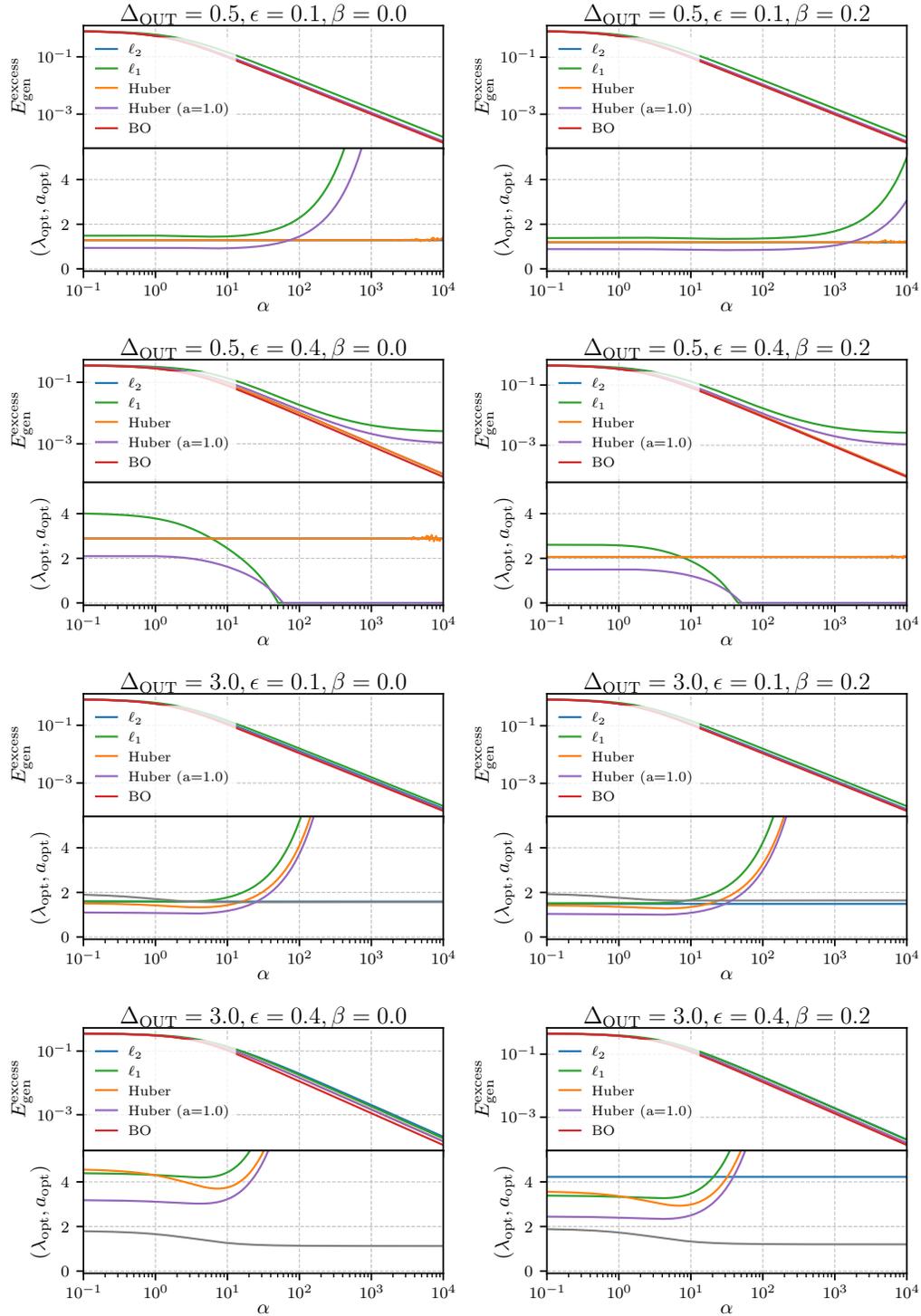
Figure 7: Excess generalisation error as a function of sample complexity for different value of the noise models' parameters. In all the plots shown we fixed $\Delta_{\mathrm{IN}} = 1$. The hyper parameters, λ and optionally a, for the ERM procedures are optimised to obtain the best generalisation error.
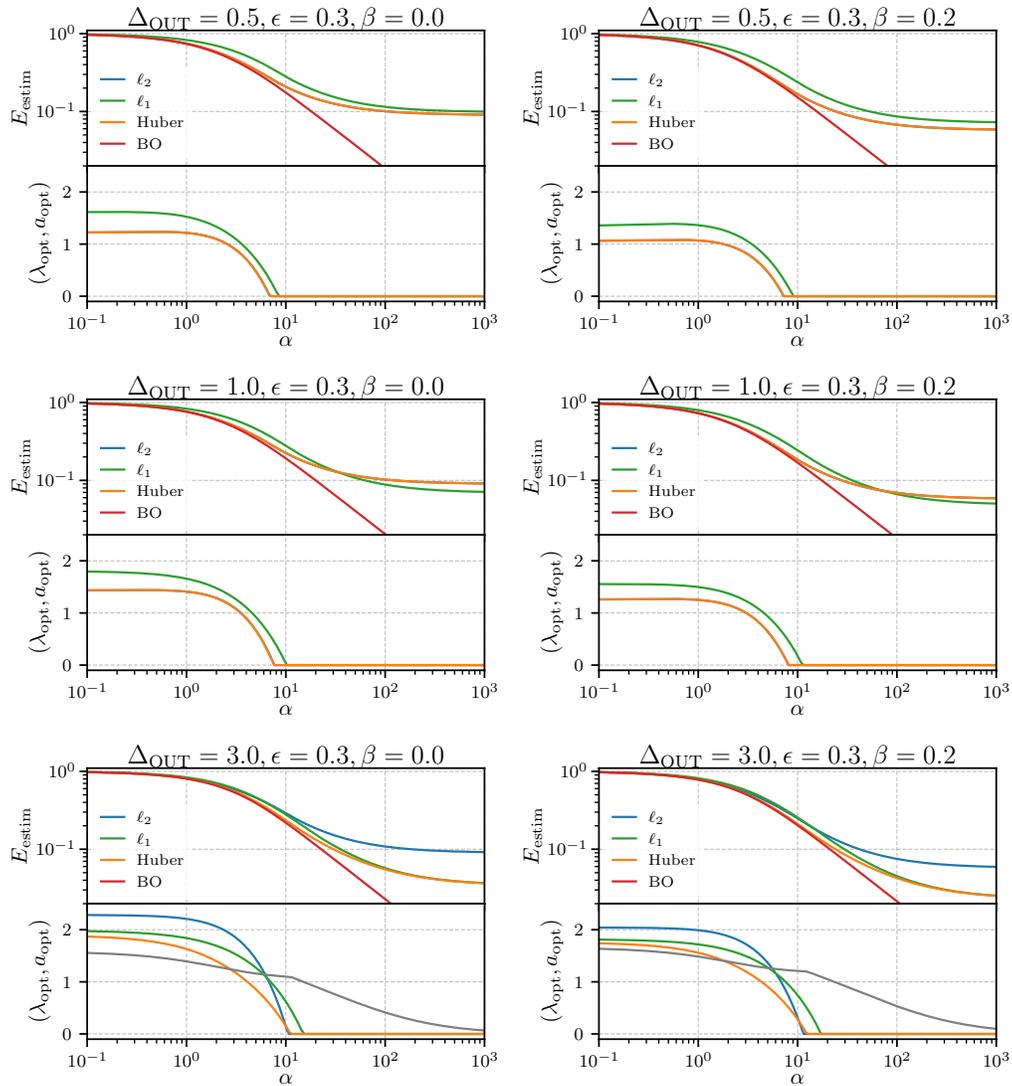
Figure 8: Estimation error as a function of sample complexity for different value of the noise models' parameters for $\beta < 1$. In all the plots shown we fixed $\Delta_{\mathrm{IN}} = 1$. The hyper parameters, λ and optionally a, for the ERM procedures are optimized to obtain the best generalisation error.
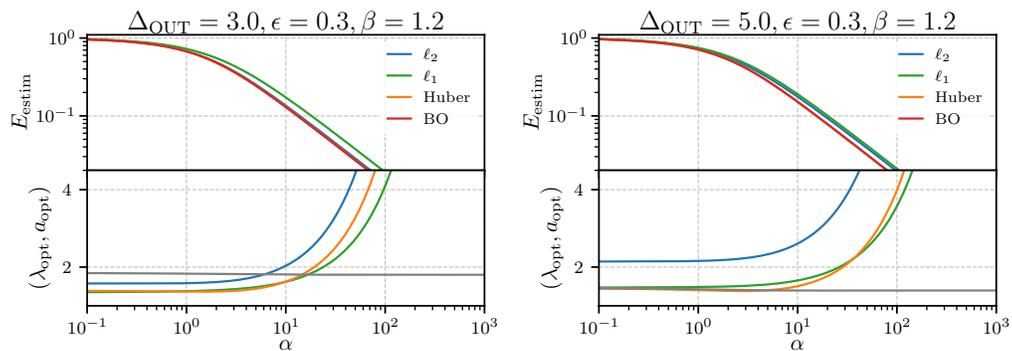
Figure 9: Estimation error as a function of sample complexity for different value of the noise models' parameters for $\beta > 1$. In all the plots shown we fixed $\Delta_{\mathrm{IN}} = 1$. The hyper parameters, λ and optionally a, for the ERM procedures are optimized to obtain the best generalisation error.
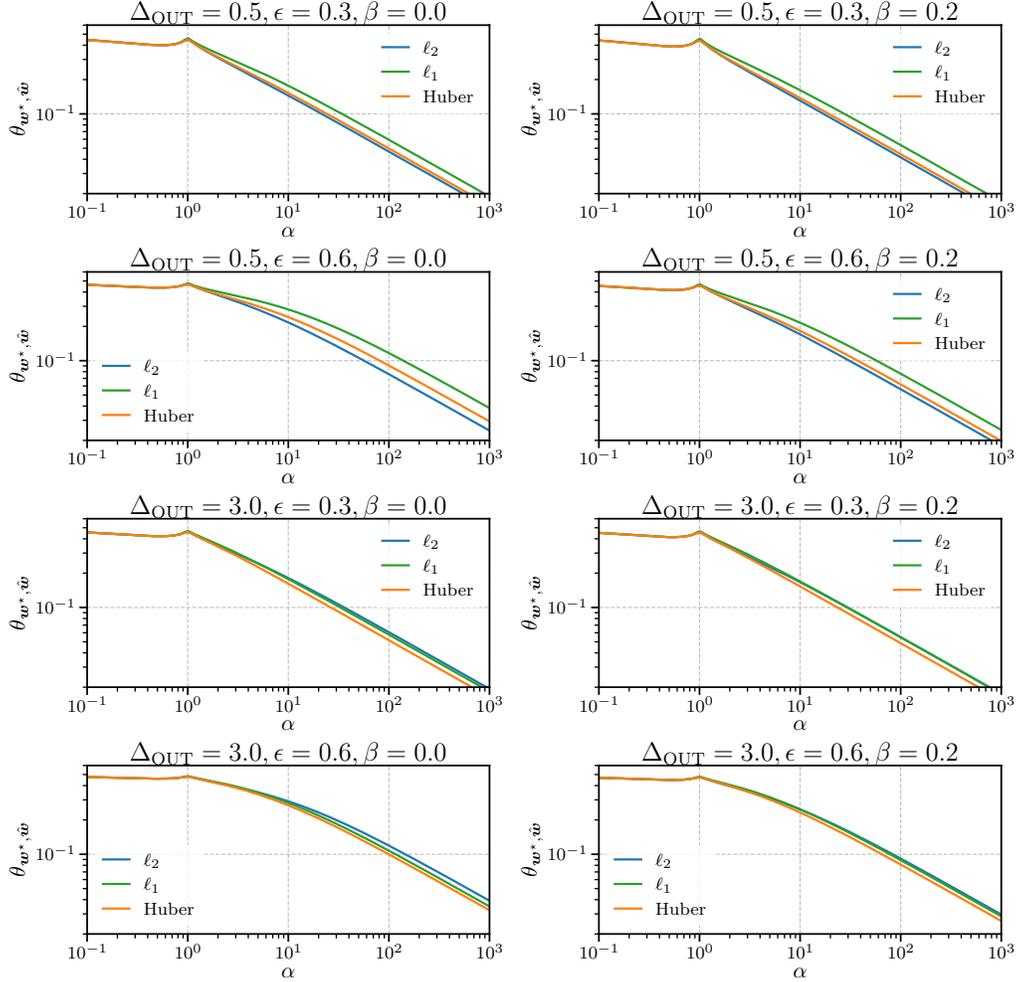
Figure 10: Angle between the teacher and the student — eq. (65) — as a function of the sample complexity α. The figures are generated for a fixed value of regularisation parameter $\lambda = 10^{-3}$ and in the case of the Huber loss for $a = 1$. For each plot we fixed $\Delta_{\mathrm{IN}} = 1$ and the other parameters chosen are indicated on top of the plot.

Matteo Vilucchio, Emanuele Troiani, Vittorio Erba, Florent Krzakala
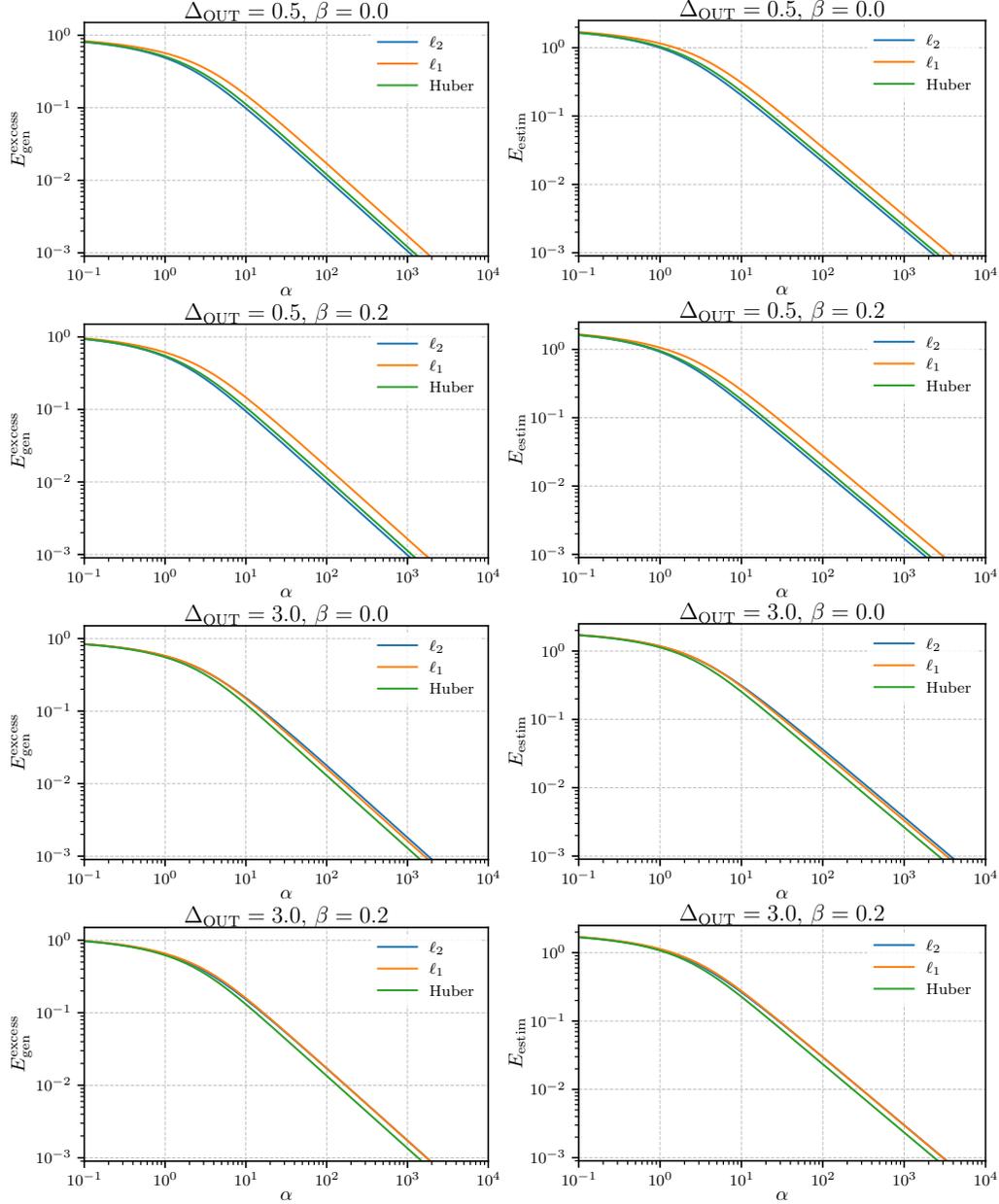
Figure 11: Excess generalisation error (first column) and estimation error (second column) after the rescale procedure explained in Section 5 as a function of the sample complexity α. The figures are generated for a fixed value of regularisation parameter $\lambda = 1$ and in the case of the Huber loss for $a = 1$. For each plot we fixed $\Delta_{\text{IN}} = 1$ and $\epsilon = 0.3$ the other parameters chosen are indicated on top of the plot.
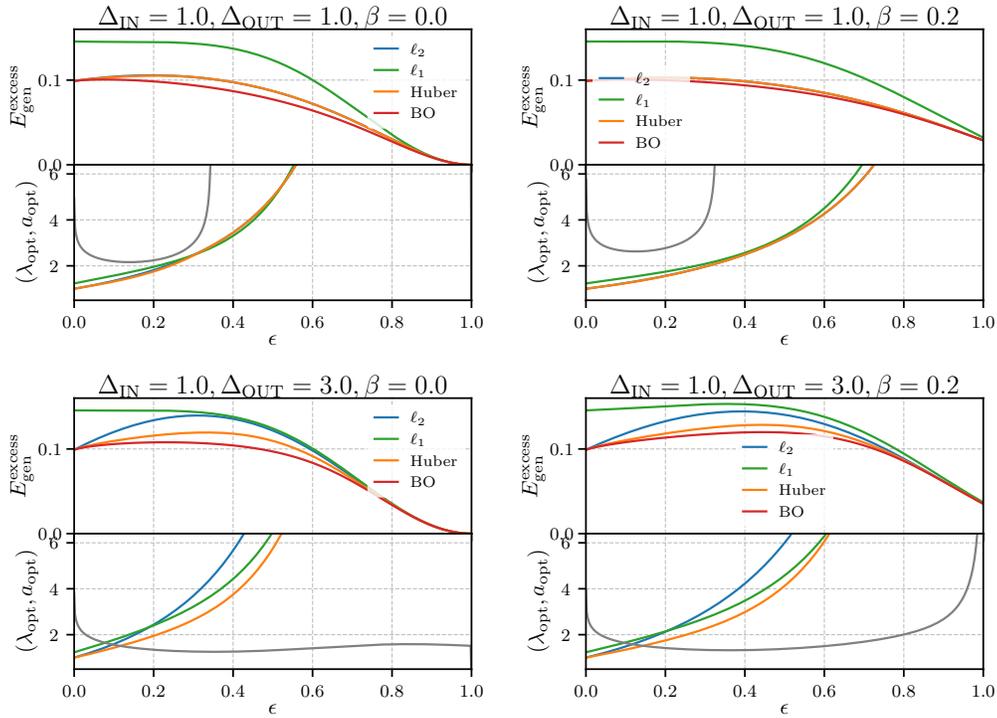
Figure 12: Excess generalisation error as a function of ϵ. In all the plots shown we fixed $\alpha = 10$. The hyper parameters, λ and optionally a, for the ERM procedures are optimized to obtain the best generalisation error.
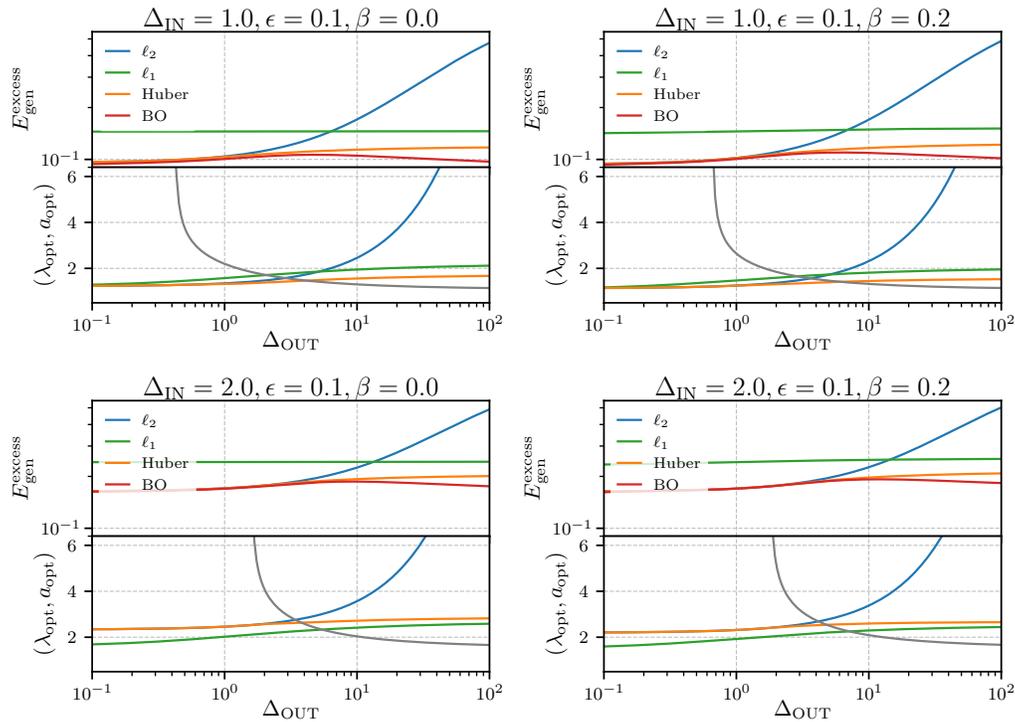
Figure 13: Excess generalisation error as a function of Δ_{OUT}. In all the plots shown we fixed $\alpha = 10$. The hyper parameters, λ and optionally a, for the ERM procedures are optimized to obtain the best generalisation error.
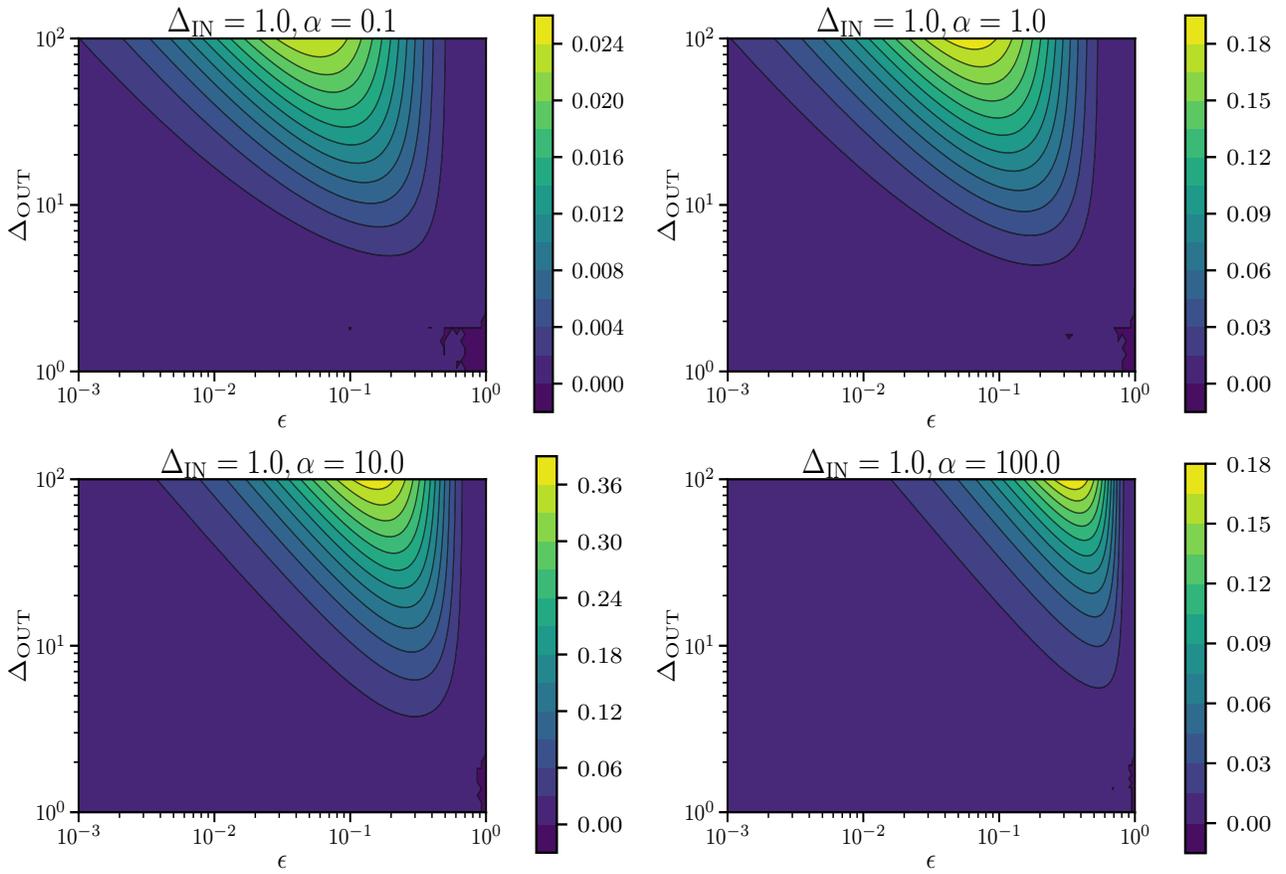
Figure 14: Difference between the Huber and ℓ_2 generalisation error as a function of Δ_{OUT} and ϵ. Here we fix $\Delta_{\mathrm{IN}} = 1$. Hyper-parameters for both losses are optimally tuned. In the dark blue region the difference is identically zero.