Generalisation error in learning with random features and the hidden manifold model
Supplementary material

Contents

Preprint. Under review.

1 Definitions and notations

In this section we recall the models introduced in the main body of the article, and introduce the notations used throughout the Supplementary Material.

1.1 The dataset

In this work we study a series of regression and classification tasks for a dataset $\{\boldsymbol{x}^\mu, y^\mu\}_{\mu=1}^n$ with labels $y^\mu \in \mathbb{R}$ sampled identically from a *generalised linear model*:

$$y^\mu \sim P_y^0\left(y^\mu \Big| \frac{\boldsymbol{c}^\mu \cdot \boldsymbol{\theta}^0}{\sqrt{d}}\right), \tag{1.1}$$

where the output-channel $P_y^0\left(\cdot\right)$ is defined as:

$$P_y^0\left(y^\mu \Big| \frac{\boldsymbol{c}^\mu \cdot \boldsymbol{\theta}^0}{\sqrt{d}}\right) = \int d\xi^\mu P\left(\xi^\mu\right)\delta\left(y^\mu - f^0\left(\frac{\boldsymbol{c}^\mu \cdot \boldsymbol{\theta}^0}{\sqrt{d}}; \xi^\mu\right)\right) \tag{1.2}$$

for some noise ξ^μ and for data points $\boldsymbol{x}^\mu \in \mathbb{R}^p$ given by:

$$\boldsymbol{x}^\mu = \sigma\left(\frac{1}{\sqrt{d}}\sum_{\rho=1}^d c_\rho^\mu \boldsymbol{f}_\rho\right). \tag{1.3}$$

The vectors $\boldsymbol{c}^\mu \in \mathbb{R}^d$ is assumed to be identically drawn from $\mathcal{N}(0, \mathrm{I}_d)$, and $\boldsymbol{\theta}^0 \in \mathbb{R}^d$ from a separable distribution P_θ. The family of vectors $\boldsymbol{f}_\rho \in \mathbb{R}^p$ and the scalar function $\sigma : \mathbb{R} \to \mathbb{R}$ can be arbitrary.

Although our results are valid for the general model introduced above, the two examples we will be exploring in this work are the noisy linear channel (for regression tasks) and the deterministic sign channel (for classification tasks):

$$y^\mu = \frac{\boldsymbol{c}^\mu \cdot \boldsymbol{\theta}^0}{\sqrt{d}} + \sqrt{\Delta}\,\xi^\mu \qquad \Leftrightarrow \qquad P_y^0\left(\boldsymbol{y} \Big| \frac{\boldsymbol{c}^\mu \cdot \boldsymbol{\theta}^0}{\sqrt{d}}\right) = \prod_{\mu=1}^n \mathcal{N}\left(y^\mu; \frac{\boldsymbol{c}^\mu \cdot \boldsymbol{\theta}^0}{\sqrt{d}}, \Delta\right) \tag{1.4}$$

$$y^\mu = \mathrm{sign}\left(\frac{\boldsymbol{c}^\mu \cdot \boldsymbol{\theta}^0}{\sqrt{d}}\right) \qquad \Leftrightarrow \qquad P_y^0\left(\boldsymbol{y} \Big| \frac{\boldsymbol{c}^\mu \cdot \boldsymbol{\theta}^0}{\sqrt{d}}\right) = \prod_{\mu=1}^n \delta\left(y^\mu - \mathrm{sign}\left(\frac{\boldsymbol{c}^\mu \cdot \boldsymbol{\theta}^0}{\sqrt{d}}\right)\right) \tag{1.5}$$

where $\xi^\mu \sim \mathcal{N}(0,1)$ and $\Delta > 0$.

This dataset can be regarded from two different perspectives.

Hidden manifold model: The dataset $\{\boldsymbol{x}^\mu, y^\mu\}_{\mu=1,\cdots,n}$ is precisely the *hidden manifold model* introduced in [1] to study the dynamics of online learning in a synthetic but structured dataset. From this perspective, although $\boldsymbol{x}^\mu$ lives in a p dimensional space, it is parametrised by a latent $d < p$-dimensional subspace spanned by the basis $\{\boldsymbol{f}_\rho\}_{\rho=1,\cdots,d}$ which is "hidden" by the application of a scalar nonlinear function σ acting component-wise. The labels $\boldsymbol{y}^\mu$ are then drawn from a generalised linear rule defined on the latent d-dimensional space.

Random features model: The dataset $\{\boldsymbol{x}^\mu, y^\mu\}_{\mu=1,\cdots,n}$ is tightly related to the Random Features model studied in [2] as a random approximation for kernel ridge regression. In this perspective, $\boldsymbol{c}^\mu \in \mathbb{R}^d$ is regarded as a collection of d-dimensional data points which are projected by a random feature matrix $\mathrm{F} = (\boldsymbol{f}_\rho)_{\rho=1}^p \in \mathbb{R}^{d\times p}$ into a higher dimensional space, followed by a non-linearity σ. In the limit of infinite number of features $d, p \to \infty$ with fixed ratio d/p, performing ridge regression of $\boldsymbol{x}^\mu$ is equivalent to kernel ridge regression with a limiting kernel depending on the distribution of the feature matrix F and on the non-linearity σ.

1.2 The task

In this work, we study the problem of learning the rule from eq. (1.1) from the dataset $\{(\boldsymbol{x}^\mu, y^\mu)\}_{\mu=1,\cdots,n}$ introduced above with a *generalised linear model*:

$$\hat{y}^\mu = \hat{f}\left(\boldsymbol{x}^\mu \cdot \hat{\boldsymbol{w}}\right) \tag{1.6}$$

where the weights $\boldsymbol{w} \in \mathbb{R}^p$ are learned by minimising a loss function with a ridge regularisation term:

$$\hat{\boldsymbol{w}} = \min_{\boldsymbol{w}} \left[\sum_{\mu=1}^{n} \ell(y^\mu, \boldsymbol{x}^\mu \cdot \boldsymbol{w}) + \frac{\lambda}{2} ||\boldsymbol{w}||_2^2 \right] . \tag{1.7}$$

for $\lambda > 0$.

It is worth stressing that our results hold for general ℓ, $\hat{f}$ and f^0 - including non-convex loss functions. However, for the purpose of the applications explored in this manuscript, we will be mostly interested in the cases $\hat{f}(x) = f^0(x) = x$ for regression and $\hat{f}(x) = f^0(x) = \text{sign}(x)$ for classification, and we will focus on the following two loss functions:

$$\ell(y^\mu, \boldsymbol{x}^\mu \cdot \boldsymbol{w}) = \begin{cases} \frac{1}{2}(y^\mu - \boldsymbol{x}^\mu \cdot \boldsymbol{w})^2, & \text{square loss} \\ \log\left(1 + e^{-y^\mu(\boldsymbol{x}^\mu \cdot \boldsymbol{w})}\right), & \text{logistic loss} \end{cases} \tag{1.8}$$

Note that these loss functions are strictly convex. Therefore, for these losses, the regularised optimisation problem in (1.7) has a unique solution.

Given a new pair $(\boldsymbol{x}^{\text{new}}, y^{\text{new}})$ drawn independently from the same distribution as $\{(\boldsymbol{x}^\mu, y^\mu)\}_{\mu=1}^n$, we define the success of our fit through the generalisation error, defined as:

$$\epsilon_g = \frac{1}{4^k} \mathbb{E}_{\boldsymbol{x}^{\text{new}}, y^{\text{new}}} \left(y^{\text{new}} - \hat{y}^{\text{new}}\right)^2 \tag{1.9}$$

where $\hat{y}^{\text{new}} = \hat{f}(\boldsymbol{x}^{\text{new}} \cdot \hat{\boldsymbol{w}})$, and for convenience we choose $k = 0$ for the regression tasks and $k = 1$ for the classification task, such that the generalisation error in this case counts misclassification. Note that for a classification problem, the generalisation error is just one minus the classification error.

Similarly, we define the *training loss* on the dataset $\{\boldsymbol{x}^\mu, y^\mu\}_{\mu=1}^n$ as:

$$\epsilon_t = \frac{1}{n} \mathbb{E}_{\{\boldsymbol{x}^\mu, y^\mu\}} \left[\sum_{\mu=1}^{n} \ell\left(y^\mu, \boldsymbol{x}^\mu \cdot \hat{\boldsymbol{w}}\right) + \frac{\lambda}{2} \|\hat{\boldsymbol{w}}\|_2^2 \right] . \tag{1.10}$$

Finally, all the results of this manuscript are derived in the *high-dimensional limit*, also known as *thermodynamic limit* in the physics literature, in which we take $p, d, n \to \infty$ while keeping the ratios $\alpha = n/p$, $\gamma = d/p$ fixed.

2 Replicated gaussian equivalences

In this section we introduce the *replicated Gaussian equivalence* (rGE), a central result we will need for our replica calculation of the generalisation error in Sec. 2.1 of the main body. The rGET is a stronger version of the Gaussian equivalence theorem (GET) that was introduced and proved in [1]. Previously, particular cases of the GET were derived in the context of random matrix theory [3, 4, 5, 6]. The gaussian equivalence has also been stated and used in [7, 8].

2.1 Gaussian equivalence theorem

Let $\mathrm{F} \in \mathbb{R}^{d\times p}$ be a fixed matrix, $\boldsymbol{w}^a \in \mathbb{R}^p$, $1 \leq a \leq r$ be a family of vectors, $\boldsymbol{\theta}^0 \in \mathbb{R}^d$ be a fixed vector and $\sigma : \mathbb{R} \to \mathbb{R}$ be a scalar function acting component-wise on vectors.

Let $\boldsymbol{c} \in \mathbb{R}^d$ be a Gaussian vector $\mathcal{N}(0, \mathrm{I}_d)$. The GET is a statement about the (joint) statistics of the following $r + 1$ random variables

$$\lambda^a = \frac{1}{\sqrt{p}}\boldsymbol{w}^a \cdot \sigma(\boldsymbol{u}) \in \mathbb{R}, \qquad \nu = \frac{1}{\sqrt{d}}\boldsymbol{c} \cdot \boldsymbol{\theta}^0 \in \mathbb{R}. \tag{2.1}$$

in the asymptotic limit where $d, p \to \infty$ with fixed p/d and fixed r. For simplicity, assume that $\sigma(x) = -\sigma(-x)$ is an odd function. Further, suppose that in the previously introduced limit the following two balance conditions hold:

Condition 1:

$$\frac{1}{\sqrt{d}}\sum_{\rho=1}^{d} F_{i\rho}F_{j\rho} = O(1), \tag{2.2}$$

for any ρ.

Condition 2:

$$S^{a_1,\dots,a_k}_{\rho_1,\dots,\rho_q} = \frac{1}{\sqrt{p}}\sum_{i=1}^{p} w_i^{a_1} w_i^{a_2} \cdots w_i^{a_k} F_{i\rho_1}F_{i\rho_2}\cdots F_{i\rho_q} = O(1), \tag{2.3}$$

for any integers $k \geq 0$, $q > 0$, for any choice of indices $\rho_1, \rho_2, \cdots, \rho_q \in \{1, \cdots, d\}$ all distinct from each other, and for any choice of indices $a_1, a_2, \cdots, a_k \in \{1, \cdots, r\}$. Under the aforementioned conditions, the following theorem holds:

Theorem 1. *In the limit $d, p \to \infty$ with fixed p/d, the random variables $\{\lambda^a, \nu\}$ are jointly normal, with zero mean and covariances:*

$$\mathbb{E}\left[\lambda^a\lambda^b\right] = \frac{\kappa_\star^2}{p}\boldsymbol{w}^a \cdot \boldsymbol{w}^b + \frac{\kappa_1^2}{d}\boldsymbol{s}^a \cdot \boldsymbol{s}^b, \qquad \mathbb{E}\left[\nu^2\right] = \frac{1}{d}||\boldsymbol{\theta}^0||^2$$

$$\mathbb{E}\left[\lambda^a\nu\right] = \frac{\kappa_1}{d}\boldsymbol{s}^a \cdot \boldsymbol{\theta}^0 \tag{2.4}$$

where:

$$\boldsymbol{s}^a = \frac{1}{\sqrt{p}}\boldsymbol{F}\boldsymbol{w}^a \in \mathbb{R}^d, \qquad a = 1, \cdots, r \tag{2.5}$$

and

$$\kappa_0 = \mathbb{E}_z\left[\sigma(z)\right], \qquad \kappa_1 = \mathbb{E}_z\left[z\sigma(z)\right], \qquad \kappa_\star^2 = \mathbb{E}_z\left[\sigma(z)^2\right] - \kappa_0^2 - \kappa_1^2 \tag{2.6}$$

where $z \sim \mathcal{N}(0, 1)$.

2.2 Replicated Gaussian equivalence

Note that the GET holds for a fixed family $\{\boldsymbol{w}^a\}_{a=1}^r$ and matrix $\mathrm{F} \in \mathbb{R}^{d\times p}$ satisfying the balance condition from eq. (2.3). In the replica setting, we will need to apply the GET under an average over r samples (refered here as *replicas*) of the Gibbs distribution μ_β, introduced in eq. 14 on the main. We therefore shall require the assumption that the balance condition eq. (2.3) holds for any sample of μ_β. We refer to this stronger version of the GET as the *replicated Gaussian equivalence* (rGE). Although proving this result is out of the scope of the present work, we check its self-consistency extensively with numerical simulations.

3 Replica analysis

In this section we give a full derivation of the result in Sec. 2.1 in the main manuscript for the generalisation error of the problem defined in Sec. 1. Our derivation follows from a Gibbs formulation of the optimisation problem in eq. (1.7) followed by a replica analysis inspired by the toolbox of the statistical physics of disordered systems.

3.1 Gibbs formulation of problem

Given the dataset $\{\boldsymbol{x}^\mu, y^\mu\}_{\mu=1}^n$ defined in Section 1.1, we define the following Gibbs measure over $\mathbb{R}^p$:

$$\mu_\beta(\boldsymbol{w}|\{\boldsymbol{x}^\mu, y^\mu\}) = \frac{1}{\mathcal{Z}_\beta} e^{-\beta\left[\sum\limits_{\mu=1}^n \ell(y^\mu, \boldsymbol{x}^\mu\cdot\boldsymbol{w}) + \frac{\lambda}{2}||\boldsymbol{w}||_2^2\right]} = \frac{1}{\mathcal{Z}_\beta} \underbrace{\prod_{\mu=1}^n e^{-\beta\ell(y^\mu, \boldsymbol{x}^\mu\cdot\boldsymbol{w})}}_{\equiv P_y(\boldsymbol{y}|\boldsymbol{w}\cdot\boldsymbol{x}^\mu)} \underbrace{\prod_{i=1}^p e^{-\frac{\beta\lambda}{2}w_i^2}}_{\equiv P_w(\boldsymbol{w})} \tag{3.1}$$

for $\beta > 0$. When $\beta \to \infty$, the Gibbs measure peaks at the solution of the optimisation problem in eq. (1.7) - which, in the particular case of a strictly convex loss, is unique. Note that in the second equality we defined the factorised distributions P_y and P_w, showing that μ_β can be interpreted as a posterior distribution of $\boldsymbol{w}$ given the dataset $\{\boldsymbol{x}^\mu, y^\mu\}$, with P_y and P_w being the likelihood and prior distributions respectively.

An exact calculation of μ_β is intractable for large values of n, p and d. However, the interest in μ_β is that in the limit $n, p, d \to \infty$ with d/p and n/p fixed, the free energy density associated to the Gibbs measure:

$$f_\beta = -\lim_{p\to\infty} \frac{1}{p}\mathbb{E}_{\{\boldsymbol{x}^\mu, y^\mu\}} \log \mathcal{Z}_\beta \tag{3.2}$$

can be computed exactly using the replica method, and at $\beta \to \infty$ give us the optimal overlaps:

$$q_w = \frac{1}{p}\mathbb{E}||\hat{\boldsymbol{w}}||^2 \qquad\qquad q_x = \frac{1}{d}\mathbb{E}||\mathrm{F}\hat{\boldsymbol{w}}||^2 \qquad\qquad m_x = \frac{1}{d}\mathbb{E}\left[\boldsymbol{\theta}^0 \cdot \mathrm{F}\hat{\boldsymbol{w}}\right] \tag{3.3}$$

that - as we will see - fully characterise the generalisation error defined in eq. (1.9).

3.2 Replica computation of the free energy density

The replica calculation of f_β is based on a large deviation principle for the free energy density. Let

$$f_\beta(\{\boldsymbol{x}^\mu, y^\mu\}) = -\frac{1}{p}\log \mathcal{Z}_\beta \tag{3.4}$$

be the free energy density for one given sample of the problem, i.e. a fixed dataset $\{\boldsymbol{x}^\mu, y^\mu\}_{\mu=1}^n$. We assume that the distribution $P(f)$ of the free energy density, seen as a random variable over different samples of the problem, satisfies a large deviation principle, in the sense that, in the thermodynamic limit:

$$P(f) \simeq e^{p\Phi(f)}\;, \tag{3.5}$$

with Φ a concave function reaching its maximum at the free energy density $f = f_\beta$, with $\Phi(f_\beta) = 0$. This hypothesis includes the notion of *self-averageness* which states that the free-energy density is the same for almost all samples in the thermodynamic limit.

The value of f_β can be computed by computing the *replicated partition function*

$$\mathbb{E}_{\{\boldsymbol{x}^\mu, y^\mu\}}\mathcal{Z}_\beta^r = \int df\; e^{p[\Phi(f) - rf]}\;, \tag{3.6}$$

and taking the limit

$$f_\beta = \lim_{r\to 0^+} \frac{\mathrm{d}}{\mathrm{d}r} \lim_{p\to\infty}\left[-\frac{1}{p}\left(\mathbb{E}_{\{\boldsymbol{x}^\mu, y^\mu\}}\mathcal{Z}_\beta^r\right)\right] \tag{3.7}$$

Although this procedure is not fully rigorous, experience from the statistical physics of disordered systems shows that it gives exact results, and in fact the resulting expression can be verified to match the numerical simulations.

Using the replica method we need to evaluate:

$$\mathbb{E}_{\{\boldsymbol{x}^\mu,y^\mu\}}\mathcal{Z}_\beta^r = \int \mathrm{d}\boldsymbol{\theta}^0\, P_\theta(\boldsymbol{\theta}^0) \int \prod_{a=1}^{r} \mathrm{d}\boldsymbol{w}\, P_w\left(\boldsymbol{w}^a\right) \times \\ \times \prod_{\mu=1}^{n} \int \mathrm{d}y^\mu\, \underbrace{\mathbb{E}_{\boldsymbol{c}^\mu}\left[P_y^0\left(y^\mu | \frac{\boldsymbol{c}^\mu \cdot \boldsymbol{\theta}^0}{\sqrt{d}}\right) \prod_{a=1}^{r} P_y\left(y^\mu | \boldsymbol{w}^a \cdot \sigma\left(\frac{1}{\sqrt{d}}\mathrm{F}^\top \boldsymbol{c}^\mu\right)\right)\right]}_{\text{(I)}} \tag{3.8}$$

where P_w and P_y have been defined in (3.1). In order to compute this quantity, we introduce, for each point μ in the database, the $r+1$ variables

$$\nu_\mu = \frac{1}{\sqrt{d}}\boldsymbol{c}^\mu \cdot \boldsymbol{\theta}^0 \;, \tag{3.9}$$

$$\lambda_\mu^a = \boldsymbol{w}^a \cdot \sigma\left(\frac{1}{\sqrt{d}}\mathrm{F}^\top \boldsymbol{c}^\mu\right) \;. \tag{3.10}$$

Choosing $\boldsymbol{c}^\mu$ at random induces a joint distribution $P(\nu_\mu, \lambda_\mu^a)$. In the thermodynamic limit $p, d \to \infty$ with fixed p/n, and for matrices F satisfying the balance condition in eq. (2.3), the *replicated Gaussian equivalence* introduced in Section 2.2 tells us that, for a given μ, the $r+1$ variables $\{\nu_\mu, \lambda_\mu^a\}_{a=1}^r$ are Gaussian random values with zero mean and covariance given by:

$$\Sigma^{ab} = \begin{pmatrix} \rho & M^a \\ M^a & Q^{ab} \end{pmatrix} \in \mathbb{R}^{(r+1)\times(r+1)} \tag{3.11}$$

The elements of the covariance matrix M^a and Q^{ab} are the rescaled version of the so-called *overlap parameters*:

$$\rho = \frac{1}{d}||\boldsymbol{\theta}^0||^2, \qquad m_s^a = \frac{1}{d}\boldsymbol{s}^a \cdot \boldsymbol{\theta}^0, \qquad q_s^{ab} = \frac{1}{d}\boldsymbol{s}^a \cdot \boldsymbol{s}^b, \qquad q_w^{ab} = \frac{1}{p}\boldsymbol{w}^a \cdot \boldsymbol{w}^b, \tag{3.12}$$

where $\boldsymbol{s}^a = \frac{1}{\sqrt{p}}\mathrm{F}\boldsymbol{w}^a$. They are thus given by:

$$M^a = \kappa_1 m_s^a, \qquad Q^{ab} = \kappa_\star^2 q_w^{ab} + \kappa_1^2 q_s^{ab}. \tag{3.13}$$

where $\kappa_1 = \mathbb{E}_z\left[z\sigma(z)\right]$ and $\kappa_\star^2 = \mathbb{E}_z\left[\sigma(z)^2\right] - \kappa_1^2$ as in eq. (2.6). With this notation, the asymptotic joint probability is simply written as:

$$P(\nu_\mu, \{\lambda_\mu^a\}_{a=1}^r) = \frac{1}{\sqrt{\det\left(2\pi\Sigma\right)}} e^{-\frac{1}{2}\sum\limits_{a,b=0}^{r} z_\mu^a \left(\Sigma^{-1}\right)^{ab} z_\mu^b} \tag{3.14}$$

with $z_\mu^0 = \nu_\mu$ and $z_\mu^a = \lambda_\mu^a$ for $a = 1, \cdots, r$. The average over the replicated partition function (3.8) therefore reads:

$$\mathbb{E}_{\{\boldsymbol{x}^\mu,y^\mu\}}\mathcal{Z}_\beta^r = \int \mathrm{d}\boldsymbol{\theta}^0\, P_\theta(\boldsymbol{\theta}^0) \int \prod_{a=1}^{r} \mathrm{d}\boldsymbol{w}\, P_w\left(\boldsymbol{w}^a\right) \prod_{\mu=1}^{n} \int \mathrm{d}y^\mu \times \\ \times \int \mathrm{d}\nu_\mu\, P_y^0\left(y^\mu|\nu_\mu\right) \int \prod_{a=1}^{r} \mathrm{d}\lambda_\mu^a\, P(\nu_\mu, \{\lambda_\mu^a\}) \prod_{a=1}^{r} P_y\left(y^\mu|\{\lambda_\mu^a\}\right). \tag{3.15}$$

Rewriting as a saddle-point problem

Note that after taking the average over $\boldsymbol{x}$, the integrals involved in the replicated partition function only couple through the overlap parameters. It is therefore useful to introduce the following Dirac

δ-functions to unconstrain them, introducing the decomposition:

$$\begin{aligned}
1 &= d^{-(r+1)^2} \int \mathrm{d}\rho \, \delta\left(d\rho - ||\boldsymbol{\theta}^0||^2\right) \int \prod_{a=1}^{r} \mathrm{d}m_s^a \, \delta\left(dm_s^a - \boldsymbol{s}^a \cdot \boldsymbol{\theta}^{\mathbf{0}}\right) \times \\
&\quad \times \int \prod_{1 \le a \le b \le r} \mathrm{d}q_s^{ab} \delta\left(dq_s^{ab} - \boldsymbol{s}^a \cdot \boldsymbol{s}^b\right) \int \prod_{1 \le a \le b \le r} \mathrm{d}q_w^{ab} \, \delta\left(pq_w^{ab} - \boldsymbol{w}^a \cdot \boldsymbol{w}^b\right) \\
&= d^{-(r+1)^2} \int \frac{\mathrm{d}\rho \mathrm{d}\hat{\rho}}{2\pi} \, e^{-i\hat{\rho}\left(d\rho - ||\boldsymbol{\theta}^0||^2\right)} \int \prod_{a=1}^{r} \frac{\mathrm{d}m_s^a \mathrm{d}\hat{m}_s^a}{2\pi} \, e^{-i\sum_{a=1}^{r} \hat{m}_s^a\left(dm_s^a - \boldsymbol{s}^a \cdot \boldsymbol{\theta}^{\mathbf{0}}\right)} \times \\
&\quad \times \int \prod_{1 \le a \le b \le r} \frac{\mathrm{d}q_s^{ab} \mathrm{d}\hat{q}_s^{ab}}{2\pi} e^{-i\sum_{1 \le a \le b \le r} \hat{q}_s^{ab}\left(dq_s^{ab} - \boldsymbol{s}^a \cdot \boldsymbol{s}^b\right)} \int \prod_{1 \le a \le b \le r} \frac{\mathrm{d}q_w^{ab} \hat{q}_w^{ab}}{2\pi} \, e^{-i\sum_{1 \le a \le b \le r} \hat{q}_w^{ab}\left(pq_w^{ab} - \boldsymbol{w}^a \cdot \boldsymbol{w}^b\right)}.
\end{aligned} \tag{3.16}$$

Introducing the above in eq. (3.15) and exchanging the integration order allows to factorise the integrals over the d, p, n dimensions and rewrite:

$$\mathbb{E}_{\{\boldsymbol{x}^\mu, y^\mu\}} \mathcal{Z}_\beta^r = \int \frac{\mathrm{d}\rho \mathrm{d}\hat{\rho}}{2\pi} \int \prod_{a=1}^{r} \frac{\mathrm{d}m_s^a \mathrm{d}\hat{m}_s^a}{2\pi} \int \prod_{1 \le a \le b \le r} \frac{\mathrm{d}q_s^{ab} \mathrm{d}\hat{q}_s^{ab}}{2\pi} \frac{\mathrm{d}q_w^{ab} \mathrm{d}\hat{q}_w^{ab}}{2\pi} e^{p\Phi^{(r)}} \tag{3.17}$$

where the integrals over the variables m_s^a, q_s^{ab} and q_w^{ab} run over $\mathbb{R}$, while those over $\hat{m}_s^a$, $\hat{q}_s^{ab}$ and $\hat{q}_w^{ab}$ run over $i\mathbb{R}$. The function $\Phi^{(r)}$, a function of all the overlap parameters, is given by:

$$\begin{aligned}
\Phi^{(r)} = &-\gamma\rho\hat{\rho} - \gamma \sum_{a=1}^{r} m_s^a \hat{m}_s^a - \sum_{1 \le a \le b \le r} \left(\gamma q_s^{ab} \hat{q}_s^{ab} + q_w \hat{q}_w\right) + \alpha \Psi_y^{(r)}\left(\rho, m_s^a, q_s^{ab}, q_w^{ab}\right) \\
&+ \Psi_w^{(r)}\left(\hat{\rho}, \hat{m}_s^a, \hat{q}_s^{ab}, \hat{q}_w^{ab}\right)
\end{aligned} \tag{3.18}$$

where we recall that $\alpha = n/p$, $\gamma = d/p$, and we have introduced:

$$\begin{aligned}
\Psi_y^{(r)} &= \log \int \mathrm{d}y \int \mathrm{d}\nu \, P_y^0\left(y|\nu\right) \int \prod_{a=1}^{r} \left[\mathrm{d}\lambda^a P_y\left(y|\lambda^a\right)\right] P(\nu, \{\lambda^a\}) \\
\Psi_w^{(r)} &= \frac{1}{p} \log \int \mathrm{d}\boldsymbol{\theta}^0 P_\theta(\boldsymbol{\theta}^0) e^{-\hat{\rho}||\boldsymbol{\theta}^0||^2} \int \prod_{a=1}^{r} \mathrm{d}\boldsymbol{w}^a \, P_w(\boldsymbol{w}^a) e^{\sum_{1 \le a \le b \le r}\left[\hat{q}_w^{ab} \boldsymbol{w}^a \cdot \boldsymbol{w}^b + \hat{q}_s^{ab} \boldsymbol{s}^a \cdot \boldsymbol{s}^b\right] - \sum_{a=1}^{r} \hat{m}_s^a \boldsymbol{s}^a \cdot \boldsymbol{\theta}^0}
\end{aligned} \tag{3.19}$$

Note that $\boldsymbol{s}^a = \frac{1}{\sqrt{p}} \mathrm{F} \boldsymbol{w}^a$ is a function of $\boldsymbol{w}^a$, and must be kept under the $\boldsymbol{w}^a$ integral. In the thermodynamic limit where $p \to \infty$ with n/p and d/p fixed, the integral in eq. (3.17) concentrates around the values of the overlap parameters that extremize $\Phi^{(r)}$, and therefore

$$f = -\lim_{r \to 0^+} \frac{1}{r} \underset{\substack{\{\rho, \hat{\rho}, m_s^a, \hat{m}_s^a\} \\ \{q_s^{ab}, \hat{q}_s^{ab}, q_w^{ab}, \hat{q}_w^{ab}\}}}{\mathbf{extr}} \Phi^{(r)}. \tag{3.20}$$

Replica symmetric Ansatz

In order to proceed with the $r \to 0^+$ limit, we restrict the extremization above to the following replica symmetric Ansatz:

$$\begin{aligned}
m_s^a &= m_s & \hat{m}^a &= \hat{m}_s & &\text{for } a = 1, \cdots, r \\
q_{s/w}^{aa} &= r_{s/w} & \hat{q}_{s/w}^{aa} &= -\frac{1}{2}\hat{r}_{s/w} & &\text{for } a = 1, \cdots, r \\
q_{s/w}^{ab} &= q_{s/w} & \hat{q}_{s/w}^{ab} &= \hat{q}_{s/w} & &\text{for } 1 \le a < b \le r
\end{aligned} \tag{3.21}$$

Note that, in the particular case of a convex loss function with $\lambda > 0$, the replica symmetric Ansatz is justified: the problem only admitting one solution, it *a fortiori* coincides with the replica symmetric one. For non-convex losses, solutions that are not replica symmetric (also known as *replica symmetry* *breaking*) are possible, and the energy landscape of the free energy needs to be carefully analysed. In the practical applications explored in this manuscript, we focus on convex losses with ridge regularisation, and therefore the replica symmetric assumption is fully justified.

Before proceeding with the limit in eq. (3.20), we need to verify that the above Ansatz is well defined - in other words, that we have not introduced a spurious order one term in Φ that would diverge. This means we need to check that $\lim_{r\to 0^+} \Phi = 0$.

First, with a bit of algebra one can check that, within our replica symmetric Ansatz:

$$\lim_{r\to 0^+} \Psi_y^{(r)} = 0. \tag{3.22}$$

Therefore,

$$\lim_{r\to 0^+} \Phi^{(r)} = -\gamma\rho\hat{\rho} + \gamma \log \int_{\mathbb{R}} \mathrm{d}\theta^0 \, P_\theta\left(\theta^0\right) e^{\hat{\rho}{\theta^0}^2} \tag{3.23}$$

where we have used the fact that P_θ is a factorised distribution to take the $p \to \infty$ limit. In order for this limit to be 0, we need that $\hat{\rho} = 0$, which also fixes ρ to be a constant given by the second moment of θ^0:

$$\rho = \mathbb{E}_{\theta^0}\left[{\theta^0}^2\right] \tag{3.24}$$

We now proceed with the limit in eq. (3.20). Let's look first at Ψ_y. The non-trivial limit comes from the fact that $\det\Sigma$ and Σ^{-1} are non-trivial functions of r. It is not hard to see, however, that Σ^{-1} itself has replica symmetric structure, with components given by:

$$\begin{aligned}
\left(\Sigma^{-1}\right)^{00} &= \tilde{\rho} = \frac{R+(r-1)Q}{\rho(R+(r-1)Q) - rM^2}, & \left(\Sigma^{-1}\right)^{aa} &= \tilde{R} = \frac{\rho R + (r-2)\rho\, Q - (r-1)M^2}{(R-Q)(\rho\, R + (r-1)\rho\, Q - r\, M^2)} \\
\left(\Sigma^{-1}\right)^{a0} &= \tilde{M} = \frac{M}{r\, M^2 - \rho\, R - (r-1)\rho\, Q}, & \left(\Sigma^{-1}\right)^{ab} &= \tilde{Q} = \frac{M^2 - \rho\, Q}{(R-Q)(\rho\, R + (r-1)\rho\, Q - r\, M^2)}
\end{aligned} \tag{3.25}$$

where M, Q and R are the rescaled overlap parameters in the replica symmetric Ansatz, that is:

$$M = \kappa_1 m_s, \qquad Q = \kappa_\star^2 q_w + \kappa_1^2 q_s, \qquad R = \kappa_\star^2 r_w + \kappa_1^2 r_s. \tag{3.26}$$

This allows us to write:

$$\begin{aligned}
\Psi_y^{(r)} = \log \int \mathrm{d}y \int \mathrm{d}\nu \, P_y^0\left(y|\nu\right) e^{-\frac{\tilde{\rho}}{2}\nu^2} \int \prod_{a=1}^{r} \mathrm{d}\lambda^a P_y\left(y|\lambda^a\right) e^{-\frac{\tilde{Q}}{2}\sum\limits_{a,b=1}^{n}\lambda^a\lambda^b - \frac{\tilde{R}-\tilde{Q}}{2}\sum\limits_{a=1}^{r}(\lambda^a)^2 - \tilde{M}\nu\sum\limits_{a=1}^{n}\lambda^a} \\
- \frac{1}{2}\log\det\left(2\pi\Sigma\right).
\end{aligned} \tag{3.27}$$

In order to completely factor the integral above in the replica space, we use the *Hubbard-Stratonovich transformation*:

$$e^{-\frac{\tilde{Q}}{2}\sum\limits_{a,b=1}^{r}\lambda^a\lambda^b} = \mathbb{E}_\xi e^{\sqrt{-\tilde{Q}}\xi\sum\limits_{a=1}^{r}\lambda^a} \tag{3.28}$$

for $\xi \sim \mathcal{N}(0,1)$, such that

$$\begin{aligned}
\Psi_y^{(r)} = \mathbb{E}_\xi \log \int \mathrm{d}y \int \mathrm{d}\nu \, P_y^0\left(y|\nu\right) e^{-\frac{\tilde{\rho}}{2}\nu^2} \left[\int \mathrm{d}\lambda P_y\left(y|\lambda\right) e^{-\frac{\tilde{R}-\tilde{Q}}{2}\lambda^2 + \left(\sqrt{-\tilde{Q}}\xi - \tilde{M}\nu\right)\lambda}\right]^r \\
- \frac{1}{2}\log\det\left(2\pi\Sigma\right).
\end{aligned} \tag{3.29}$$

Taking into account the r dependence of the inverse elements and of the determinant, we can take the limit to get:

$$\begin{aligned}
\lim_{r\to 0^+} \frac{1}{r}\Psi_y^{(r)} = \mathbb{E}_\xi \int_{\mathbb{R}} \mathrm{d}y \int \frac{\mathrm{d}\nu}{\sqrt{2\pi\rho}} P_y^0\left(y|\nu\right) e^{-\frac{1}{2\rho}\nu^2} \log \int \frac{\mathrm{d}\lambda}{\sqrt{2\pi}} \, P_y\left(y|\lambda\right) e^{-\frac{1}{2}\frac{\lambda^2}{R-Q} + \left(\frac{\sqrt{Q-M^2/\rho}}{R-Q}\xi + \frac{M/\rho}{R-Q}\nu\right)\lambda} \\
- \frac{1}{2}\log\left(R-Q\right) - \frac{1}{2}\frac{Q}{R-Q}
\end{aligned} \tag{3.30}$$

Finally, making a change of variables and defining:

$$\mathcal{Z}_y^{\cdot/0}(y;\omega,V) = \int \frac{\mathrm{d}x}{\sqrt{2\pi V}} e^{-\frac{1}{2V}(x-\omega)^2} P_y^{\cdot/0}\left(y|x\right) \tag{3.31}$$

allows us to rewrite the limit of Ψ_y - which abusing notation we still denote Ψ_y - as:

$$\Psi_y = \mathbb{E}_\xi\left[\int_{\mathbb{R}} \mathrm{d}y\, \mathcal{Z}_y^0\left(y; \frac{M}{\sqrt{Q}}\xi, \rho\, - \frac{M^2}{Q}\right) \log \mathcal{Z}_y\left(y; \sqrt{Q}\xi, R-Q\right)\right]. \tag{3.32}$$

One can follow a very similar approach for the limit of Ψ_w, although in this case the limit is much simpler, since there is no r dependence on the hat variables. The limit can be written as:

$$\Psi_w = \lim_{p\to\infty} \frac{1}{p}\mathbb{E}_{\xi,\eta,\theta^0} \log \int_{\mathbb{R}^d} \mathrm{d}\boldsymbol{s}\, P_s(\boldsymbol{s};\eta) e^{-\frac{\hat{V}_s}{2}||\boldsymbol{s}||^2 + \left(\sqrt{\hat{q}_s}\xi\mathbf{1}_d + \hat{m}_s\boldsymbol{\theta}^0\right)^\top \boldsymbol{s}} \tag{3.33}$$

for $\xi, \eta \sim \mathcal{N}(0,1)$, and we have defined:

$$P_s(\boldsymbol{s};\eta) = \int_{\mathbb{R}^p} \mathrm{d}\boldsymbol{w}\, P_w(\boldsymbol{w}) e^{-\frac{\hat{V}_w}{2}||\boldsymbol{w}||^2 + \sqrt{\hat{q}_w}\eta\mathbf{1}_p^\top\boldsymbol{w}} \delta\left(\boldsymbol{s} - \frac{1}{\sqrt{p}}\mathrm{F}\boldsymbol{w}\right) \tag{3.34}$$

and we have defined the shorthands $\hat{V}_w = \hat{r}_w + \hat{q}_w$ and $\hat{V}_s = \hat{r}_s + \hat{q}_s$.

Summary of the replica symmetric free energy density

Summarising the calculation above, the replica symmetric free energy density reads:

$$\begin{aligned} f = \textbf{extr}\Big\{ &-\frac{\gamma}{2}r_s\hat{r}_s - \frac{\gamma}{2}q_s\hat{q}_s + \gamma m_s\hat{m}_s - \frac{1}{2}r_w\hat{r}_w - \frac{1}{2}q_w\hat{q}_w \\ &- \alpha\Psi_y(R,Q,M) - \Psi_w\left(\hat{r}_s,\hat{q}_s,\hat{m}_s,\hat{r}_w,\hat{q}_w\right)\Big\} \end{aligned} \tag{3.35}$$

with $\alpha = \frac{n}{p}$, $\gamma = \frac{d}{p}$, and:

$$Q = \kappa_1^2 q_s + \kappa_\star^2 q_w, \qquad R = \kappa_1^2 r_s + \kappa_\star^2 r_w \qquad M = \kappa_1 m_s. \tag{3.36}$$

The so-called potentials (Ψ_y, Ψ_w) are given by:

$$\Psi_w = \lim_{p\to\infty} \frac{1}{p}\mathbb{E}_{\xi,\eta,\theta^0} \log \int_{\mathbb{R}^d} \mathrm{d}\boldsymbol{s} P_s(\boldsymbol{s};\eta) e^{-\frac{\hat{V}_s}{2}||\boldsymbol{s}||^2 + \left(\sqrt{\hat{q}_s}\xi\mathbf{1}_d + \hat{m}_s\boldsymbol{\theta}^0\right)^\top \boldsymbol{s}} \tag{3.37}$$

$$\Psi_y = \mathbb{E}_\xi\left[\int_{\mathbb{R}} \mathrm{d}y\, \mathcal{Z}_y^0\left(y; \frac{M}{\sqrt{Q}}\xi, \rho\, - \frac{M^2}{Q}\right) \log \mathcal{Z}_y\left(y; \sqrt{Q}\xi, R-Q\right)\right]. \tag{3.38}$$

where:

$$\begin{aligned} P_s(\boldsymbol{s};\eta) &= \int_{\mathbb{R}^p} \mathrm{d}\boldsymbol{w}\, P_w(\boldsymbol{w}) e^{-\frac{\hat{V}_w}{2}||\boldsymbol{w}||^2 + \sqrt{\hat{q}_w}\eta\mathbf{1}_p^\top\boldsymbol{w}} \delta\left(\boldsymbol{s} - \frac{1}{\sqrt{p}}\mathrm{F}\boldsymbol{w}\right) \\ \mathcal{Z}_y^{\cdot/0}(y;\omega,V) &= \int \frac{\mathrm{d}x}{\sqrt{2\pi V}} e^{-\frac{1}{2V}(x-\omega)^2} P_y^{\cdot/0}\left(y|x\right) \end{aligned} \tag{3.39}$$

3.3 Evaluating Ψ_w for ridge regularisation and Gaussian prior

Note that as long as the limit in Ψ_w is well defined, the eq. (3.35) holds for any P_θ and P_w. However, as discussed in Sec. 1.1, we are interested in $\boldsymbol{\theta}^0 \sim \mathcal{N}(0, \mathrm{I}_d)$ and ridge regularisation so that $P_w = \exp\left(-\frac{\beta\lambda}{2}||\boldsymbol{w}||^2\right)$. In this case, we simply have:

$$P(\boldsymbol{s};\eta) = \frac{e^{\frac{p}{2}\frac{\eta^2\hat{q}_w}{\beta\lambda+\hat{V}_w}}}{(\beta\lambda+\hat{V}_w)^{p/2}}\mathcal{N}(\boldsymbol{s};\boldsymbol{\mu},\Sigma) \tag{3.40}$$

with:

$$\boldsymbol{\mu} = \frac{\sqrt{\hat{q}_w}\eta}{\beta\lambda+\hat{V}_w}\frac{\mathrm{F}\mathbf{1}_p}{\sqrt{p}} \in \mathbb{R}^d, \qquad \Sigma = \frac{1}{\beta\lambda+\hat{V}_w}\frac{\mathrm{F}\mathrm{F}^\top}{p} \in \mathbb{R}^{d\times d}. \tag{3.41}$$

Therefore the argument of the logarithm in Ψ_w is just another Gaussian integral we can do explicitly:

$$\mathbb{E}_s e^{-\frac{\hat{V}_s}{2}||\boldsymbol{s}||^2+\boldsymbol{b}^\top \boldsymbol{s}} = \frac{e^{\frac{p}{2}\frac{\eta^2\hat{q}_w}{\beta\lambda+\hat{V}_w}}}{\left(\beta\lambda+\hat{V}_w\right)^{p/2}} \frac{e^{-\frac{1}{2}\boldsymbol{\mu}^\top\Sigma^{-1}\boldsymbol{\mu}+\frac{1}{2\hat{V}_s}||\boldsymbol{b}+\Sigma^{-1}\boldsymbol{\mu}||^2}}{\sqrt{\det\left(\mathrm{I}_d+\hat{V}_s\Sigma\right)}} e^{-\frac{1}{2\hat{V}_s}\left(b+\Sigma^{-1}\mu\right)^\top\left(\mathrm{I}_d+\hat{V}_s\Sigma\right)^{-1}\left(b+\Sigma^{-1}\mu\right)} \tag{3.42}$$

where we have defined the shorthand $\boldsymbol{b} = \left(\sqrt{\hat{q}_s}\xi\mathbf{1}_d + \hat{m}_s\boldsymbol{\theta}^0\right) \in \mathbb{R}^d$. Inserting back in eq. (3.37) and taking the log,

$$\begin{aligned}\Psi_w = \lim_{p\to\infty} \mathbb{E}_{\theta^0,\xi,\eta}\Bigg[&\frac{1}{2}\frac{\eta^2\hat{q}_w}{\beta\lambda+\hat{V}_w} - \frac{1}{2}\log\left(\beta\lambda+\hat{V}_w\right) - \frac{1}{2p}\operatorname{tr}\log\left(\mathrm{I}_d+\hat{V}_s\Sigma\right) - \frac{1}{2p}\boldsymbol{\mu}^\top\Sigma^{-1}\boldsymbol{\mu} \\ &+\frac{1}{2p\hat{V}_s}||\boldsymbol{b}+\Sigma^{-1}\boldsymbol{\mu}||^2 - \frac{1}{2p\hat{V}_s}\left(b+\Sigma^{-1}\mu\right)^\top\left(\mathrm{I}_d+\hat{V}_s\Sigma\right)^{-1}\left(b+\Sigma^{-1}\mu\right)\Bigg]\end{aligned} \tag{3.43}$$

The averages over $\eta, \xi, \boldsymbol{\theta}^0$ simplify this expression considerably:

$$\begin{aligned}\mathbb{E}_\eta\left[\boldsymbol{\mu}^\top\Sigma^{-1}\boldsymbol{\mu}\right] &= \frac{1}{p}\frac{\hat{q}_w}{(\beta\lambda+\hat{V}_w)^2}\left(\mathbf{F}\mathbf{1}_p\right)^\top\Sigma^{-1}\left(\mathbf{F}\mathbf{1}_p\right) = d\frac{\hat{q}_w}{\beta\lambda+\hat{V}_w} \\ \mathbb{E}_{\eta,\xi,\theta^0}||b+\Sigma^{-1}\boldsymbol{\mu}||^2 &= d(\hat{m}_s^2+\hat{q}_s) + \frac{1}{p}\hat{q}_w\operatorname{tr}\left(\mathrm{FF}^\top\right)^{-1} \\ \mathbb{E}_{\eta,\xi,\theta^0}\left(b+\Sigma^{-1}\boldsymbol{\mu}\right)^\top\left(\mathrm{I}_d+\hat{V}_s\Sigma\right)^{-1}\left(b+\Sigma^{-1}\boldsymbol{\mu}\right) &= \frac{1}{p}\hat{q}_w\operatorname{tr}\left[\mathrm{FF}^\top\left(\mathrm{I}_d+\hat{V}_s\Sigma\right)^{-1}\right] \\ &\quad+(\hat{m}_s^2+\hat{q}_s)\operatorname{tr}\left(\mathrm{I}_d+\hat{V}_s\Sigma\right)^{-1}\end{aligned} \tag{3.44}$$

Finally, we can combine the two terms:

$$\operatorname{tr}\frac{\mathrm{FF}^\top}{p} + \operatorname{tr}\left[\frac{\mathrm{FF}^\top}{p}\left(\mathrm{I}_d+\hat{V}_s\Sigma\right)^{-1}\right] = \frac{\hat{V}_s}{\beta\lambda+\hat{V}_w}\operatorname{tr}\left(\mathrm{I}_d+\hat{V}_s\Sigma\right)^{-1}, \tag{3.45}$$

and write:

$$\begin{aligned}\Psi_w = &-\frac{1}{2}\log\left(\beta\lambda+\hat{V}_w\right) - \frac{1}{2}\lim_{p\to\infty}\frac{1}{p}\operatorname{tr}\log\left(\mathrm{I}_d+\frac{\hat{V}_s}{\beta\lambda+\hat{V}_w}\frac{\mathrm{FF}^\top}{p}\right) \\ &+\frac{\hat{m}_s^2+\hat{q}_s}{2\hat{V}_s}\left[\gamma - \lim_{p\to\infty}\frac{1}{p}\operatorname{tr}\left(\mathrm{I}_d+\frac{\hat{V}_s}{\beta\lambda+\hat{V}_w}\frac{\mathrm{FF}^\top}{p}\right)^{-1}\right] \\ &+\frac{1}{2}\frac{\hat{q}_w}{\beta\lambda+\hat{V}_w}\left[1-\gamma+\lim_{p\to\infty}\frac{1}{p}\operatorname{tr}\left(\mathrm{I}_d+\frac{\hat{V}_s}{\beta\lambda+\hat{V}_w}\frac{\mathrm{FF}^\top}{p}\right)^{-1}\right]\end{aligned} \tag{3.46}$$

Note that Ψ only depends on the spectral properties of the matrix $\frac{1}{p}\mathrm{FF}^\top \in \mathbb{R}^{d\times d}$, and more specifically on its resolvent in the asymptotic limit. A case of particular interest is when $\mathrm{FF}^\top$ has a well defined spectral measure μ on the $p, d \to \infty$ limit with $\gamma = d/p$ fixed. In that case, we can write:

$$\lim_{p\to\infty}\frac{1}{p}\operatorname{tr}\left(\mathrm{I}_d+\frac{\hat{V}_s}{\beta\lambda+\hat{V}_w}\frac{\mathrm{FF}^\top}{p}\right)^{-1} = \gamma\frac{\beta\lambda+\hat{V}_w}{\hat{V}_s}g_\mu\left(-\frac{\beta\lambda+\hat{V}_w}{\hat{V}_s}\right) \tag{3.47}$$

$$\tag{3.48}$$

where g_μ is the Stieltjes transform of μ, defined by:

$$g_\mu(z) = \int\frac{\mathrm{d}\mu(t)}{t-z}. \tag{3.49}$$

Similarly, the logarithm term can be expressed as the logarithm potential of μ - although for the purpose of evaluating the generalisation error we will only need the derivative of these terms, and therefore only the Stieltjes transforms and its derivative.

In what follows, we will mostly focus on two kinds of projection matrices F:

Gaussian projections: For $\mathrm{F} \in \mathbb{R}^{d\times p}$ a random matrix with i.i.d. Gaussian entries with zero mean and variance 1, μ is given by the well-known Marchenko-Pastur law, and the corresponding Stieltjes transform is given by:

$$g_\mu(z) = \frac{1-z-\gamma-\sqrt{(z-1-\gamma)^2-4\gamma}}{2z\gamma}, \qquad z<0 \tag{3.50}$$

Orthogonally invariant projection: For $\mathrm{F} = \mathrm{U}^\top \mathrm{D} \mathrm{V}$ with $\mathrm{U} \in \mathbb{R}^{d\times d}$ and $\mathrm{V} \in \mathbb{R}^{p\times p}$ two orthogonal matrices and $\mathrm{D} \in \mathbb{R}^{d\times p}$ a rectangular diagonal matrix of rank $\min(d,p)$ and diagonal entries d_k, the empirical spectral density μ_p is given by:

$$\mu_d(\lambda) = \frac{1}{d}\sum_{k=1}^{\min(r,p)} \delta(\lambda-\lambda_k) = \left(1-\min\left(1,\frac{1}{\gamma}\right)\right)\delta(\lambda) + \frac{1}{p}\sum_{k=1}^{\min(d,p)} \delta(\lambda-d_k^2) \tag{3.51}$$

Therefore the choice of diagonal elements d_k fully characterise the spectrum of $\mathrm{F}\mathrm{F}^\top$. In order for the orthogonally invariant case to be comparable to the Gaussian case, we fix d_k in such a way that the projected vector $\mathrm{F}\boldsymbol{w}$ is of the same order in both cases, i.e.

$$d_k^2 = \begin{cases} \gamma & \text{for } \gamma > 1 \\ 1 & \text{for } \gamma \leq 1 \end{cases} \tag{3.52}$$

With this choice, the Stieltjes transform of μ reads:

$$g_\mu(z) = \begin{cases} -(1-\frac{1}{\gamma})\frac{1}{z} + \frac{1}{\gamma}\frac{1}{\gamma-z} & \text{for } \gamma > 1 \\ \frac{1}{1-z} & \text{for } \gamma \leq 1 \end{cases} \tag{3.53}$$

3.4 Gaussian equivalent model

It is interesting to note that the average over the dataset $\{\boldsymbol{x}^\mu, y^\mu\}_{\mu=1}^n$ of the replicated partition function $\mathcal{Z}_\beta^r$ in eq. (3.15), obtained after the application of the GET, is identical to the replicated partition function of the same task over the following dual dataset $\{\tilde{\boldsymbol{x}}^\mu, y^\mu\}_{\mu=1}^n$, where:

$$\tilde{\boldsymbol{x}}^\mu = \kappa_0 \mathbf{1}_p + \kappa_1 \frac{1}{\sqrt{d}} \mathrm{F}^\top \boldsymbol{c}^\mu + \kappa_\star \boldsymbol{z}^\mu \tag{3.54}$$

where $\boldsymbol{z}^\mu \sim \mathcal{N}(\mathbf{0}, \mathrm{I}_p)$, and the labels $y^\mu \sim P_y$ are the same. Indeed, calling $\tilde{\mathcal{Z}}_\beta^r$ the replicated partition function for this equivalent dataset, and considering κ_0 we have:

$$\begin{aligned} \mathbb{E}_{\{\tilde{\boldsymbol{x}}^\mu, y^\mu\}} \tilde{\mathcal{Z}}_\beta^r = \int \mathrm{d}\boldsymbol{\theta}^0 \, P_\theta(\boldsymbol{\theta}^0) \int \prod_{a=1}^r \mathrm{d}\boldsymbol{w} \, P_w\left(\boldsymbol{w}^a\right) \times \\ \times \prod_{\mu=1}^n \int \mathrm{d}y^\mu \, \underbrace{\mathbb{E}_{\boldsymbol{c}^\mu, \boldsymbol{z}^\mu}\left[P_y^0\left(y^\mu | \frac{\boldsymbol{c}^\mu \cdot \boldsymbol{\theta}^0}{\sqrt{d}}\right) \prod_{a=1}^r P_y\left(y^\mu | \boldsymbol{w}^a \cdot \left(\frac{\kappa_1}{\sqrt{d}} \mathrm{F}^\top \boldsymbol{c}^\mu + \kappa_\star \boldsymbol{z}^\mu\right)\right)\right]}_{\text{(I)}}. \end{aligned} \tag{3.55}$$

Rewriting (I):

$$\begin{aligned} \text{(I)} = \int \mathrm{d}\nu_\mu \, P_y^0\left(y^\mu | \nu_\mu\right) \int \prod_{a=1}^r \mathrm{d}\lambda_\mu^a \, P_y\left(y^\mu | \lambda_\mu^a\right) \times \\ \times \underbrace{\mathbb{E}_{\boldsymbol{c}^\mu, \boldsymbol{z}^\mu}\left[\delta\left(\nu_\mu - \frac{1}{\sqrt{d}} \boldsymbol{c}^\mu \cdot \boldsymbol{\theta}^0\right) \prod_{a=1}^r \delta\left(\lambda_\mu^a - \frac{\kappa_1}{\sqrt{d}} \boldsymbol{w}^a \cdot \mathrm{F}^\top \boldsymbol{c}^\mu + \kappa_\star \boldsymbol{w}^a \cdot \boldsymbol{z}^\mu\right)\right]}_{\equiv P(\nu, \lambda)}. \end{aligned} \tag{3.56}$$

It is easy to show that taking (κ_0, κ_1) to match those from eq. (2.6), the variables $\left(\nu_\mu, \{\lambda_\mu^a\}\right)$ are jointly Gaussian variables with correlation matrix given by Σ exactly as in eq. (3.11). This establishes the equivalence

$$\tilde{\mathcal{Z}}_\beta^r = \mathcal{Z}_\beta^r \tag{3.57}$$

from which follows the equivalence between the asymptotic generalisation and test error of these two models.

4 Saddle-point equations and the generalisation error

The upshot of the replica analysis is to exchange the p-dimensional minimisation problem for $\boldsymbol{w} \in \mathbb{R}^p$ in eq. (1.7) for a one-dimensional minimisation problem for the parameters $\{r_s, q_s, m_s, r_w, q_w\}$ and their conjugate in eq. (3.35). In particular, note that by construction at the limit $\beta \to \infty$ the solution $\{q_s^\star, m_s^\star, q_w^\star\}$ of eq. (3.35) corresponds to:

$$q_w^\star = \frac{1}{p}||\hat{\boldsymbol{w}}||^2 \qquad\qquad q_s^\star = \frac{1}{d}||\mathrm{F}\hat{\boldsymbol{w}}||^2 \qquad\qquad m_s^\star = \frac{1}{d}\left(\mathrm{F}\hat{\boldsymbol{w}}\right)\cdot\boldsymbol{\theta}^0 \tag{4.1}$$

where $\hat{\boldsymbol{w}}$ is the solution of the solution of eq. (1.7). As we will see, both the generalisation error defined in eq. (1.9) and the training loss can be expressed entirely in terms of these overlap parameters.

4.1 Generalisation error as a function of the overlaps

Let $\{\boldsymbol{x}^{\text{new}}, y^{\text{new}}\}$ be a new sample independently drawn from the same distribution of our data $\{\boldsymbol{x}^\mu, y^\mu\}_{\mu=1}^n$. The generalisation error can then be written as:

$$\begin{aligned}\epsilon_g &= \frac{1}{4^k}\mathbb{E}_{\boldsymbol{x}^{\text{new}},y^{\text{new}}}\left(y^{\text{new}} - \hat{f}\left(\sigma\left(\mathrm{F}^\top\boldsymbol{c}^{\text{new}}\right)\cdot\hat{\boldsymbol{w}}\right)\right)^2\\ &= \frac{1}{4^k}\int \mathrm{d}y\int \mathrm{d}\nu\, P_y^0(y|\nu)\int\mathrm{d}\lambda\,(y-\hat{f}(\lambda))^2\mathbb{E}_{\boldsymbol{c}^{\text{new}}}\left[\delta\left(\nu-\boldsymbol{c}^{\text{new}}\cdot\boldsymbol{\theta}^0\right)\delta\left(\lambda-\sigma\left(\mathrm{F}^\top\boldsymbol{c}^{\text{new}}\right)\cdot\hat{\boldsymbol{w}}\right)\right].\end{aligned} \tag{4.2}$$

where for convenience, we normalise $k = 0$ for the regression task and $k = 1$ for the classification task. Again, we apply the GET from Sec. 2 to write the joint distribution over $\{\nu, \lambda\}$:

$$P(\nu,\lambda) = \frac{1}{\sqrt{\det\left(2\pi\Sigma\right)}}e^{-\frac{1}{2}\boldsymbol{z}^\top\Sigma^{-1}\boldsymbol{z}}, \tag{4.3}$$

where $\boldsymbol{z} = (\nu, \lambda)^\top \in \mathbb{R}^2$ and Σ is given by

$$\Sigma = \begin{pmatrix}\rho & M^\star\\ M^\star & Q^\star\end{pmatrix}, \quad \rho = \frac{1}{d}||\boldsymbol{\theta^0}||^2 \quad M^\star = \frac{\kappa_1}{d}\left(\mathrm{F}\hat{\boldsymbol{w}}\right)\cdot\boldsymbol{\theta}^0, \quad Q^\star = \frac{\kappa_1^2}{d}||\mathrm{F}\hat{\boldsymbol{w}}||^2 + \frac{\kappa_\star^2}{p}||\hat{\boldsymbol{w}}||^2. \tag{4.4}$$

Inserting in eq. (4.2) gives the desired representation of the generalisation error in terms of the optimal overlap parameters:

$$\epsilon_g = \frac{1}{4^k}\int\mathrm{d}y\int\mathrm{d}\nu\,P_y^0(y|\nu)\int\mathrm{d}\lambda\,P(\nu,\lambda)(y-\hat{f}(\lambda))^2 \tag{4.5}$$

For linear labels $y = \boldsymbol{c}\cdot\boldsymbol{\theta}^0$ in the regression problem, we simply have:

$$\epsilon_g = \rho + Q^\star - 2M^\star \tag{4.6}$$

while for the corresponding classification problem with $y = \text{sign}\left(\boldsymbol{c}\cdot\boldsymbol{\theta}^0\right)$:

$$\epsilon_g = \frac{1}{\pi}\cos^{-1}\left(\frac{M^\star}{\sqrt{Q^\star}}\right) \tag{4.7}$$

which, as expected, only depend on the angle between $\mathrm{F}\hat{\boldsymbol{w}}$ and $\boldsymbol{\theta}^0$.

4.2 Training loss

Similarly to the generalisation error, the asymptotic of the training loss, defined for the training data $\{\boldsymbol{x}^\mu, y^\mu\}_{\mu=1}^n$ as:

$$\epsilon_t = \frac{1}{n}\mathbb{E}_{\{\boldsymbol{x}^\mu,y^\mu\}}\left[\sum_{\mu=1}^n \ell\left(y^\mu, \boldsymbol{x}^\mu\cdot\hat{\boldsymbol{w}}\right) + \frac{\lambda}{2}||\hat{\boldsymbol{w}}||_2^2\right], \tag{4.8}$$

can also be written only in terms of the overlap parameters. Indeed, it is closely related to the free energy density defined in eq. (3.2). A close inspection on this definition tells us that:

$$\lim_{n\to\infty}\epsilon_t = \lim_{\beta\to\infty}\partial_\beta f_\beta. \tag{4.9}$$

Taking the derivative of the free energy with respect to the parameter β and recalling that $p = \alpha n$, we can then get:

$$\lim_{n\to\infty} \epsilon_t = \frac{\lambda}{2\alpha} \lim_{p\to\infty} \mathbb{E}_{\{\boldsymbol{x}^\mu, y^\mu\}} \left[\frac{\|\hat{\boldsymbol{w}}\|_2^2}{p} \right] - \lim_{\beta\to\infty} \partial_\beta \Psi_y. \tag{4.10}$$

For what concerns the contribution of the regulariser, we simply note that in the limit of $p \to \infty$, the average concentrates around the overlap parameter $q_w^\star$. Instead, for what concerns the contribution of the loss function, we can start by explicitly taking the derivative with respect to β of Ψ_y in eq. (3.32), i.e.:

$$\partial_\beta \Psi_y = -\mathbb{E}_\xi \left[\int_{\mathbb{R}} \mathrm{d}y \, \frac{\mathcal{Z}_y^0 \left(y, \omega_0^\star\right)}{\mathcal{Z}_y \left(y, \omega_1^\star\right)} \int \frac{\mathrm{d}x}{\sqrt{2\pi V_1^\star}} e^{-\frac{1}{2V_1^\star}(x-\omega_1^\star)^2 - \beta\ell(y,x)} \ell\left(y, x\right) \right], \tag{4.11}$$

with $\mathcal{Z}_y^{\cdot/0}$ defined in eq. (3.31). At this point, as explained more in details in section 4.4, we can notice that in the limit of $\beta \to \infty$, it holds:

$$\lim_{\beta\to\infty} \partial_\beta \Psi_y = -\mathbb{E}_\xi \left[\int_{\mathbb{R}} \mathrm{d}y \; \mathcal{Z}_y^0 \left(y, \omega_0^\star\right) \ell\left(y, \eta\left(y, \omega_1^\star\right)\right) \right], \tag{4.12}$$

with $\eta\left(y, \omega_1^\star\right)$ given in eq. (4.21). Combining the two results together we then finally get:

$$\lim_{n\to\infty} \epsilon_t \to \frac{\lambda}{2\alpha} q_w^\star + \mathbb{E}_\xi \left[\int_{\mathbb{R}} \mathrm{d}y \; \mathcal{Z}_y^0 \left(y, \omega_0^\star\right) \ell\left(y, \eta\left(y, \omega_1^\star\right)\right) \right]. \tag{4.13}$$

4.3 Solving for the overlaps

As we showed above, both the generalisation error and the training loss are completely determined by the $\beta \to \infty$ solution of the extremization problem in eq. (3.35). For strictly convex losses ℓ, there is a unique solution to this problem, that can be found by considering the derivatives of the replica potential. This leads to a set of self-consistent saddle-point equations that can be solved iteratively:

$$\begin{cases} \hat{r}_s = -2\frac{\alpha\kappa_1^2}{\gamma}\partial_{r_s}\Psi_y\left(R, Q, M\right) \\ \hat{q}_s = -2\frac{\alpha\kappa_1^2}{\gamma}\partial_{q_s}\Psi_y\left(R, Q, M\right) \\ \hat{m}_s = \frac{\alpha\kappa_1}{\gamma}\partial_{m_s}\Psi_y\left(R, Q, M\right) \\ \\ \hat{r}_w = -2\alpha\kappa_\star^2\partial_{r_w}\Psi_y\left(R, Q, M\right) \\ \hat{q}_w = -2\alpha\kappa_\star^2\partial_{q_w}\Psi_y\left(R, Q, M\right) \end{cases} \qquad \begin{cases} r_s = -\frac{2}{\gamma}\partial_{\hat{r}_s}\Psi_w\left(\hat{r}_s, \hat{q}_s, \hat{m}_s, \hat{r}_w, \hat{q}_w\right) \\ q_s = -\frac{2}{\gamma}\partial_{\hat{q}_s}\Psi_w\left(\hat{r}_s, \hat{q}_s, \hat{m}_s, \hat{r}_w, \hat{q}_w\right) \\ m_s = \frac{1}{\gamma}\partial_{\hat{m}_s}\Psi_w\left(\hat{r}_s, \hat{q}_s, \hat{m}_s, \hat{r}_w, \hat{q}_w\right) \\ \\ r_w = -2\partial_{\hat{r}_w}\Psi_w\left(\hat{r}_s, \hat{q}_s, \hat{m}_s, \hat{r}_w, \hat{q}_w\right) \\ q_w = -2\partial_{\hat{q}_w}\Psi_w\left(\hat{r}_s, \hat{q}_s, \hat{m}_s, \hat{r}_w, \hat{q}_w\right) \end{cases} \tag{4.14}$$

In the case of a F with well-defined spectral density μ, we can be more explicit and write:

$$\begin{cases} V_s = \frac{1}{\hat{V}_s}\left(1 - z\, g_\mu(-z)\right) \\ q_s = \frac{\hat{m}_s^2+\hat{q}_s}{\hat{V}_s^2}\left[1 - 2zg_\mu(-z) + z^2g_\mu'(-z)\right] - \frac{\hat{q}_w}{(\beta\lambda+\hat{V}_w)\hat{V}_s}\left[-zg_\mu(-z) + z^2g_\mu'(-z)\right] \\ m_s = \frac{\hat{m}_s}{\hat{V}_s}\left(1 - z\, g_\mu(-z)\right) \\ \\ V_w = \frac{\gamma}{\beta\lambda+\hat{V}_w}\left[\frac{1}{\gamma} - 1 + zg_\mu(-z)\right] \\ q_w = \gamma\frac{\hat{q}_w}{(\beta\lambda+\hat{V}_w)^2}\left[\frac{1}{\gamma} - 1 + z^2g_\mu'(-z)\right] - \gamma\frac{\hat{m}_s^2+\hat{q}_s}{(\beta\lambda+\hat{V}_w)\hat{V}_s}\left[-zg_\mu(-z) + z^2g_\mu'(-z)\right] \end{cases} \tag{4.15}$$

where:

$$V_{s/w} = r_{s/w} - q_{r/w} \qquad \hat{V}_{s/w} = \hat{r}_{s/w} + \hat{q}_{r/w} \qquad z = \frac{\beta\lambda + \hat{V}_w}{\hat{V}_s} \tag{4.16}$$

We can also simplify slightly the derivatives of Ψ_y without loosing generality by applying Stein's lemma, yielding:

$$\begin{cases} \hat{V}_s = -\frac{\alpha\kappa_1^2}{\gamma}\mathbb{E}_\xi\left[\int_{\mathbb{R}} \mathrm{d}y\ \mathcal{Z}_y^0\left(y; \frac{M}{\sqrt{Q}}\xi, \rho - \frac{M^2}{Q}\right)\partial_\omega f_y\left(y; \sqrt{Q}\xi, R-Q\right)\right] \\ \hat{q}_s = \frac{\alpha\kappa_1^2}{\gamma}\mathbb{E}_\xi\left[\int_{\mathbb{R}} \mathrm{d}y\ \mathcal{Z}_y^0\left(y; \frac{M}{\sqrt{Q}}\xi, \rho - \frac{M^2}{Q}\right) f_y\left(y; \sqrt{Q}\xi, R-Q\right)^2\right] \\ \hat{m}_s = \frac{\alpha\kappa_1}{\gamma}\mathbb{E}_\xi\left[\int_{\mathbb{R}} \mathrm{d}y\ \mathcal{Z}_y^0\left(y; \frac{M}{\sqrt{Q}}\xi, \rho - \frac{M^2}{Q}\right) f_y^0\left(y; \frac{M}{\sqrt{Q}}\xi, \rho - \frac{M^2}{Q}\right) f_y\left(y; \sqrt{Q}\xi, R-Q\right)\right] \\ \\ \hat{V}_w = -\alpha\kappa_\star^2\mathbb{E}_\xi\left[\int_{\mathbb{R}} \mathrm{d}y\ \mathcal{Z}_y^0\left(y; \frac{M}{\sqrt{Q}}\xi, \rho - \frac{M^2}{Q}\right)\partial_\omega f_y\left(y; \sqrt{Q}\xi, R-Q\right)\right] \\ \hat{q}_w = \alpha\kappa_\star^2\mathbb{E}_\xi\left[\int_{\mathbb{R}} \mathrm{d}y\ \mathcal{Z}_y^0\left(y; \frac{M}{\sqrt{Q}}\xi, \rho - \frac{M^2}{Q}\right) f_y\left(y; \sqrt{Q}\xi, R-Q\right)^2\right] \end{cases} \tag{4.17}$$

with $f_y^{\cdot/0}(y;\omega,V) = \partial_\omega \log \mathcal{Z}_y^{\cdot/0}$. For a given choice of spectral density μ (corresponding to a choice of projection F), label rule P_y^0 and loss function ℓ, the auxiliary functions $(\mathcal{Z}^0, \mathcal{Z})$ can be computed, and from them the right-hand side of the update equations above. The equations can then be iterated until the convergence to the fixed point minimising the free energy at fixed (α, γ, β). For convex losses and $\beta \to \infty$, the fixed point of these equations gives the overlap corresponding to the estimator solving eq. (1.7).

4.4 Taking $\beta \to \infty$ explicitly

Although the saddle-point equations above can be iterated explicitly for any $\beta > 0$, it is envisageable to take the limit $\beta \to \infty$ explicitly, since β is an auxiliary parameter we introduced, and that was not present in the original problem defined in eq. (1.7).

Since the overlap parameters depend on β only implicitly through $\mathcal{Z}_y$ and its derivatives, we proceed with the following ansatz for their $\beta \to \infty$ scaling:

$$\begin{aligned} V_{s/w}^\infty &= \beta V_{s/w} & q_{s/w}^\infty &= q_{s/w} & m_s^\infty &= m_s \\ \hat{V}_{s/w}^\infty &= \frac{1}{\beta}\hat{V}_{s/w} & \hat{q}_{s/w}^\infty &= \frac{1}{\beta^2}\hat{q}_{s/w} & \hat{m}_s^\infty &= \hat{m}_s. \end{aligned} \tag{4.18}$$

This ansatz can be motivated as follows. Recall that:

$$\mathcal{Z}_y(y;\omega,V) = \int \frac{\mathrm{d}x}{\sqrt{2\pi V}} e^{-\beta\left[\frac{(x-\omega)^2}{2\beta V} + \ell(x,y)\right]} = \int \frac{\mathrm{d}x}{\sqrt{2\pi V}} e^{-\beta\mathcal{L}(x)}. \tag{4.19}$$

Therefore, letting $V = \mu_1^2 V_s + \mu_\star^2 V_w$ scale as $V^\infty = \beta V$, at $\beta \to \infty$:

$$\mathcal{Z}_y(y;\omega,V) \underset{\beta\to\infty}{=} e^{-\beta\mathcal{L}(\eta)} \tag{4.20}$$

where:

$$\eta(y;\omega,V) = \underset{x\in\mathbb{R}}{\operatorname{argmin}}\left[\frac{(x-\omega)^2}{2V^\infty} + \ell(x,y)\right]. \tag{4.21}$$

For convex losses ℓ with $\lambda > 0$, this one-dimensional minimisation problem has a unique solution that can be easily evaluated. Intuitively, this ansatz translates the fact the variance of our estimator goes to zero as a power law at $\beta \to \infty$, meaning the Gibbs measure concentrates around the solution of the optimisation problem eq. (1.7). The other scalings in eq. (4.19) follow from analysing the dependence of the saddle-point equations in V.

The ansatz in eq. (4.18) allow us to take the $\beta \to \infty$ and rewrite the saddle-point equations as:

$$\begin{cases} \hat{V}_s^\infty = \frac{\alpha\mu_1^2}{\gamma}\mathbb{E}_\xi\left[\int_{\mathbb{R}} \mathrm{d}y\ \mathcal{Z}_y^0\left(\frac{1-\partial_\omega\eta}{V^\infty}\right)\right] \\ \hat{q}_s^\infty = \frac{\alpha\mu_1^2}{\gamma}\mathbb{E}_\xi\left[\int_{\mathbb{R}} \mathrm{d}y\ \mathcal{Z}_y^0\left(\frac{\eta-\omega}{V^\infty}\right)^2\right] \\ \hat{m}_s^\infty = \frac{\alpha\mu_1}{\gamma}\mathbb{E}_\xi\left[\int_{\mathbb{R}} \mathrm{d}y\ \partial_\omega\mathcal{Z}_y^0\left(\frac{\eta-\omega}{V^\infty}\right)\right] \\ \\ \hat{V}_w^\infty = \alpha\mu_\star^2\mathbb{E}_\xi\left[\int_{\mathbb{R}} \mathrm{d}y\ \mathcal{Z}_y^0\left(\frac{1-\partial_\omega\eta}{V^\infty}\right)\right] \\ \hat{q}_w^\infty = \alpha\mu_\star^2\mathbb{E}_\xi\left[\int_{\mathbb{R}} \mathrm{d}y\ \mathcal{Z}_y^0\left(\frac{\eta-\omega}{V^\infty}\right)^2\right] \end{cases} \tag{4.22}$$

$$\begin{cases} V_s^\infty = \frac{1}{\hat{V}_s^\infty}\left(1 - z\ g_\mu(-z)\right) \\ q_s^\infty = \frac{(\hat{m}_s^\infty)^2+\hat{q}_s^\infty}{\left(\hat{V}_s^\infty\right)^2}\left[1-2zg_\mu(-z)+z^2g'_\mu(-z)\right] - \frac{\hat{q}_w^\infty}{(\lambda+\hat{V}_w)\hat{V}_s}\left[-zg_\mu(-z)+z^2g'_\mu(-z)\right] \\ m_s^\infty = \frac{\hat{m}_s^\infty}{\hat{V}_s^\infty}\left(1 - z\ g_\mu(-z)\right) \\ \\ V_w^\infty = \frac{\gamma}{\lambda+\hat{V}_w^\infty}\left[\frac{1}{\gamma}-1+zg_\mu(-z)\right] \\ q_w^\infty = \gamma\frac{\hat{q}_w^\infty}{(\lambda+\hat{V}_w^\infty)^2}\left[\frac{1}{\gamma}-1+z^2g'_\mu(-z)\right] - \gamma\frac{(\hat{m}_s^\infty)^2+\hat{q}_s^\infty}{(\lambda+\hat{V}_w^\infty)\hat{V}_s^\infty}\left[-zg_\mu(-z)+z^2g'_\mu(-z)\right] \end{cases} \tag{4.23}$$

where $\mathcal{Z}_y^0(y;\omega,V)$ is always evaluated at $(\omega,V) = \left(\frac{M^\infty}{\sqrt{Q^\infty}}\xi, \rho - \frac{M^{\infty\,2}}{Q^\infty}\right)$, $\eta(y;\omega,V)$ at $(\omega,V) = \left(\sqrt{Q^\infty}\xi, V^\infty\right)$ and $z = \frac{\lambda+\hat{V}_w^\infty}{\hat{V}_s^\infty}$.

4.5 Examples

In this section we exemplify our general result in two particular cases for which the integrals in the right-hand side of eq. (4.22) can be analytically performed: the ridge regression task with linear labels and a classification problem with square loss and ridge regularisation term. The former example appears in Fig. 5 (left) and the later in Figs. 2 (blue curve), 6, 7 of the main.

Ridge regression with linear labels: Consider the task of doing ridge regression $\ell(y,x) = \frac{1}{2}\left(y-x\right)^2$, $\lambda > 0$ on the linear patterns $\boldsymbol{y} = \frac{1}{\sqrt{d}}\mathrm{C}\boldsymbol{\theta}^0 + \sqrt{\Delta}\boldsymbol{z}$, with $\boldsymbol{z} \sim \mathcal{N}(\mathbf{0}, \mathrm{I}_n)$ and $\boldsymbol{\theta}^\star \sim \mathcal{N}(\mathbf{0}, \mathrm{I}_d)$. In this case, we have:

$$\eta(y;\omega,V) = \frac{\omega + yV}{1+V} \tag{4.24}$$

and the saddle-point equations for the hat overlap read:

$$\hat{V}_s^\infty = \frac{\alpha}{\gamma}\frac{\kappa_1^2}{1+V^\infty} \qquad \hat{q}_s^0 = \frac{\alpha\kappa_1^2}{\gamma}\frac{1+\Delta+Q^\infty-2M^\infty}{\left(1+V^\infty\right)^2} \qquad \hat{m}_s = \frac{\alpha}{\gamma}\frac{\kappa_1}{1+V^\infty}$$

$$\hat{V}_w^\infty = \frac{\alpha\kappa_\star^2}{1+V^\infty} \qquad \hat{q}_w^\infty = \alpha\kappa_\star^2\frac{1+\Delta+Q^\infty-2M^\infty}{\left(1+V^\infty\right)^2} \tag{4.25}$$

This particular example corresponds precisely to the setting studied in [8].

Classification with square loss and ridge regularisation: Consider a classification task with square loss $\ell(y,x) = \frac{1}{2}\left(y-x\right)^2$ and labels generated as $\boldsymbol{y} = \mathrm{sign}\left(\frac{1}{\sqrt{d}}\mathrm{C}\boldsymbol{\theta}^0\right)$, with $\boldsymbol{\theta}^0 \sim \mathcal{N}(\mathbf{0}, \mathrm{I}_d)$. Then the saddle-point equations are simply:

$$\hat{V}_s^\infty = \frac{\alpha}{\gamma}\frac{\kappa_1^2}{1+V^\infty} \qquad \hat{q}_s^\infty = \frac{\alpha}{\gamma}\kappa_1^2\frac{1+Q^\infty-\frac{2\sqrt{2}M^\infty}{\sqrt{\pi}}}{\left(1+V^\infty\right)^2} \qquad \hat{m}_s = \frac{\alpha}{\gamma}\sqrt{\frac{2}{\pi}}\frac{\kappa_1}{1+V^\infty}$$

$$\hat{V}_w^\infty = \frac{\alpha\kappa_\star^2}{1+V^\infty} \qquad \hat{q}_w^\infty = \alpha\kappa_\star^2\frac{1+Q^\infty-\frac{2M^\infty}{\sqrt{\pi}}}{\left(1+V^\infty\right)^2} \tag{4.26}$$

5 Numerical Simulations

In this section, we provide more details on how the numerical simulations in the main manuscript have been performed.

First, the dataset $\{\boldsymbol{x}^\mu, y^\mu\}_{\mu=1}^n$ is generated according to the procedure described in Section 1.1 of the main, which we summarise here for convenience in algorithm 1:

Algorithm 1 Generating dataset $\{\boldsymbol{x}^\mu, y^\mu\}_{\mu=1}^n$

```
**Input:** Integer $d$, parameters $\alpha, \gamma \in \mathbb{R}_+$, matrix $\mathrm{F} \in \mathbb{R}^{d\times p}$, vector $\boldsymbol{\theta}^0 \in \mathbb{R}^d$ non-linear functions $\sigma, f^0 : \mathbb{R} \to \mathbb{R}$.
Assign $p \leftarrow \lfloor d/\gamma \rfloor$, $n \leftarrow \lfloor \alpha p \rfloor$
Draw $\mathrm{C} \in \mathbb{R}^{n\times d}$ with entries $c_\rho^\mu \sim \mathcal{N}(0,1)$ i.i.d.
Assign $\boldsymbol{y} \leftarrow f^0\left(\mathrm{C}\boldsymbol{\theta}^0\right) \in \mathbb{R}^n$ component-wise.
Assign $\mathrm{X} \leftarrow \sigma\left(\mathrm{CF}\right) \in \mathbb{R}^{n\times p}$ component-wise.
**Return:** $\mathrm{X}, \boldsymbol{y}$
```

In all the examples from the main, we have drawn $\boldsymbol{\theta}^0 \sim \mathcal{N}(0, \mathrm{I}_d)$. For the regression task in Fig. 5 we have taken $f^0(x) = x$, while in the remaining classification tasks $f^0(x) = \text{sign}(x)$. For Gaussian projections, the components of F are drawn from $\mathcal{N}(0,1)$ i.i.d., and in for the random orthogonal projections we draw two orthogonal matrices $\mathrm{U} \in \mathbb{R}^{d\times d}$, $\mathrm{V} \in \mathbb{R}^{p\times p}$ from the Haar measure and we let $\mathrm{F} = \mathrm{U}^\top \mathrm{D} \mathrm{V}$ with $\mathrm{D} \in \mathbb{R}^{d\times p}$ a diagonal matrix with diagonal entries $d_k = \max(\sqrt{\gamma}, 1)$, $k = 1, \cdots, \min(n, p)$.

Given this dataset, the aim is to infer the configuration $\hat{\boldsymbol{w}}$, minimising a given loss function with a ridge regularisation term. In the following, we describe how to accomplish this task for both square and logistic loss.

Square Loss: In this case, the goal is to solve the following optimisation problem:

$$\hat{\boldsymbol{w}} = \min_{\boldsymbol{w}} \left[\frac{1}{2}\sum_{\mu=1}^{n}\left(y^\mu - \boldsymbol{x}^\mu \cdot \boldsymbol{w}\right)^2 + \frac{\lambda}{2}||\boldsymbol{w}||_2^2\right] . \tag{5.1}$$

which has a simple closed-form solution given in terms of the Moore-Penrose inverse:

$$\hat{\boldsymbol{w}} = \begin{cases} \left(\mathrm{X}^\top \mathrm{X} + \lambda \mathrm{I}_p\right)^{-1} \mathrm{X}^\top \boldsymbol{y}, & \text{if } n > p \\ \\ \mathrm{X}^\top \left(\mathrm{X}\mathrm{X}^T + \lambda \mathrm{I}_n\right)^{-1} \boldsymbol{y}, & \text{if } p > n \end{cases} \tag{5.2}$$

Logistic Loss: In this case, the goal is to solve the following optimisation problem:

$$\hat{\boldsymbol{w}} = \min_{\boldsymbol{w}} \left[\sum_{\mu=1}^{n} \log\left(1 + e^{-y^\mu(\boldsymbol{x}^\mu \cdot \boldsymbol{w})}\right) + \frac{\lambda}{2}||\boldsymbol{w}||_2^2\right] . \tag{5.3}$$

To solve the above, we use the *Gradient Descent* (GD) on the regularised loss. In our simulations, we took advantage of Scikit-learn 0.22.1, an out-of-the-box open source library for machine learning tasks in Python [9, 10]. The library provides the class *sklearn.linear_model.LogisticRegression*, which implements GD with logistic loss and a further ℓ_2-regularisation, if the parameter 'penalty' is set to 'l2'. GD stops either if the following condition is satisfied:

$$\max\{(\nabla \boldsymbol{w})_i \, | i = 1, ..., p\} \leqslant \text{tol}, \tag{5.4}$$

with $\nabla \boldsymbol{w}$ being the gradient, or if a maximum number of iterations is reached. We set tol to 10^{-4} and the maximum number of iterations to 10^4.

In both cases described above, the algorithm returns the estimator $\hat{\boldsymbol{w}} \in \mathbb{R}^p$, from which all the quantities of interest can be evaluated. For instance, the generalisation error can be simply computed by drawing a new and independent sample $\{\mathrm{X}^{\text{new}}, \boldsymbol{y}^{\text{new}}\}$ using algorithm 1 with the same inputs F, σ, f^0 and $\boldsymbol{\theta}^0$ and computing:

$$\epsilon_g(n,p,d) = \frac{1}{4^k n} ||\boldsymbol{y}^{\text{new}} - \hat{f}\left(\mathrm{X}^{\text{new}}\hat{\boldsymbol{w}}\right)||_2^2 \tag{5.5}$$

with $\hat{f}(x) = x$ for the regression task and $\hat{f}(x) = \text{sign}(x)$ for the classification task.

The procedure outlined above is repeated n_{seeds} times, for different and independent draws of the random quantities $\mathrm{F}, \boldsymbol{\theta}^0$, and a simple mean is taken in order to obtain the ensemble average of the different quantities. In most of the examples from the main, we found that $n_{\text{seeds}} = 30$ was enough to obtain a very good agreement with the analytical prediction from the replica analysis. The full pipeline for computing the averaged generalisation error is exemplified in algorithm 2.

Algorithm 2 Averaged generalisation error.

```
Input: Integer $d$, parameters $\alpha, \gamma, \lambda \in \mathbb{R}_+$, non-linear functions $\sigma, f^0, \hat{f}$ and integer $n_{\text{seeds}}$.
Assign $p \leftarrow \lfloor d/\gamma \rfloor$, $n \leftarrow \lfloor \alpha p \rfloor$
Initialise $E_g = 0$.
for $i = 1$ to $n_{\text{seeds}}$ do
    Draw F, $\boldsymbol{\theta}^0$.
    Assign $\mathrm{X}, \boldsymbol{y} \leftarrow$ Alg. 1.
    Compute $\hat{\boldsymbol{w}}$ from eq. (5.1) or (5.3) with $\mathrm{X}, \boldsymbol{y}$ and $\lambda$.
    Generate new dataset $\mathrm{X}^{\text{new}}, \boldsymbol{y}^{\text{new}}$ from Alg. 1.
    Assign $E_g \leftarrow E_g + \frac{1}{4^k n}||\boldsymbol{y}^{\text{new}} - \hat{f}\left(\mathrm{X}^{\text{new}}\hat{\boldsymbol{w}}\right)||_2^2$
end for
Return: $\epsilon_g = \frac{E_g}{n_{\text{seeds}}}$
```

References

[1] Sebastian Goldt, Marc Mézard, Florent Krzakala, and Lenka Zdeborová. Modelling the influence of data structure on learning in neural networks. *arXiv preprint arXiv:1909.11500*, 2019.

[2] Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems 20*, pages 1177–1184. 2008.

[3] Walid Hachem, Philippe Loubaton, and Jamal Najim. Deterministic equivalents for certain functionals of large random matrices. *Ann. Appl. Probab.*, 17(3):875–930, 06 2007.

[4] Xiuyuan Cheng and Amit Singer. The spectrum of random inner-product kernel matrices. *Random Matrices: Theory and Applications*, 02(04):1350010, 2013.

[5] Zhou Fan and Andrea Montanari. The spectral norm of random inner-product kernel matrices. *Probability Theory and Related Fields*, 173(1):27–85, Feb 2019.

[6] Jeffrey Pennington and Pratik Worah. Nonlinear random matrix theory for deep learning. In *Advances in Neural Information Processing Systems 30*, pages 2637–2646. 2017.

[7] Song Mei and Andrea Montanari. The generalization error of random features regression: precise asymptotics and double descent curve. *arXiv preprint arXiv:1908.05355*, 2019.

[8] Andrea Montanari, Feng Ruan, Youngtak Sohn, and Jun Yan. The generalization error of max-margin linear classifiers: high-dimensional asymptotics in the overparametrized regime. *arXiv preprint arXiv:1911.01544*, 2019.

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

[10] Lars Buitinck, Gilles Louppe, Mathieu Blondel, Fabian Pedregosa, Andreas Mueller, Olivier Grisel, Vlad Niculae, Peter Prettenhofer, Alexandre Gramfort, Jaques Grobler, Robert Layton, Jake VanderPlas, Arnaud Joly, Brian Holt, and Gaël Varoquaux. API design for machine learning software: experiences from the scikit-learn project. In *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, pages 108–122, 2013.