Generalised linear task on G^{3}M model

Bruno Loureiro

December 16, 2020

Contents

1 Setting

We are interested in the supervised problem of fitting a dataset $(\boldsymbol{x}^\mu, y^\mu)_{\mu=1}^n$ with a parametric model $\hat{y} = f_{\boldsymbol{w}}(\boldsymbol{x})$ by minimising a loss function ℓ:

$$\hat{\boldsymbol{w}} = \underset{\boldsymbol{w}\in\mathbb{R}^p}{\operatorname{argmin}} \left[\sum_{\mu=1}^{n} \ell\left(y^\mu, \hat{y}^\mu\right) + \frac{\lambda}{2} \sum_{i=1}^{p} r(w_i) \right], \tag{1.1}$$

where $\lambda > 0$ and $r(\boldsymbol{w})$ is a regularisation term, e.g. $r(w) = w^2$ for the ℓ_2-penalty.

The G^{3}M model: The G^{3}M model consists of a model where both teacher and student are generalised linear models, but over different input spaces:

$$\textbf{Teacher: } y = f^0\left(\frac{\boldsymbol{z}\cdot\boldsymbol{\theta}^0}{\sqrt{k}}\right), \qquad \boldsymbol{z}\in\mathbb{R}^k, \qquad \boldsymbol{\theta}^0 \sim P_\theta \tag{1.2}$$

$$\textbf{Student: } \hat{y} = \hat{f}\left(\frac{\boldsymbol{x}\cdot\boldsymbol{w}}{\sqrt{p}}\right), \qquad \boldsymbol{x}\in\mathbb{R}^p, \tag{1.3}$$

where $\boldsymbol{z}, \boldsymbol{x} \sim \mathcal{N}(\boldsymbol{0}, \Sigma)$ are jointly Gaussian random variables with covariance matrix $\Sigma \in \mathbb{R}^{(k+p)\times(k+p)}$ given by:

$$\Sigma = \begin{pmatrix} \Psi & \Phi \\ \Phi^\top & \Omega \end{pmatrix} \tag{1.4}$$

Let $\gamma = k/p$ the ratio between these two dimensions, and define the sample complexity with respect to the student dimension $\alpha = n/p$.

Gibbs minimisation

To apply the usual statistical physics tool, we define the Gibbs measure over weights $\boldsymbol{w} \in \mathbb{R}^p$:

$$\mu_\beta(\boldsymbol{w}) = \frac{1}{\mathcal{Z}_\beta} e^{-\beta\left[\sum_{\mu=1}^{n} \ell(y^\mu, \hat{y}^\mu) + \frac{\lambda}{2}\sum_{i=1}^{p} r(w_i)\right]} = \frac{1}{\mathcal{Z}_\beta} \underbrace{\prod_{\mu=1}^{n} e^{-\beta \sum_{\mu=1}^{n} \ell(y^\mu, \hat{y}^\mu)}}_{P_y} \underbrace{\prod_{i=1}^{p} e^{-\frac{\beta\lambda}{2} r(w_i)}}_{P_w} \tag{1.5}$$

Note that P_y and P_w can be interpreted as a (unormalised) likelihood and prior distribution respectively. In the limit $\beta \to \infty$, the measure μ_β concentrates around solutions of the minimisation in eq. (1.1).

2 Replica computation of the free energy

In the statistical physics framework, the aim is to compute the free energy density, defined as

$$f_\beta = \lim_{p\to\infty} \frac{1}{p} \mathbb{E}_{\boldsymbol{x},y} \log \mathcal{Z}_\beta \tag{2.1}$$

using the replica trick:

$$\log \mathcal{Z}_\beta = \lim_{r\to 0^+} \frac{1}{r} \partial_r \mathcal{Z}_\beta^r \tag{2.2}$$

the motivation being that this quantity gives all the information needed to compute the asymptotic training and generalisation error in the problem.

Averaging

The computation of the free energy density thus boils down to the evaluation of the averaged replicated partition function:

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{x},y} \mathcal{Z}_\beta^r &= \prod_{\mu=1}^{n} \mathbb{E}_{\boldsymbol{c}^\mu} \int \mathrm{d}y^\mu \int_{\mathbb{R}^d} \mathrm{d}\boldsymbol{\theta}^0 \; P_{\boldsymbol{\theta}^0}(\boldsymbol{\theta}^0) \prod_{a=1}^{r} \int_{\mathbb{R}^p} \mathrm{d}\boldsymbol{w}^a \; P_w(\boldsymbol{w}^a) P_y\left(y^\mu \middle| \frac{\boldsymbol{x}^\mu \cdot \boldsymbol{w}^a}{\sqrt{p}}\right) P_y^0\left(y^\mu \middle| \frac{\boldsymbol{z}^\mu \cdot \boldsymbol{\theta}^0}{\sqrt{k}}\right) \\
&= \prod_{\mu=1}^{n} \int \mathrm{d}y^\mu \int_{\mathbb{R}^d} \mathrm{d}\boldsymbol{\theta}^0 \; P_{\boldsymbol{\theta}^0}(\boldsymbol{\theta}^0) \int_{\mathbb{R}^{p\times r}} \left(\prod_{a=1}^{r} \mathrm{d}\boldsymbol{w}^a \; P_w(\boldsymbol{w}^a)\right) \underbrace{\mathbb{E}_{\boldsymbol{c}^\mu}\left[P_y^0\left(y^\mu \middle| \frac{\boldsymbol{z}^\mu \cdot \boldsymbol{\theta}^0}{\sqrt{k}}\right) \prod_{a=1}^{r} P_y\left(y^\mu \middle| \frac{\boldsymbol{x}^\mu \cdot \boldsymbol{w}^a}{\sqrt{p}}\right)\right]}_{(\star)}
\end{aligned} \tag{2.3}$$

Focusing on the average term:

$$\begin{aligned}
(\star) &= \mathbb{E}_{\boldsymbol{c}^\mu}\left[P_y^0\left(y^\mu \middle| \frac{\boldsymbol{z}^\mu \cdot \boldsymbol{\theta}^0}{\sqrt{k}}\right) \prod_{a=1}^{r} P_y\left(y^\mu \middle| \frac{\boldsymbol{x}^\mu \cdot \boldsymbol{w}^a}{\sqrt{p}}\right)\right] \\
&= \int_{\mathbb{R}} \mathrm{d}\nu_\mu P_y^0(y|\nu_\mu) \int_{\mathbb{R}^r} \left(\prod_{a=1}^{r} \mathrm{d}\lambda_\mu^a P_y(y^\mu|\lambda_\mu^a)\right) \underbrace{\mathbb{E}_{\boldsymbol{c}^\mu}\left[\delta\left(\nu_\mu - \frac{\boldsymbol{z}^\mu \cdot \boldsymbol{\theta}^0}{\sqrt{k}}\right) \prod_{a=1}^{r} \delta\left(\lambda_\mu^a - \frac{\boldsymbol{x}^\mu \cdot \boldsymbol{w}^a}{\sqrt{p}}\right)\right]}_{P(\nu,\lambda)}
\end{aligned} \tag{2.4}$$

Note that since $\boldsymbol{x}^\mu = \mathcal{G}(\boldsymbol{c}^\mu)$ the last term defines the joint probability between the random variables (ν_μ, λ_μ^a). The *Gaussian Equivalence Principle* states that for certain architectures $\mathcal{G}$, the random

variables (ν_μ, λ_μ^a) are asymptotically jointly Gaussian, with zero mean and covariance matrix given by:

$$\Sigma^{ab} = \begin{pmatrix} \rho & m^a \\ m^a & Q^{ab} \end{pmatrix}. \tag{2.5}$$

where the so-called overlap parameters (ρ, m^a, Q^{ab}) are related to the weights $\boldsymbol{\theta}^0, \boldsymbol{w}$:

$$\rho \equiv \mathbb{E}\left[\nu_\mu^2\right] = \frac{1}{k}\boldsymbol{\theta}^{0\top}\Psi\boldsymbol{\theta}^0, \qquad m^a \equiv \mathbb{E}\left[\lambda_\mu^a \nu_\mu\right] = \frac{1}{\sqrt{pk}}\boldsymbol{w}^{a\top}\Phi\boldsymbol{\theta}^0, \qquad Q^{ab} \equiv \mathbb{E}\left[\lambda_\mu^a \lambda_\mu^b\right] = \frac{1}{p}\boldsymbol{w}^{a\top}\Omega\boldsymbol{w}^b$$

We can therefore write the averaged replicated partition function as:

$$\mathbb{E}_{\boldsymbol{x},y}\mathcal{Z}_\beta^r = \prod_{\mu=1}^n \int \mathrm{d}y^\mu \int_{\mathbb{R}^d} \mathrm{d}\boldsymbol{\theta}^0 \; P_{\boldsymbol{\theta}^0}(\boldsymbol{\theta}^0) \int_{\mathbb{R}^{p\times r}} \left(\prod_{a=1}^r \mathrm{d}\boldsymbol{w}^a \; P_w(\boldsymbol{w}^a)\right) \mathcal{N}(\nu_\mu, \lambda_\mu^a; \mathbf{0}, \Sigma^{ab}) \tag{2.6}$$

Rewriting as a saddle-point problem

The next step is to free the overlap parameters by introducing delta functions:

$$\begin{aligned} 1 &\propto \int_{\mathbb{R}} \mathrm{d}\rho \; \delta\left(k\rho - \boldsymbol{\theta}^{0\top}\Psi\boldsymbol{\theta}^0\right) \int_{\mathbb{R}^r} \prod_{a=1}^r \mathrm{d}m^a \; \delta\left(\sqrt{kp}m^a - \boldsymbol{w}^a\Phi\boldsymbol{\theta}^0\right) \int_{\mathbb{R}^{r\times r}} \prod_{1\le a\le b\le r} \mathrm{d}Q^{ab} \; \delta\left(pQ^{ab} - \boldsymbol{w}^{a\top}\Omega\boldsymbol{w}^b\right) \\ &= \int_{\mathbb{R}} \frac{\mathrm{d}\rho\mathrm{d}\hat{\rho}}{2\pi} \; e^{i\hat{\rho}\left(k\rho - \boldsymbol{\theta}^{0\top}\Psi\boldsymbol{\theta}^0\right)} \int_{\mathbb{R}^r} \prod_{a=1}^r \frac{\mathrm{d}m^a\mathrm{d}\hat{m}^a}{2\pi} \; e^{i\sum_{a=1}^r \hat{m}^a\left(\sqrt{kp}m^a - \boldsymbol{w}^{a\top}\Phi\boldsymbol{\theta}^0\right)} \times \\ &\quad \times \int_{\mathbb{R}^{r\times r}} \prod_{1\le a\le b\le r} \frac{\mathrm{d}Q^{ab}\mathrm{d}\hat{Q}^{ab}}{2\pi} \; e^{i\sum_{1\le a\le b\le r}\hat{Q}^{ab}\left(pQ^{ab} - \boldsymbol{w}^{a\top}\Omega\boldsymbol{w}^b\right)} \end{aligned} \tag{2.7}$$

Inserting this in eq. (2.6) allow us to rewrite:

$$\mathbb{E}_{\boldsymbol{x},y}\mathcal{Z}_\beta^r = \int_{\mathbb{R}} \frac{\mathrm{d}\rho\mathrm{d}\hat{\rho}}{2\pi} \int_{\mathbb{R}^r} \prod_{a=1}^r \frac{\mathrm{d}m^a\mathrm{d}\hat{m}^a}{2\pi} \int_{\mathbb{R}^{r\times r}} \prod_{1\le a\le b\le r} \frac{\mathrm{d}Q^{ab}\mathrm{d}\hat{Q}^{ab}}{2\pi} e^{p\Phi^{(r)}} \tag{2.8}$$

where we have absorbed a $-i$ factor in the integrals (this won't matter since we will look to the saddle-point) and defined the potential:

$$\Phi^{(r)} = -\gamma\rho\hat{\rho} - \sqrt{\gamma}\sum_{a=1}^r m^a\hat{m}^a - \sum_{1\le a\le b\le r} Q^{ab}\hat{Q}^{ab} + \alpha\Psi_y^{(r)}(\rho, m^a, Q^{ab}) + \Psi_w^{(r)}(\hat{\rho}, \hat{m}^a, \hat{Q}^{ab}) \tag{2.9}$$

with $\alpha = n/p$, $\gamma = k/p$ and:

$$\Psi_w^{(r)} = \frac{1}{p}\log \int_{\mathbb{R}^d} \mathrm{d}\boldsymbol{\theta}^0 P_{\boldsymbol{\theta}^0}\left(\boldsymbol{\theta}^0\right) \int_{\mathbb{R}^{p\times r}} \prod_{a=1}^r \mathrm{d}\boldsymbol{w}^a P_w\left(\boldsymbol{w}^a\right) e^{\hat{\rho}\boldsymbol{\theta}^{0\top}\Psi\boldsymbol{\theta}^0 + \sum_{a=1}^r \hat{m}^a\boldsymbol{w}^{a\top}\Phi\boldsymbol{\theta}^0 + \sum_{1\le a\le b\le r}\hat{Q}^{ab}\boldsymbol{w}^{a\top}\Omega\boldsymbol{w}^b} \tag{2.10}$$

$$\Psi_y^{(r)} = \log \int_{\mathbb{R}} \mathrm{d}y \int_{\mathbb{R}} \mathrm{d}\nu \; P_y^0(y|\nu) \int \prod_{a=1}^r \mathrm{d}\lambda^a P_y(y|\lambda^a) \; \mathcal{N}(\nu, \lambda^a; \mathbf{0}, \Sigma^{ab}) \tag{2.11}$$

In the high-dimensional limit where $p \to \infty$ while $\alpha = n/p$ and $\gamma = k/p$ stay finite, the integral in eq. (2.8) concentrate around the values of the overlaps that extremise $\Phi^{(r)}$, and therefore we can write:

$$f = -\lim_{r\to 0^+} \frac{1}{r}\mathbf{extr}\; \Phi^{(r)}\left(\hat{\rho}, \hat{m}^a, \hat{Q}^{ab}; \rho, m^a, Q^{ab}\right) \tag{2.12}$$

Replica symmetric ansatz

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

$$\begin{aligned} m^a &= m, & \hat{m}^a &= \hat{m}, & &\text{for } a = 1, \dots, r \\ q^{aa} &= r, & \hat{q}^{aa} &= -\frac{1}{2}\hat{r}, & &\text{for } a = 1, \dots, r \\ Q^{ab} &= q, & \hat{Q}^{ab} &= \hat{q}, & &\text{for } 1 \leq a < b \leq r \end{aligned} \tag{2.13}$$

The steps in the $r \to 0^+$ limit for the trace and channel terms are exactly the same as in the single layer hidden-manifold model:

$$\Psi_y \equiv \lim_{r \to 0^+} \frac{1}{r} \Psi_w^{(r)} = \mathbb{E}_\xi \left[\int_{\mathbb{R}} \mathrm{d}y \; \mathcal{Z}_y^0 \left(y, \frac{m}{\sqrt{q}} \xi, \rho - \frac{m^2}{q} \right) \log \mathcal{Z}_y(y, \sqrt{q}\xi, V) \right] \tag{2.14}$$

where $\xi \sim \mathcal{N}(0,1)$, $V = r - q$ and:

$$\mathcal{Z}_y^{\cdot/0}(y, \omega, V) = \int_{\mathbb{R}} \frac{\mathrm{d}x}{\sqrt{2\pi V}} e^{-\frac{(x-\omega)^2}{2V}} P_y^{\cdot/0}(y|x) \tag{2.15}$$

As before, the consistency condition of the zeroth order term in the free energy fix $\rho = \mathbb{E}_{\boldsymbol{\theta}^0} \left[\frac{1}{k} \boldsymbol{\theta}^{0\top} \Psi \boldsymbol{\theta}^0 \right]$ and $\hat{\rho} = 0$. For the prior term, we need to be a bit more careful. First, inserting the replica symmetric ansatz:

$$\Psi_w^{(r)} = \frac{1}{p} \log \int_{\mathbb{R}^d} \mathrm{d}\boldsymbol{\theta}^0 P_{\boldsymbol{\theta}^0}\left(\boldsymbol{\theta}^0\right) \int_{\mathbb{R}^{p \times r}} \prod_{a=1}^{r} \mathrm{d}\boldsymbol{w}^a P_w\left(\boldsymbol{w}^a\right) e^{-\frac{\hat{V}}{2} \sum\limits_{a=1}^{r} \boldsymbol{w}^{a\top} \Omega \boldsymbol{w}^a + \hat{m} \sum\limits_{a=1}^{r} \boldsymbol{w}^{a\top} \Phi \boldsymbol{\theta}^0 + \hat{q} \sum\limits_{a,b=1}^{r} \boldsymbol{w}^{a\top} \Omega \boldsymbol{w}^b} \tag{2.16}$$

where we have defined $\hat{V} = \hat{r} + \hat{q}$. Now using that:

$$e^{\hat{q} \sum\limits_{a,b=1}^{r} \boldsymbol{w}^{a\top} \Omega \boldsymbol{w}^b} = \mathbb{E}_\xi \left[e^{\sqrt{\hat{q}} \xi \Omega^{1/2} \sum\limits_{a=1}^{r} \boldsymbol{w}^a} \right] \tag{2.17}$$

for $\xi \sim \mathcal{N}(0,1)$, we can write:

$$\Psi_w^{(r)} = \frac{1}{p} \log \mathbb{E}_\xi \int_{\mathbb{R}^k} \mathrm{d}\boldsymbol{\theta}^0 P_{\boldsymbol{\theta}^0}\left(\boldsymbol{\theta}^0\right) \prod_{a=1}^{r} \int_{\mathbb{R}^p} \mathrm{d}\boldsymbol{w}^a P_w\left(\boldsymbol{w}^a\right) e^{-\frac{\hat{V}}{2} \boldsymbol{w}^{a\top} \Omega \boldsymbol{w}^a + \boldsymbol{w}^{a\top} \left(\hat{m} \Phi \boldsymbol{\theta}^0 + \hat{q} \Omega^{1/2} \xi \right)} \tag{2.18}$$

$$= \frac{1}{p} \log \mathbb{E}_\xi \int_{\mathbb{R}^k} \mathrm{d}\boldsymbol{\theta}^0 P_{\boldsymbol{\theta}^0}\left(\boldsymbol{\theta}^0\right) \left[\int_{\mathbb{R}^p} \mathrm{d}\boldsymbol{w} \; P_w\left(\boldsymbol{w}\right) e^{-\frac{\hat{V}}{2} \boldsymbol{w}^{\top} \Omega \boldsymbol{w} + \boldsymbol{w}^{\top} \left(\hat{m} \Phi \boldsymbol{\theta}^0 + \hat{q} \mathbf{1}^{\top} \Omega^{1/2} \xi \right)} \right]^r \tag{2.19}$$

and therefore:

$$\Psi_w \equiv \lim_{r \to 0^+} \frac{1}{r} \Psi_w^{(r)} = \frac{1}{p} \mathbb{E}_{\xi, \boldsymbol{\theta}^0} \log \int_{\mathbb{R}^p} \mathrm{d}\boldsymbol{w} \; P_w\left(\boldsymbol{w}\right) e^{-\frac{\hat{V}}{2} \boldsymbol{w}^{\top} \Omega \boldsymbol{w} + \boldsymbol{w}^{\top} \left(\hat{m} \Phi \boldsymbol{\theta}^0 + \hat{q} \mathbf{1}^{\top} \Omega^{1/2} \xi \right)} \tag{2.20}$$

Summary

The replica symmetric free energy density is simply given by:

$$f_\beta = \underset{q, m, \hat{q}, \hat{m}}{\mathbf{extr}} \left\{ -\frac{1}{2} r\hat{r} - \frac{1}{2} q\hat{q} + \sqrt{\gamma} \; m\hat{m} - \alpha \Psi_y(r, m, q) - \Psi_w(\hat{r}, \hat{m}, \hat{q}) \right\} \tag{2.21}$$

where

$$\Psi_w = \lim_{p\to\infty} \frac{1}{p}\mathbb{E}_{\xi,\boldsymbol{\theta}^0} \log \int_{\mathbb{R}^p} \mathrm{d}\boldsymbol{w}\ P_w\left(\boldsymbol{w}\right) e^{-\frac{\hat{V}}{2}\boldsymbol{w}^\top\Omega\boldsymbol{w}+\boldsymbol{w}^\top\left(\hat{m}\Phi\boldsymbol{\theta}^0+\hat{q}\mathbf{1}^\top\Omega^{1/2}\xi\right)}$$
$$\Psi_y = \mathbb{E}_\xi\left[\int_{\mathbb{R}} \mathrm{d}y\ \mathcal{Z}_y^0\left(y, \frac{m}{\sqrt{q}}\xi, \rho - \frac{m^2}{q}\right)\log \mathcal{Z}_y(y,\sqrt{q}\xi,V)\right] \tag{2.22}$$

and

$$\mathcal{Z}_y^{\cdot/0}(y,\omega,V) = \int_{\mathbb{R}} \frac{\mathrm{d}x}{\sqrt{2\pi V}} e^{-\frac{(x-\omega)^2}{2V}} P_y^{\cdot/0}(y|x) \tag{2.23}$$

3 Ridge regression and Gaussian student

For an ℓ_2-regularisation term, we have:

$$P_w(\boldsymbol{w}) = \frac{1}{(2\pi)^{p/2}} e^{-\frac{\beta\lambda}{2}||\boldsymbol{w}||_2^2} \tag{3.1}$$

where we have included a convenient constant, and therefore:

$$\begin{aligned}\int_{\mathbb{R}^p} \mathrm{d}\boldsymbol{w}\ P_w(\boldsymbol{w}) e^{-\frac{\hat{V}}{2}\boldsymbol{w}^\top\Omega\boldsymbol{w}+\boldsymbol{w}^\top\left(\hat{m}\Phi\boldsymbol{\theta}^0+\sqrt{\hat{q}}\mathbf{1}^\top\Omega^{1/2}\xi\right)} &= \int_{\mathbb{R}^p} \frac{\mathrm{d}\boldsymbol{w}}{(2\pi)^{p/2}} e^{-\frac{1}{2}\boldsymbol{w}^\top\left(\beta\lambda \mathrm{I}_p+\hat{V}\Omega\right)\boldsymbol{w}+\boldsymbol{w}^\top\left(\hat{m}\Phi\boldsymbol{\theta}^0+\sqrt{\hat{q}}\mathbf{1}^\top\Omega^{1/2}\xi\right)} \\ &= \frac{\exp\left(\frac{1}{2}\left(\hat{m}\Phi\boldsymbol{\theta}^0+\sqrt{\hat{q}}\mathbf{1}^\top\Omega^{1/2}\xi\right)^\top\left(\beta\lambda\mathrm{I}_p+\hat{V}\Omega\right)^{-1}\left(\hat{m}\Phi\boldsymbol{\theta}^0+\sqrt{\hat{q}}\mathbf{1}^\top\Omega^{1/2}\xi\right)^\top\right)}{\sqrt{\det\left(\beta\lambda\mathrm{I}_p+\hat{V}\Omega\right)}}\end{aligned} \tag{3.2}$$

taking the log and using $\log\det = \operatorname{tr}\log$, up to the limit:

$$\Psi_w = -\frac{1}{2p}\operatorname{tr}\log\left(\beta\lambda\mathrm{I}_p+\hat{V}\Omega\right) + \frac{1}{2p}\mathbb{E}_{\xi,\boldsymbol{\theta}^0}\left[\left(\hat{m}\Phi\boldsymbol{\theta}^0+\sqrt{\hat{q}}\mathbf{1}^\top\Omega^{1/2}\xi\right)^\top\left(\beta\lambda\mathrm{I}_p+\hat{V}\Omega\right)^{-1}\left(\hat{m}\Phi\boldsymbol{\theta}^0+\sqrt{\hat{q}}\mathbf{1}^\top\Omega^{1/2}\xi\right)\right] \tag{3.3}$$

Defining the shorthand $\mathrm{A} = \left(\beta\lambda\mathrm{I}_p+\hat{V}\Omega\right)^{-1}$, we can now take the averages over ξ explicitly:

$$\mathbb{E}_\xi\left[\left(\hat{m}\Phi\boldsymbol{\theta}^0+\sqrt{\hat{q}}\mathbf{1}^\top\Omega^{1/2}\xi\right)^\top\mathrm{A}\left(\hat{m}\Phi\boldsymbol{\theta}^0+\hat{q}\mathbf{1}^\top\Omega^{1/2}\xi\right)\right] = \hat{m}^2{\boldsymbol{\theta}^0}^\top\Phi^\top\mathrm{A}\Phi\boldsymbol{\theta}^0 + \hat{q}\operatorname{tr}\Omega^{1/2}\mathrm{A}\Omega^{1/2} \tag{3.4}$$

Now taking the average with respect to $\boldsymbol{\theta}^0$:

$$\mathbb{E}_{\boldsymbol{\theta}^0}\left[{\boldsymbol{\theta}^0}^\top\Phi^\top\mathrm{A}\Phi\boldsymbol{\theta}^0\right] = \operatorname{tr}\Phi^\top\mathrm{A}\Phi \tag{3.5}$$

Putting together,

$$\Psi_w = -\frac{1}{2p}\operatorname{tr}\log\left(\beta\lambda\mathrm{I}_p+\hat{V}\Omega\right) + \hat{m}^2\frac{1}{2p}\operatorname{tr}\Phi\Phi^\top\left(\beta\lambda\mathrm{I}_p+\hat{V}\Omega\right)^{-1} + \hat{q}\frac{1}{2p}\operatorname{tr}\Omega\left(\beta\lambda\mathrm{I}_p+\hat{V}\Omega\right)^{-1} \tag{3.6}$$

Also note that $\boldsymbol{\theta}^0 \sim \mathcal{N}(\mathbf{0}, \mathrm{I}_d)$ imply that:

$$\rho = \mathbb{E}_{\boldsymbol{\theta}^0}\left[\frac{1}{k}{\boldsymbol{\theta}^0}^\top\Psi\boldsymbol{\theta}^0\right] = \frac{1}{k}\operatorname{tr}\Psi \tag{3.7}$$

4 Zero-temperature state evolution equations

The state evolution equations related to the likelihood Ψ_y are the same as before. Therefore we only need to compute the derivatives of Ψ_w with respect to the overlaps $(\hat{r}, \hat{q}, \hat{m})$. Recalling that $\hat{V} = \hat{r} + \hat{q}$:

$$\begin{aligned}
\partial_{\hat{r}} \Psi_w &= -\frac{1}{2p} \operatorname{tr}\left(\beta\lambda \mathrm{I}_p + \hat{V}\Omega\right)^{-1} \Omega - \frac{\hat{m}^2}{2p} \operatorname{tr}\left(\beta\lambda \mathrm{I}_p + \hat{V}\Omega\right)^{-2} \Omega\Phi\Phi^\top - \frac{\hat{q}}{2p} \operatorname{tr}\left(\beta\lambda \mathrm{I}_p + \hat{V}\Omega\right)^{-2} \Omega^2 \\
\partial_{\hat{q}} \Psi_w &= -\frac{\hat{m}^2}{2p} \operatorname{tr}\left(\beta\lambda \mathrm{I}_p + \hat{V}\Omega\right)^{-2} \Omega\Phi\Phi^\top - \frac{\hat{q}}{2p} \operatorname{tr}\left(\beta\lambda \mathrm{I}_p + \hat{V}\Omega\right)^{-2} \Omega^2 \qquad (4.1) \\
\partial_{\hat{m}} \Psi_w &= \frac{\hat{m}}{p} \operatorname{tr} \Phi\Phi^\top \left(\beta\lambda \mathrm{I}_p + \hat{V}\Omega\right)^{-1} \qquad (4.2)
\end{aligned}$$

Therefore the finite temperature state evolution equations read:

$$\begin{cases} \hat{V} = \alpha \mathbb{E}_\xi \left[\int_{\mathbb{R}} \mathrm{d}y \; \mathcal{Z}_y^0 \partial_\omega f_{\text{out}}\right] \\ \hat{q} = \alpha \mathbb{E}_\xi \left[\int_{\mathbb{R}} \mathrm{d}y \; \mathcal{Z}_y^0 f_{\text{out}}^2\right] \\ \hat{m} = \frac{\alpha}{\sqrt{\gamma}} \mathbb{E}_\xi \left[\int_{\mathbb{R}} \mathrm{d}y \; \partial_\omega \mathcal{Z}_y^0 f_{\text{out}}\right] \end{cases} \qquad \begin{cases} V = \frac{1}{p} \operatorname{tr}\left(\beta\lambda \mathrm{I}_p + \hat{V}\Omega\right)^{-1} \Omega \\ q = \frac{1}{p} \operatorname{tr}\left[\left(\hat{q}\Omega + \hat{m}^2 \Phi\Phi^\top\right) \Omega \left(\beta\lambda \mathrm{I}_p + \hat{V}\Omega\right)^{-2}\right] \\ m = \frac{1}{\sqrt{\gamma}} \frac{\hat{m}}{p} \operatorname{tr} \Phi\Phi^\top \left(\beta\lambda \mathrm{I}_p + \hat{V}\Omega\right)^{-1} \end{cases} \qquad (4.3)$$

To take the zero temperature limit, we let:

$$\begin{aligned}
V^\infty &= \beta V & q^\infty &= q & m^\infty &= m \\
\hat{V}^\infty &= \frac{1}{\beta} \hat{V} & \hat{q}^\infty &= \frac{1}{\beta^2} \hat{q} & \hat{m}^\infty &= \frac{1}{\beta} \hat{m}.
\end{aligned} \qquad (4.4)$$

Inserting this in the equations above and taking $\beta \to 0$:

$$\begin{cases} \hat{V} = \alpha \mathbb{E}_\xi \left[\int_{\mathbb{R}} \mathrm{d}y \; \mathcal{Z}_y^0 \left(\frac{1 - \partial_\omega \eta}{V}\right)\right] \\ \hat{q} = \alpha \mathbb{E}_\xi \left[\int_{\mathbb{R}} \mathrm{d}y \; \mathcal{Z}_y^0 \left(\frac{\eta - \omega}{V}\right)^2\right] \\ \hat{m} = \frac{\alpha}{\sqrt{\gamma}} \mathbb{E}_\xi \left[\int_{\mathbb{R}} \mathrm{d}y \; \partial_\omega \mathcal{Z}_y^0 \left(\frac{\eta - \omega}{V}\right)\right] \end{cases} \qquad \begin{cases} V = \frac{1}{p} \operatorname{tr}\left(\lambda \mathrm{I}_p + \hat{V}\Omega\right)^{-1} \Omega \\ q = \frac{1}{p} \operatorname{tr}\left[\left(\hat{q}\Omega + \hat{m}^2 \Phi\Phi^\top\right) \Omega \left(\lambda \mathrm{I}_p + \hat{V}\Omega\right)^{-2}\right] \\ m = \frac{1}{\sqrt{\gamma}} \frac{\hat{m}}{p} \operatorname{tr} \Phi\Phi^\top \left(\lambda \mathrm{I}_p + \hat{V}\Omega\right)^{-1} \end{cases} \qquad (4.5)$$

where η is the proximal operator:

$$\eta(y, \omega, V) = \underset{x \in \mathbb{R}}{\operatorname{argmin}} \left[\frac{(x - \omega)^2}{2V} + \ell\left(y, x\right)\right] \qquad (4.6)$$

5 Training and test errors

Training loss: We define the training loss as the prediction loss the data set $\{\boldsymbol{x}^\mu, y^\mu\}_{\mu=1}^n$:

$$\epsilon_t = \lim_{n \to \infty} \frac{1}{n} \mathbb{E}_{\boldsymbol{x}, y} \sum_{\mu=1}^{n} \ell\left(y^\mu, \hat{f}(\hat{\boldsymbol{w}} \cdot \boldsymbol{x}^\mu)\right) = -\lim_{\beta \to \infty} \partial_\beta \Psi_y(V^*, q^*, m^*) \qquad (5.1)$$

where $V^\star, q^\star$ and $m^\star$ are the solutions of the extremisation in eq. (2.21). For instance, for the ridge task, we simply have:

$$\epsilon_t = \frac{\rho + q^\star - 2m^\star}{(1 + V^\star)^2} \qquad (5.2)$$

Test error: We define the test error as the prediction ℓ_2 loss on a new pair of samples $\{\boldsymbol{x}^{\text{new}}, y^{\text{new}}\}$:

$$\begin{aligned}\epsilon_g &= \lim_{p\to\infty} \mathbb{E}_{\boldsymbol{x}^{\text{new}},y^{\text{new}}} \left(y^{\text{new}} - \hat{f}(\hat{\boldsymbol{w}}\cdot\boldsymbol{x}^{\text{new}})\right)^2 = \lim_{p\to\infty} \mathbb{E}_{\boldsymbol{z},\boldsymbol{\theta}^0} \left[\left(f^0(\boldsymbol{z}\cdot\boldsymbol{\theta}^0) - \hat{f}\left(\boldsymbol{x}\cdot\hat{\boldsymbol{w}}\right)\right)^2\right] \\ &= \mathbb{E}_{\nu,\lambda}\left[\left(f^0(\nu) - \hat{f}(\lambda)\right)^2\right] \end{aligned} \tag{5.3}$$

where $\nu, \lambda \sim \mathcal{N}(0, \Sigma)$ are jointly Gaussian random variables with zero mean and covariance matrix

$$\Sigma = \begin{pmatrix} \rho & m^\star \\ m^\star & q^\star \end{pmatrix}. \tag{5.4}$$

For instance, for the ridge task we simply have $\epsilon_t = \rho + q^\star - 2m^\star$ while for a classification task we have $\epsilon_t = \pi^{-1}\cos^{-1}\left(m/\sqrt{q}\right)$.

6 Numerical experiments

As a check, consider the simplest setting where both teacher and student are drawn from a single-layer hidden manifold model with different dimensions and different activations:

$$\boldsymbol{z} = \bar{\sigma}\left(\frac{1}{\sqrt{d}}\bar{\mathrm{F}}\boldsymbol{c}\right), \qquad\qquad \boldsymbol{x} = \sigma\left(\frac{1}{\sqrt{d}}\mathrm{F}\boldsymbol{c}\right), \tag{6.1}$$

with $\bar{\mathrm{F}} \in \mathbb{R}^{k\times d}$, $\mathrm{F} \in \mathbb{R}^{p\times d}$ two Gaussian matrices. Due to the GET, we know that this is equivalent to the G^{3}M model with the following covariances:

$$\Psi = \bar{\kappa}_1^2\bar{\mathrm{F}}\bar{\mathrm{F}}^\top + \bar{\kappa}_\star^2\mathrm{I}_k, \qquad\qquad \Phi = \bar{\kappa}_1\kappa_1\mathrm{F}\bar{\mathrm{F}}^\top, \qquad\qquad \Omega = \kappa_1^2\mathrm{F}\mathrm{F}^\top + \kappa_\star^2\mathrm{I}_p \tag{6.2}$$

with $\kappa_1 \equiv \mathbb{E}\left[\xi\sigma(\xi)\right]$ and $\kappa_\star^2 \equiv \mathbb{E}\left[\sigma(\xi)\right]^2 - \kappa_1^2$ for $\xi \sim \mathcal{N}(0,1)$ (idem for the bar). Figure 1 shows good agreement between the replica result and the simulations done directly on the non-linear models, for two tasks: ridge and logistic regression.

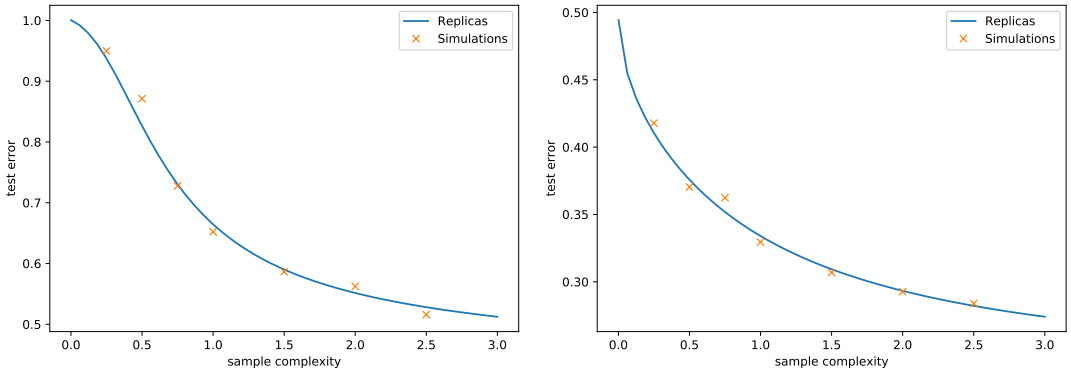

Figure 1: Generalisation curves for the model defined in eq. (6.1) for **(left)** ridge regression and **(right)** logistic regression, with $d/k = 1$, $d/p = 0.5$ and $\lambda = 0.01$.