Title to be decided

Kasimir Tanner,[1] Matteo Vilucchio,[1] and Florent Krzakala[1]
[1] *Information, Learning and Physics Laboratory, EPFL, 1015 Lausanne, Switzerland*

Sample Abstract

I. INTRODUCTION

M: Important questions that we can think about:

1. What can we predict and do novel with this model?
2. What is the human-like model that can correctly classify the perturbed images?
3. Why in the TPIV we had to consider q instead of Q?

K:

1. With this model we can try to find settings in which adversarial training does not help. I presume this exists. The alternative would also be an interesting finding.
2. Is a human-like model a denoised model that only uses the principal components? I.e. could a case be made that overparametrization hurts sometimes?
3. With these questions I wonder if we can only provide experimental evidence or if we can show something more mathy.

II. RELATED WORKS

III. DATA MODEL

M: We can also prove that the new minimisation is convex. This could lead to a more strong justification of Replica Symmetric solution. K: I would not call myself familiar with proofs, would it go along the lines of: the dot product is linear and hence trivially convex, the L-2 norm (with positive definite covariance matrix) is also trivially convex; the sum of convex functions is convex; a convex function (logistic loss) of a convex function is convex (is it though?); the sum of convex functions (over the training samples) is convex? K: As an alternative, we might look at the hessian of the function and show that it is positive-definite.

$$x \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{x}}) \,, \quad \boldsymbol{\theta}_0 \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{\theta}}) \tag{1}$$

–

$$\begin{aligned} y(\boldsymbol{x}) &= f\left(\frac{1}{\sqrt{d}} \boldsymbol{\theta}_0^\top \boldsymbol{x}\right) \\ \hat{y}(\boldsymbol{x}) &= f\left(\frac{1}{\sqrt{d}} \boldsymbol{w}^\top \boldsymbol{x}\right) \end{aligned} \tag{2}$$

We suppose that the classification loss is a decreasing function of its single argument.
We are interested in looking at the following

$$\hat{\boldsymbol{w}} = \arg\min_{\boldsymbol{w}} \sum_{\mu=1}^{n} \max_{\|\boldsymbol{\delta}\|_{\boldsymbol{\Sigma}_{\boldsymbol{\delta}}^2} \leq \frac{\varepsilon_t}{\sqrt{d}}} g\left(y_\mu \frac{\boldsymbol{w}^\top (\boldsymbol{x} + \boldsymbol{\delta})}{\sqrt{d}}\right) + \frac{\lambda}{2} r(\boldsymbol{w}) \tag{3}$$

It is important to notice the dimensional scaling of the adversarial training constant ε_t has been made explicit and it is $1/\sqrt{d}$.

As a loss function g we choose the logistic loss

$$g(x) = \log(1 + \exp(-x)) \tag{4}$$

The inner maximisation is solved by

$$\delta = -y \frac{\boldsymbol{\Sigma}_{\boldsymbol{\delta}}^{-1/2} \boldsymbol{w}}{\|\boldsymbol{w}\|_2} \tag{5}$$

leading to this modified problem which is

$$\hat{\boldsymbol{w}} = \arg\min_{\boldsymbol{w}} \sum_{\mu=1}^{n} g\left(y_\mu \frac{\boldsymbol{w}^\top \boldsymbol{x}}{\sqrt{d}} - \varepsilon_t \frac{\boldsymbol{w}^\top \boldsymbol{\Sigma}_{\boldsymbol{\delta}} \boldsymbol{w}}{\sqrt{d}\,\|\boldsymbol{w}\|_2} \right) + \frac{\lambda}{2} r(\boldsymbol{w}) \tag{6}$$

M: Two things that I would like to be extremely sure of: 1. That the one that we have chose is actually the maximiser of the internal max 2. That the equivalent problem is still convex for a suitable choice of the matrices

K: Explicitly carry the 1/2 from the regularisation!

A. A few notes on convexity

We would like to prove that the problem in eq. (6) is still a convex problem in the components of $\boldsymbol{w}$.

ACKNOWLEDGEMENTS

[1] Jean Barbier, Florent Krzakala, Nicolas Macris, Léo Miolane, and Lenka Zdeborová. Optimal Errors and Phase Transitions in High-Dimensional Generalized Linear Models. *Proceedings of the National Academy of Sciences*, 116(12):5451–5460, March 2019. ISSN 0027-8424, 1091-6490. doi:10.1073/pnas.1802705116. URL `http://arxiv.org/abs/1708.03395`. arXiv:1708.03395 [cond-mat, physics:math-ph].

[2] Federica Gerace, Bruno Loureiro, Florent Krzakala, Marc Mézard, and Lenka Zdeborová. Generalisation error in learning with random features and the hidden manifold model*. *Journal of Statistical Mechanics: Theory and Experiment*, 2021(12): 124013, dec 2021. doi:10.1088/1742-5468/ac3ae6. URL `https://dx.doi.org/10.1088/1742-5468/ac3ae6`.

[3] Bruno Loureiro, Cédric Gerbelot, Hugo Cui, Sebastian Goldt, Florent Krzakala, Marc Mézard, and Lenka Zdeborová. Learning curves of generic features maps for realistic datasets with a teacher-student model*. *Journal of Statistical Mechanics: Theory and Experiment*, 2022(11):114001, nov 2022. doi:10.1088/1742-5468/ac9825. URL `https://dx.doi.org/10.1088/1742-5468/ac9825`.

[4] Martin Mächler. Accurately Computing log(1 exp(|a|)) Assessed by the Rmpfr package.

[5] 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.

[6] Fabian Pedregosa. How to Evaluate the Logistic Loss and not NaN trying, September 2019. URL `http://fa.bianp.net/blog/2019/evaluate_logistic/`. Section: coding.

Appendix A: Replica Computation

We start by defining the Gibbs measure over the weights $\boldsymbol{w}$

$$\mu_\beta(\mathrm{d}\boldsymbol{w}) = \frac{1}{\mathcal{Z}_\beta} e^{-\beta\left[\sum_{\mu=1}^n g\left(y^\mu, \boldsymbol{w}^\top \boldsymbol{x}^\mu, \boldsymbol{w}, \boldsymbol{\Sigma_\delta}, \varepsilon_t\right) + \frac{\lambda}{2}\boldsymbol{w}^T \boldsymbol{\Sigma_w} \boldsymbol{w}\right]} \mathrm{d}\boldsymbol{w} = \frac{1}{\mathcal{Z}_\beta} \underbrace{\prod_{\mu=1}^n e^{-\beta g\left(y^\mu, \boldsymbol{w}^\top \boldsymbol{x}^\mu, \boldsymbol{w}, \boldsymbol{\Sigma_\delta}, \varepsilon_t\right)}}_{P_g} \underbrace{e^{-\frac{\beta\lambda}{2}\boldsymbol{w}^T \boldsymbol{\Sigma_w} \boldsymbol{w}} \, \mathrm{d}w_i}_{P_w} \tag{A1}$$

Here, $\mathcal{Z}_\beta$, is the partition function that normalizes the Gibbs measure and it is given by

$$\mathcal{Z}_\beta = \int_{\mathbb{R}^d} \mathrm{d}\boldsymbol{w} e^{-\frac{\beta\lambda}{2}\boldsymbol{w}^T \boldsymbol{\Sigma_w} \boldsymbol{w}} \prod_{\mu=1}^n e^{-\beta g\left(y^\mu, \boldsymbol{w}^\top \boldsymbol{x}^\mu, \boldsymbol{w}, \boldsymbol{\Sigma_\delta}, \varepsilon_t\right)} \tag{A2}$$

You do need attention, but the the free energy density is truly all you need. In the zero temperature limit, $\beta \to \infty$ the Gibbs measure A1 concentrates around the solutions of the ERM problem. With the replica method, we can compute the free energy density, it is given by:

$$\beta f_\beta = -\lim_{d\to\infty} \frac{1}{d} \mathbb{E}_{\mathcal{D}} \log \mathcal{Z}_\beta \tag{A3}$$

It turns out that the free energy density which is equivalent to the free entropy up to a sign, is also equal to the conditional entropy density $\frac{1}{d}H(y \mid \boldsymbol{w})$ up to a sign, and also to the mutual information density between the data and the target labels $\frac{1}{d}I\left(x; x \mid \boldsymbol{w}\right)$ [1].

To start the computation, we need the Replica-trick: K: This is more of an analogy, in practice, if you say look at Bruno's work, you will find that for instance the $r \to 0$ limit is taken later when taking the quenched free entropy as a saddle point equation

$$\begin{aligned}
\lim_{d\to\infty} \frac{1}{d} \mathbb{E}_{\mathcal{D}} \log \mathcal{Z}_\beta &= \lim_{d\to\infty} \frac{1}{d} \mathbb{E}_{\mathcal{D}} \partial_r \left(1 + r \log \mathcal{Z}_\beta\right) \approx \lim_{d\to\infty} \frac{1}{d} \partial_r \mathbb{E}_{\mathcal{D}} e^{r \log \mathcal{Z}_\beta} \\
&= \lim_{d\to\infty} \frac{1}{d} \partial_r \mathbb{E}_{\mathcal{D}} e^{r \log \mathcal{Z}_\beta} = \lim_{r\to 0} \lim_{d\to\infty} \frac{1}{d} \frac{\partial_r \mathbb{E}_{\mathcal{D}} \mathcal{Z}^r}{1} \\
&= \lim_{r\to 0} \lim_{d\to\infty} \frac{1}{d} \frac{\mathbb{E}_{\mathcal{D}} \mathcal{Z}^r}{r}
\end{aligned} \tag{A4}$$

M: I already have a problem with what is written in [3]. From what I knew the Replica trick is each one of the following

$$\overline{\log Z} = \lim_{n\to 0} \frac{\overline{Z^n} - 1}{n} = \lim_{n\to 0} \frac{\log\left(\overline{Z^n}\right)}{n} = \lim_{n\to 0} \partial_n \overline{Z^n} \tag{A5}$$

but it seems that they use the derivative divided by n. With the last one of the previous equation I can make sense of eq. (A24), otherwise not.

Note that we introduced three limits up to here. The first is the zero temperature limit ensuring that we find the ground state of our Gibbs measure which corresponds to the minimum of our ERM problem. The second is the thermodynamic limit of very large dimension whilst keeping the sampling ratio fixed. And the third limit stems from the replica trick allowing us to compute the logarithm of the partition function, it corresponds to setting the number of replicated systems to zero.

This computation follows for the first part the one in [3]. So we start with the initial definition of replicated partition function the difference we have in our case is that we have a dependence on ε_t on the output probability.

$$\begin{aligned}
\mathbb{E}_{\mathcal{D}} \mathcal{Z}_\beta^r &= \prod_{\mu=1}^n \mathbb{E}_{\boldsymbol{x}^\mu} \prod_{a=1}^r \int_{\mathbb{R}^d} P_w\left(\mathrm{d}\boldsymbol{w}^a\right) P\left(y^\mu \mid \frac{\boldsymbol{x}^\mu \cdot \boldsymbol{w}^a}{\sqrt{d}}\right) \\
&= \prod_{\mu=1}^n \int_{\mathbb{R}} \mathrm{d}y^\mu \int_{\mathbb{R}^p} P_{\boldsymbol{\theta}_0}\left(\mathrm{d}\boldsymbol{\theta}_0\right) \int_{\mathbb{R}^{d\times r}} \left(\prod_{a=1}^r P_w\left(\mathrm{d}\boldsymbol{w}^a\right)\right) \mathbb{E}_{\boldsymbol{x}^\mu}\left[P_0\left(y^\mu \mid \frac{\boldsymbol{x}^\mu \cdot \boldsymbol{\theta}_0}{\sqrt{p}}\right) \prod_{a=1}^r P_g\left(y^\mu \mid \frac{\boldsymbol{x}^\mu \cdot \boldsymbol{w}^a}{\sqrt{d}}, \boldsymbol{\Sigma_\delta}, \boldsymbol{w}^a, \varepsilon_t\right)\right]
\end{aligned} \tag{A6}$$

explicitly we have that the term in P_g is

$$P_g\left(y^\mu \mid \frac{\boldsymbol{x}^\mu\cdot\boldsymbol{w}^a}{\sqrt{d}}, \boldsymbol{\Sigma_\delta}, \boldsymbol{w}^a, \varepsilon_t\right) = \frac{\sqrt{\beta}}{\sqrt{2\pi}}\exp\left(-\beta g\left(y\frac{\boldsymbol{x}^\mu\cdot\boldsymbol{w}^a}{\sqrt{d}} - \frac{\varepsilon_t}{\sqrt{d}}\frac{\boldsymbol{w}^a\boldsymbol{\Sigma}_{\boldsymbol{\delta}}^{-1/2}\boldsymbol{w}^a}{\|\boldsymbol{w}^a\|_2}\right)\right) \tag{A7}$$

the last part it is equal to:

$$\begin{aligned}&\mathbb{E}_{\boldsymbol{x}^\mu}\left[P_0\left(y^\mu \mid \frac{\boldsymbol{x}^\mu\cdot\boldsymbol{\theta}_0}{\sqrt{p}}\right)\prod_{a=1}^r P_g\left(y^\mu \mid \frac{\boldsymbol{x}^\mu\cdot\boldsymbol{w}^a}{\sqrt{d}}, \boldsymbol{\Sigma_\delta}, \boldsymbol{w}^a, \varepsilon_t\right)\right]\\ &= \int_{\mathbb{R}} \mathrm{d}\nu_\mu P_0\left(y \mid \nu_\mu\right)\int_{\mathbb{R}^r}\left(\prod_{a=1}^r \mathrm{d}\lambda_\mu^a P_g\left(y^\mu \mid \lambda_\mu^a, \boldsymbol{\Sigma_\delta}, \boldsymbol{w}^a, \varepsilon_t\right)\right)\mathbb{E}_{\boldsymbol{x}^\mu}\left[\delta\left(\nu_\mu - \frac{\boldsymbol{x}^\mu\cdot\boldsymbol{\theta}_0}{\sqrt{p}}\right)\prod_{a=1}^r\delta\left(\lambda_\mu^a - \frac{\boldsymbol{x}^\mu\cdot\boldsymbol{w}^a}{\sqrt{d}}\right)\right]\end{aligned} \tag{A8}$$

We can still perform the average over the dataset. We have that the new variables will behave again as Gaussians with the following covariances:

$$\rho \equiv \mathbb{E}\left[\nu_\mu^2\right] = \frac{1}{d}\boldsymbol{\theta}_0^\top\boldsymbol{\Sigma_x}\boldsymbol{\theta}_0, \quad m^a \equiv \mathbb{E}\left[\lambda_\mu^a\nu_\mu\right] = \frac{1}{d}\boldsymbol{\theta}_0^\top\boldsymbol{\Sigma_x}\boldsymbol{w}^a, \quad Q^{ab} \equiv \mathbb{E}\left[\lambda_\mu^a\lambda_\mu^b\right] = \frac{1}{d}\boldsymbol{w}^{a\top}\boldsymbol{\Sigma_x}\boldsymbol{w}^b \tag{A9}$$

where one can organise them in a single covariance matrix.

Now we want to perform several change of variables. The first one is the one in the matrix of overlaps:

$$\begin{aligned}1 &\propto \int_{\mathbb{R}}\mathrm{d}\rho\,\delta\left(d\rho - \boldsymbol{\theta}_0^\top\boldsymbol{\Sigma_x}\boldsymbol{\theta}_0\right)\int_{\mathbb{R}^r}\prod_{a=1}^r\mathrm{d}m^a\,\delta\left(dm^a - \boldsymbol{\theta}_0^\top\boldsymbol{\Sigma_x}\boldsymbol{w}^a\right)\int_{\mathbb{R}^{r\times r}}\prod_{1\le a\le b\le r}\mathrm{d}Q^{ab}\,\delta\left(dQ^{ab} - \boldsymbol{w}^{a\top}\boldsymbol{\Sigma_x}\boldsymbol{w}^b\right)\\ &= \int_{\mathbb{R}}\frac{\mathrm{d}\rho\,\mathrm{d}\hat{\rho}}{2\pi}e^{i\hat{\rho}\left(d\rho - \boldsymbol{\theta}_0^\top\boldsymbol{\Sigma_x}\boldsymbol{\theta}_0\right)}\int_{\mathbb{R}^r}\prod_{a=1}^r\frac{\mathrm{d}m^a\,\mathrm{d}\hat{m}^a}{2\pi}e^{i\sum_{a=1}^r\hat{m}^a\left(dm^a - \boldsymbol{\theta}_0^\top\boldsymbol{\Sigma_x}\boldsymbol{w}^a\right)}\int_{\mathbb{R}^{r\times r}}\prod_{1\le a\le b\le r}\frac{\mathrm{d}Q^{ab}\,\mathrm{d}\hat{Q}^{ab}}{2\pi}e^{i\hat{Q}^{ab}\left(dQ^{ab} - \boldsymbol{w}^{a\top}\boldsymbol{\Sigma_x}\boldsymbol{w}^b\right)}\end{aligned} \tag{A10}$$

the other one is

$$\begin{aligned}1 &\propto \int\prod_{a=1}^r\mathrm{d}A^a\,\delta(dA^a - \boldsymbol{w}^a\boldsymbol{\Sigma_\delta}\boldsymbol{w}^a)\int\prod_{1\le a\le b\le r}\mathrm{d}N^{ab}\,\delta\big(dN^{ab} - \boldsymbol{w}^a\cdot\boldsymbol{w}^a\big)\\ &= \int\prod_{a=1}^r\frac{\mathrm{d}A^a\,\mathrm{d}\hat{A}^a}{2\pi}e^{i\hat{A}^a(dA^a - \boldsymbol{w}^a\boldsymbol{\Sigma_\delta}\boldsymbol{w}^a)}\int\prod_{1\le a\le b\le r}\frac{\mathrm{d}N^{ab}\,\mathrm{d}\hat{N}^{ab}}{2\pi}e^{i\hat{N}^{ab}(dN^{ab} - \boldsymbol{w}^a\cdot\boldsymbol{w}^a)}\end{aligned} \tag{A11}$$

We finally can write our replicated partition function as the integral of a functional as follows

$$\mathbb{E}_{\mathcal{D}}\mathcal{Z}_\beta^r = \int\frac{\mathrm{d}\rho\,\mathrm{d}\hat{\rho}}{2\pi}\prod_{a=1}^r\frac{\mathrm{d}A^a\,\mathrm{d}\hat{A}^a}{2\pi}\frac{\mathrm{d}m^a\,\mathrm{d}\hat{m}^a}{2\pi}\prod_{a\le b=1}^r\frac{\mathrm{d}Q^{ab}\,\mathrm{d}\hat{Q}^{ab}}{2\pi}\frac{\mathrm{d}N^{ab}\,\mathrm{d}\hat{N}^{ab}}{2\pi}e^{d\Phi^{(r)}} \tag{A12}$$

where the r times replicated functional $\Phi^{(r)}$ is

$$\begin{aligned}\Phi^{(r)} = &-\rho\hat{\rho} - \sum_{a=1}^r m^a\hat{m}^a - \sum_{1\le a\le b\le r}Q^{ab}\hat{Q}^{ab} - \sum_{a=1}^r A^a\hat{A}^a - \sum_{1\le a\le b\le r}N^{ab}\hat{N}^{ab}\\ &+\alpha\Psi_y^{(r)}\left(\rho, m^a, Q^{ab}, A^a, N^{ab}\right) + \Psi_w^{(r)}\left(\hat{\rho}, \hat{m}^a, \hat{Q}^{ab}, \hat{A}^a, \hat{N}^{ab}\right)\end{aligned} \tag{A13}$$

we will refer to the elements in the first line of eq. (A13) as the trace term. We also have defined the prior part of the free energy Ψ_w to be

$$\Psi_w^{(r)} = \frac{1}{d}\log\left[\int_{\mathbb{R}^p}P_{\boldsymbol{\theta}_0}\left(\mathrm{d}\boldsymbol{\theta}_0\right)\int_{\mathbb{R}^{d\times r}}\prod_{a=1}^r P_w\left(\mathrm{d}\boldsymbol{w}^a\right)e^{\hat{\rho}\boldsymbol{\theta}_0^\top\boldsymbol{\Sigma_x}\boldsymbol{\theta}_0 + \sum_{a=1}^r\left(\hat{m}^a\boldsymbol{\theta}_0^\top\boldsymbol{\Sigma_x}\boldsymbol{w}^a + \hat{A}^a\boldsymbol{w}^a\boldsymbol{\Sigma_\delta}\boldsymbol{w}^a\right) + \sum_{1\le a\le b\le r}\left(\hat{Q}^{ab}\boldsymbol{w}^{a\top}\boldsymbol{\Sigma_x}\boldsymbol{w}^b + \hat{N}^{ab}\boldsymbol{w}^a\cdot\boldsymbol{w}^b\right)}\right] \tag{A14}$$

and the channel part of the free energy Ψ_y as

$$\Psi_y^{(r)} = \log\left[\int_{\mathbb{R}} \mathrm{d}y \int_{\mathbb{R}} \mathrm{d}\nu P_0(y \mid \nu) \int \prod_{a=1}^{r} \mathrm{d}\lambda^a P_g\left(y \mid \lambda^a, A^a, N^{ab}, \varepsilon_t\right) \mathcal{N}\left(\nu, \lambda^a; \mathbf{0}, \Sigma^{ab}\right)\right] \tag{A15}$$

where we have factored out the fact that $(\nu_\mu, \lambda_\mu)\ \mu = 1...n$ factors over all the data points.

In the thermodynamic limit where $d \to \infty$ with n/d fixed, the integral in eq. (A12) concentrates around the values of the overlap parameters that extremize the free entropy $\Phi^{(r)}$ and hence we can get the free energy density as:

$$\beta f_\beta = -\lim_{r\to 0^+} \frac{1}{r} \operatorname{extr} \Phi^{(r)} = -\lim_{r\to 0^+} \partial_r \operatorname{extr} \Phi^{(r)} \tag{A16}$$

1. Replica Symmetric Ansatz

We propose the following Ansatz for the variables that we have to extremise over

$$\begin{array}{llll} m^a = m & \hat{m}^a = \hat{m} & \text{for } a = 1, \cdots, r \\ q^{aa} = Q & \hat{q}^{aa} = -\frac{1}{2}\hat{Q} & \text{for } a = 1, \cdots, r \\ q^{ab} = q & \hat{q}^{ab} = \hat{q} & \text{for } 1 \le a < b \le r \\ A^a = A & \hat{A}^a = \hat{A} & \text{for } a = 1, \cdots, r \\ N^a = N & \hat{N}^a = \hat{N} & \text{for } a = 1, \cdots, r \end{array} \tag{A17}$$

M: We should add a justification as in [2] that explains why this ansatz is correct.

M: We should find a justification that tells us that we don't need A^{ab} nor N^{ab}. This justification imo could just be because it works :)

The trace term becomes

$$\rho\hat{\rho} + rm\hat{m} + \frac{r(r-1)}{2} q\hat{q} - \frac{r}{2} Q\hat{Q} + rA\hat{A} + rN\hat{N} \tag{A18}$$

As in [2] we should check that the ansatz is well defined and for that to be the case we have that we obtain

$$\rho = \mathbb{E}_{\boldsymbol{\theta}_0}\left[\boldsymbol{\theta}_0^2\right], \quad \hat{\rho} = 0 \tag{A19}$$

Now we take the limit $r \to 0$ after dividing the trace term by r and defining $V = Q - q$ and $\hat{V} = \hat{Q} + \hat{q}$

$$\begin{aligned} &-\frac{1}{2}(q\hat{V} - \hat{q}V) + m\hat{m} + A\hat{A} + N\hat{N} \\ &-\frac{1}{2}(q\hat{q} + Q\hat{Q}) + m\hat{m} + A\hat{A} + N\hat{N} = \end{aligned} \tag{A20}$$

the role of these definitions will be clear afterwards. K: This is not correct, there is a VV term missing. M: Yes, I agree but in the end it should not matter since it is constant in β.

Thus we can proceed plug these ansätze inside eqs. (A14) and (A15) we obtain the following

$$\Psi_w^{(r)} = \frac{1}{d} \log\left[\int_{\mathbb{R}^p} P_{\boldsymbol{\theta}_0}\left(\mathrm{d}\boldsymbol{\theta}_0\right) \int_{\mathbb{R}^{d\times r}} \prod_{a=1}^{r} P_w\left(\mathrm{d}\boldsymbol{w}^a\right) e^{\hat{\rho}\boldsymbol{\theta}_0^\top \boldsymbol{\Sigma}_{\boldsymbol{x}} \boldsymbol{\theta}_0 + \hat{m}\sum_{a=1}^r \boldsymbol{\theta}_0^\top \boldsymbol{\Sigma}_{\boldsymbol{x}} \boldsymbol{w}^a - \frac{1}{2}\hat{Q}\sum_{1\le a<b\le r} \boldsymbol{w}^{a\top} \boldsymbol{\Sigma}_{\boldsymbol{x}} \boldsymbol{w}^b + \hat{N}\sum_{a=1}^r \boldsymbol{w}^a\cdot\boldsymbol{w}^a + \hat{A}\sum_{a=1}^r \boldsymbol{w}^a \boldsymbol{\Sigma}_{\boldsymbol{\delta}} \boldsymbol{w}^a}\right] \tag{A21}$$

to perform in the following the $r \to 0^+$ limit we can change a bit the integral by factoring out all the terms. To perform this simplificaiton we will define $\hat{V} = \hat{Q} + \hat{q}$ and the multidimansional Hubbard-Stratonovic identity which reads

$$e^{\hat{q}\sum_{a,b=1}^r \boldsymbol{w}^{a\top} \boldsymbol{\Sigma}_{\boldsymbol{x}} \boldsymbol{w}^b} = \mathbb{E}_{\boldsymbol{\xi}}\left[e^{\boldsymbol{\xi}^\top \sqrt{\hat{q}} \boldsymbol{\Sigma}_{\boldsymbol{x}}^{1/2} \sum_{a=1}^r \boldsymbol{w}^a}\right] \tag{A22}$$

where $\boldsymbol{\xi} \sim \mathcal{N}(\mathbf{0}, 1_d)$. Thus by calling the part inside the log in eq. (A21) with the letter $\mathcal{A}$ we have that

$$\begin{aligned}
\mathcal{A} &= \int_{\mathbb{R}^d} P_{\boldsymbol{\theta}_0}(\mathrm{d}\boldsymbol{\theta}_0)\, e^{\hat{\rho}\boldsymbol{\theta}_0^\top \boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{\theta}_0} \int_{\mathbb{R}^{d\times r}} \prod_{a=1}^{r} P_w(\mathrm{d}\boldsymbol{w}^a)\, e^{\hat{m}\sum_{a=1}^r \boldsymbol{\theta}_0^\top \boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{w}^a - \frac{1}{2}\hat{Q}\sum_{1\le a<b\le r}\boldsymbol{w}^{a\top}\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{w}^b + \hat{N}\sum_{a=1}^r \boldsymbol{w}^a\cdot\boldsymbol{w}^a + \hat{A}\sum_{a=1}^r \boldsymbol{w}^a\boldsymbol{\Sigma}_{\boldsymbol{\delta}}\boldsymbol{w}^a} \\
&= \int_{\mathbb{R}^d} P_{\boldsymbol{\theta}_0}(\mathrm{d}\boldsymbol{\theta}_0)\, e^{\hat{\rho}\boldsymbol{\theta}_0^\top \boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{\theta}_0} \int_{\mathbb{R}^{d\times r}} \prod_{a=1}^{r} P_w(\mathrm{d}\boldsymbol{w}^a)\, e^{\sum_{a=1}^r \left(-\frac{\hat{V}}{2}\boldsymbol{w}^a\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{w}^a + \hat{A}\boldsymbol{w}^a\boldsymbol{\Sigma}_{\boldsymbol{\delta}}\boldsymbol{w}^a + \hat{N}\boldsymbol{w}^a\boldsymbol{w}^a + \hat{m}\boldsymbol{\theta}_0\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{w}^a\right) + \hat{q}\sum_{a,b=1}^r \boldsymbol{w}^a\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{w}^b} \\
&= \int_{\mathbb{R}^d} P_{\boldsymbol{\theta}_0}(\mathrm{d}\boldsymbol{\theta}_0)\, e^{\hat{\rho}\boldsymbol{\theta}_0^\top \boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{\theta}_0} \int_{\mathbb{R}^{d\times r}} \prod_{a=1}^{r} P_w(\mathrm{d}\boldsymbol{w}^a)\, \mathbb{E}_{\boldsymbol{\xi}}\left[e^{-\frac{\hat{V}}{2}\boldsymbol{w}^a\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{w}^a + \hat{A}\boldsymbol{w}^a\boldsymbol{\Sigma}_{\boldsymbol{\delta}}\boldsymbol{w}^a + \hat{N}\boldsymbol{w}^a\boldsymbol{w}^a + \boldsymbol{w}^a\left(\hat{m}\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{\theta}_0 + \sqrt{\hat{q}}\boldsymbol{\Sigma}_{\boldsymbol{x}}^{1/2}\boldsymbol{\xi}\right)}\right] \\
&= \mathbb{E}_{\boldsymbol{\xi}}\left[\int_{\mathbb{R}^d} P_{\boldsymbol{\theta}_0}(\mathrm{d}\boldsymbol{\theta}_0)\, e^{\hat{\rho}\boldsymbol{\theta}_0^\top \boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{\theta}_0}\left[\int_{\mathbb{R}^d} P_w(\mathrm{d}\boldsymbol{w})\, e^{-\frac{\hat{V}}{2}\boldsymbol{w}\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{w} + \hat{A}\boldsymbol{w}\boldsymbol{\Sigma}_{\boldsymbol{\delta}}\boldsymbol{w} + \hat{N}\boldsymbol{w}^\top\boldsymbol{w} + \boldsymbol{w}\left(\hat{m}\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{\theta}_0 + \sqrt{\hat{q}}\boldsymbol{\Sigma}_{\boldsymbol{x}}^{1/2}\boldsymbol{\xi}\right)}\right]^r\right]
\end{aligned} \tag{A23}$$

Then we can take the derivative and limit and obtain

$$\Psi_w = \lim_{r\to 0^+}\Psi_w^{(r)} = \frac{1}{d}\mathbb{E}_{\boldsymbol{\xi},\boldsymbol{\theta}_0}\left[\log\int_{\mathbb{R}^d} P_w(\mathrm{d}\boldsymbol{w})\, e^{-\frac{\hat{V}}{2}\boldsymbol{w}\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{w} + \hat{A}\boldsymbol{w}\boldsymbol{\Sigma}_{\boldsymbol{\delta}}\boldsymbol{w} + \hat{N}\boldsymbol{w}^\top\boldsymbol{w} + \boldsymbol{w}\left(\hat{m}\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{\theta}_0 + \sqrt{\hat{q}}\boldsymbol{\Sigma}_{\boldsymbol{x}}^{1/2}\boldsymbol{\xi}\right)}\right] \tag{A24}$$

where we still need to take the limit $d \to \infty$.

Now we can focus on the channel term and rewrite it in a more suitable way for taking the $r \to 0^+$ limit. In a very similar fashion as before we would like to simplify

$$\Psi_y^{(r)} = \log\left[\int_{\mathbb{R}}\mathrm{d}y\int_{\mathbb{R}}\mathrm{d}\nu\, P_0(y\mid\nu)\int\prod_{a=1}^{r}\mathrm{d}\lambda^a\, P_g(y\mid\lambda^a, A, N, \varepsilon_t)\,\mathcal{N}\left(\nu, \lambda^a; \mathbf{0}, \Sigma^{ab}\right)\right] \tag{A25}$$

We will indicate the argument of the log with $\mathcal{B}$. Additionally we have that the martix of covariances is

$$\Sigma = \begin{pmatrix} \rho & m & m & \dots & m \\ m & Q & q & \dots & q \\ m & q & Q & \dots & q \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ m & q & q & \dots & Q \end{pmatrix} \tag{A26}$$

and in addition also the inverse matrix has a Replica Symmetric Structure which is given from the following elements

$$\begin{aligned}
\left(\Sigma^{-1}\right)_{00} &\equiv \tilde{\rho} = \frac{Q + (r-1)q}{\rho(Q+(r-1)q) - rm^2}\,, & \left(\Sigma^{-1}\right)_{0a} &\equiv \tilde{m} = \frac{m}{rm^2 - \rho(Q+(r-1)q)}\,, \\
\left(\Sigma^{-1}\right)_{aa} &\equiv \tilde{Q} = \frac{\rho(Q+(r-2)q) - (r-1)m^2}{(Q-q)\left(\rho(Q+(r-1)q) - rm^2\right)}\,, & \left(\Sigma^{-1}\right)_{ab} &\equiv \tilde{q} = \frac{m^2 - \rho q}{(Q-q)\left(\rho(Q+(r-1)q) - rm^2\right)}
\end{aligned} \tag{A27}$$

and thus there is an implicit dependence on r in the covariance. To check that the inverse matrix has a RS structure as well one can think of the formula that is used to evaluate the inverse of a matrix from the cofactors.

Also we look at the determinant of the matrix. There are three different eigenvalue types

$$\begin{aligned}
&\lambda_1 = Q - q\,, \qquad \lambda_2 = \frac{1}{2m}\left(-Q - q(r-1) + \rho - \tilde{\Delta}\right), \qquad \lambda_3 = \frac{1}{2m}\left(-Q - q(r-1) + \rho + \tilde{\Delta}\right), \\
&d_1 = r-1\,, \qquad d_2 = 1\,, \qquad\qquad\qquad\qquad d_3 = 1\,,
\end{aligned} \tag{A28}$$

with $\tilde{\Delta} = \sqrt{4m^2 r + (Q + q(r-1) - \rho)^2}$ and thus one obtains the determinant. More explicitly we have that

$$\det(2\pi\Sigma) = (2\pi)^{r+1}(Q-q)^{r-1}(2m)^{-4}(-Q - q(r-1) + \rho - \tilde{\Delta})(-Q - q(r-1) + \rho + \tilde{\Delta}) \tag{A29}$$

Thus we have that

$$\begin{aligned}
\mathcal{B} &= \int_{\mathbb{R}}\mathrm{d}y\int_{\mathbb{R}}\mathrm{d}\nu\, P_0(y\mid\nu)e^{-\frac{1}{2}\tilde{\rho}\nu^2}\int\prod_{a=1}^{r}\mathrm{d}\lambda^a\, P_g(y\mid\lambda^a, A, N, \varepsilon_t)\, e^{-\tilde{m}\nu\sum_{a=1}^r\lambda^a - \frac{1}{2}\tilde{Q}\sum_{a=1}^r(\lambda^a)^2 - \frac{1}{2}\tilde{q}\sum_{1\le a<b\le r}\lambda^a\lambda^b - \frac{1}{2}\log\det(2\pi\Sigma)} \\
&= \mathbb{E}_\xi\int_{\mathbb{R}}\mathrm{d}y\, e^{-\frac{1}{2}\log\det(2\pi\Sigma)}\int_{\mathbb{R}}\mathrm{d}\nu\, P_0(y\mid\nu)e^{-\frac{1}{2}\tilde{\rho}\nu^2}\left[\int\mathrm{d}\lambda\, P_g(y\mid\lambda, A, N, \varepsilon_t)\, e^{-\frac{\tilde{Q}-\tilde{q}}{2}\lambda^2 + (\sqrt{-\tilde{q}}\xi - \tilde{m}\nu)\lambda}\right]^r
\end{aligned} \tag{A30}$$

Now we can follow a similar procedure as before and define $V = Q - q$ we have that and the limit is

$$\Psi_y = \lim_{r\to 0^+} \Psi_y^{(r)} = \mathbb{E}_\xi \left[\int_{\mathbb{R}} \mathrm{d}y \int \frac{\mathrm{d}\nu}{\sqrt{2\pi\rho}} P_0(y \mid \nu) e^{-\frac{1}{2\rho}\nu^2} \log \left[\int \frac{\mathrm{d}\lambda}{\sqrt{2\pi}} P_y(y \mid \lambda, A, N, \varepsilon_t) e^{-\frac{1}{2}\frac{\lambda^2}{V} + \left(\frac{\sqrt{q-m^2/\rho}}{V}\xi + \frac{m/\rho}{V}\nu\right)\lambda} \right] \right] - \frac{1}{2}\log V - \frac{1}{2}\frac{q}{V} \tag{A31}$$

the term outside with a log gives the normalisation constant and the other term

We would like to rewrite the quantities with the help of the following definition

$$\mathcal{Z}_0(y, \omega, V) = \int \frac{\mathrm{d}x}{\sqrt{2\pi V}} e^{-\frac{1}{2V}(x-\omega)^2} P_0(y \mid x)\,, \quad \mathcal{Z}_y(y, \omega, V, A, N) = \int \frac{\mathrm{d}x}{\sqrt{2\pi V}} e^{-\frac{1}{2V}(x-\omega)^2} P_y(y \mid x, A, N, \varepsilon_t) \tag{A32}$$

thus the first step is to complete the square to have

M: I am pretty sure that the result is the following but I am not sure how to show it lol. K: I would start by looking at section I.3 in Aubin. He cites also Barbier and introduces denoising functions related to AMP. At first glance the multi-class perceptron of Cornacchia et al. might also be useful.

$$\mathbb{E}_\xi \left[\int_{\mathbb{R}} \mathrm{d}y\, \mathcal{Z}_0\left(y, \frac{m}{\sqrt{q}}\xi, \rho - \frac{m^2}{q}\right) \log \mathcal{Z}_y(y, \sqrt{q}\xi, V, A, N) \right] \tag{A33}$$

Now there are two things that we still need to do : find the form for the prior term and take the limit $\beta \to \infty$.

2. Prior term for ℓ_2 regularisation

To be as general as possible we would like to include the case of a possible non isotropic regularisation. Thus

$$P_w(\mathrm{d}\boldsymbol{w}) = \frac{1}{(2\pi)^{d/2}} \exp\left(-\frac{\beta\lambda}{2}\boldsymbol{w}\boldsymbol{\Sigma}_{\boldsymbol{w}}\boldsymbol{w}\right)\mathrm{d}\boldsymbol{w} \tag{A34}$$

We want to calculate the term inside the log in eq. (A24)

$$\int_{\mathbb{R}^d} P_w(\mathrm{d}\boldsymbol{w}) e^{-\frac{\hat{V}}{2}\boldsymbol{w}\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{w} + \hat{A}\boldsymbol{w}\boldsymbol{\Sigma}_{\boldsymbol{\delta}}\boldsymbol{w} + \hat{N}\boldsymbol{w}^\top\boldsymbol{w} + \boldsymbol{w}\left(\hat{m}\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{\theta}_0 + \sqrt{\hat{q}}\boldsymbol{\Sigma}_{\boldsymbol{x}}^{1/2}\boldsymbol{\xi}\right)} = \frac{\exp\left(\frac{1}{2}\left(\hat{m}\Phi^\top\boldsymbol{\theta}_0 + \sqrt{\hat{q}}\Omega^{1/2}\boldsymbol{\xi}\right)^\top \boldsymbol{\Lambda}^{-1}\left(\hat{m}\Phi^\top\boldsymbol{\theta}_0 + \sqrt{\hat{q}}\Omega^{1/2}\boldsymbol{\xi}\right)^\top\right)}{\sqrt{\det\boldsymbol{\Lambda}}} \tag{A35}$$

where we defined $\boldsymbol{\Lambda} = \beta\lambda\boldsymbol{\Sigma}_{\boldsymbol{w}} + \hat{V}\boldsymbol{\Sigma}_{\boldsymbol{x}} + \hat{A}\boldsymbol{\Sigma}_{\boldsymbol{\delta}} + \hat{N}\boldsymbol{I}$. Now the prior term becomes after taking the log and using an identity K: note the identity for the determinant.

$$\begin{aligned}\Psi_w &= \frac{1}{2d}\mathbb{E}_{\boldsymbol{\xi},\boldsymbol{\theta}_0}\left[\frac{1}{2}\left(\hat{m}\Phi^\top\boldsymbol{\theta}_0 + \sqrt{\hat{q}}\Omega^{1/2}\boldsymbol{\xi}\right)^\top\boldsymbol{\Lambda}^{-1}\left(\hat{m}\Phi^\top\boldsymbol{\theta}_0 + \sqrt{\hat{q}}\Omega^{1/2}\boldsymbol{\xi}\right)^\top\right] - \frac{1}{2d}\operatorname{tr}\log\boldsymbol{\Lambda} \\ &= \frac{1}{2d}\operatorname{tr}\left[\left(\hat{m}^2\Phi^\top\boldsymbol{\theta}_0\boldsymbol{\theta}_0^\top\Phi + \hat{q}\Omega\right)\boldsymbol{\Lambda}^{-1}\right] - \frac{1}{2d}\operatorname{tr}\log\boldsymbol{\Lambda}\end{aligned} \tag{A36}$$

3. Saddle-point equations

As we pre-announced we would like to find the stationary values that dominate the intergal and to do so we should derive the exponent with respect to all the order parameters.

The saddle points that depend on $m, q, V, \hat{m}, \hat{q}$ and $\hat{V}$ are of a similar form as those found already in [3]. We need thus to derive with respect to $A, N, \hat{A}$ and $\hat{N}$.

We start by taking the derivative wrt $\hat{A}$. By the definition of $\mathcal{Z}$

$$\begin{aligned}\partial_A \mathcal{Z}_y(y, \omega, V) &= \frac{\partial}{\partial A}\sqrt{\beta}\int \frac{\mathrm{d}x}{\sqrt{2\pi V}} e^{-\frac{1}{2V}(x-\omega)^2} \frac{e^{-\beta g\left(yx - \varepsilon_t A/\sqrt{N}\right)}}{\sqrt{2\pi}} \\ &= \sqrt{\beta}\int \frac{\mathrm{d}x}{\sqrt{2\pi V}} e^{-\frac{1}{2V}(x-\omega)^2} e^{-\beta g\left(yx - \varepsilon_t A/\sqrt{N}\right)} \left(\beta\frac{\varepsilon_t}{\sqrt{N}} g'\left(yx - \varepsilon_t\frac{A}{\sqrt{N}}\right)\right)\end{aligned} \tag{A37}$$

We also have for the derivative with respect to N is

$$\begin{aligned}\partial_N \mathcal{Z}_y(y,\omega,V) &= \frac{\partial}{\partial N}\sqrt{\beta}\int \frac{\mathrm{d}x}{\sqrt{2\pi V}} e^{-\frac{1}{2V}(x-\omega)^2} \frac{e^{-\beta g\left(yx-\varepsilon_t A/\sqrt{N}\right)}}{\sqrt{2\pi}} \\ &= \sqrt{\beta}\int \frac{\mathrm{d}x}{\sqrt{2\pi V}} e^{-\frac{1}{2V}(x-\omega)^2} e^{-\beta g\left(yx-\varepsilon_t A/\sqrt{N}\right)} \left(-\beta \frac{\varepsilon_t A}{2N^{3/2}} g'\left(yx-\varepsilon_t \frac{A}{\sqrt{N}}\right)\right)\end{aligned} \tag{A38}$$

The factor $\sqrt{\beta}$ in front of everything is taken care with the temperature scalings of the order parameter V chosen in eq. (A42).

Also we will need the derivate of the term Ψ_w with respect to the hat-variables $\hat{A}$ and $\hat{N}$

$$\partial_{\hat{A}}\Psi_w = \frac{1}{2d}\,\mathrm{tr}\left[\left(\hat{m}^2\Phi^\top \boldsymbol{\theta}_0\boldsymbol{\theta}_0^\top\Phi + \hat{q}\Omega\right)\boldsymbol{\Sigma}_{\boldsymbol{\delta}}\boldsymbol{\Lambda}^{-2}\right] \tag{A39}$$

and

$$\partial_{\hat{N}}\Psi_w = \frac{1}{2d}\,\mathrm{tr}\left[\left(\hat{m}^2\Phi^\top \boldsymbol{\theta}_0\boldsymbol{\theta}_0^\top\Phi + \hat{q}\Omega\right)\boldsymbol{\Lambda}^{-2}\right] \tag{A40}$$

with the same definition of $\boldsymbol{\Lambda}$ as in the previous section. M: In the previous one I am not ultra sure of the factor 2 in red

$$\begin{aligned}\hat{A} &= -\alpha\partial_A\Psi_y \qquad A = -\partial_{\hat{A}}\Psi_w \\ \hat{N} &= -\alpha\partial_N\Psi_y \qquad N = -\partial_{\hat{N}}\Psi_w\end{aligned} \tag{A41}$$

4. Zero temperature limit

We now need to take the zero temperature limit for this case. The explicit scalings of the parameters are

$$\begin{aligned} V &\to \beta^{-1}V \quad & q &\to q \quad & m &\to m \quad & A &\to \beta^2 A \quad & N &\to \beta^2 N \\ \hat{V} &\to \beta\hat{V} \quad & \hat{q} &\to \beta^2\hat{q} \quad & \hat{m} &\to \beta\hat{m} \quad & \hat{A} &\to \beta\hat{A} \quad & \hat{N} &\to \beta\hat{N}\end{aligned} \tag{A42}$$

M: Is there another way to check these scalings make sense? Do they have some interpretation of some kind? What I have done right now is just that I have chosen the scalings such that $\partial\mathcal{Z}$ is finite in the $\beta\to\infty$ limit. K: For me there is also some stuff that is not clear. In Gerace they say that the scaling should be such that in the channel distribution the exp factor is beta * Loss, allowing us to find the ground state. But they also consider the limit in all derivatives of the channel distribution. The rest of the free entropy (trace and prior terms) remain unmentioned. Which leads me to question our saddle point equations.

The limit of the prior term is

$$\Psi_w = \lim_{\beta\to\infty}\frac{1}{\beta}\Psi_w = \frac{1}{2d}\,\mathrm{tr}\left[\left(\hat{m}^2\Phi^\top \boldsymbol{\theta}_0\boldsymbol{\theta}_0^\top\Phi + \hat{q}\Omega\right)\boldsymbol{\Lambda}^{-1}\right] \tag{A43}$$

and then the limit of the channel term becomes

$$\Psi_y = \lim_{\beta\to\infty}\frac{1}{\beta}\Psi_y = -\mathbb{E}_{\xi\sim\mathcal{N}(0,1)}\left[\int \mathrm{d}y\mathcal{Z}_0\left(y,\frac{m}{\sqrt{q}}\xi,\rho-\frac{m^2}{q}\right)\mathcal{M}_{Vg(y,\cdot;A,N,\varepsilon_t)}(\sqrt{q}\xi)\right] \tag{A44}$$

where $\mathcal{M}_{Vg(y,\cdot;A,N,\varepsilon_t)}$ is the Moreau envelope of the modified loss function defined in eq. (6) with the relevant quantities changed for their overlaps. On the other hand the limit of the quantities in eqs. (A37) and (A38) will be taken as evaluating $\propto g'(...)$ on the value of the proximal.

Now we can take the previous equations and take the limit $\beta\to\infty$. We report for completeness the whole set of equations

$$\begin{cases}\hat{V} = -\alpha\mathbb{E}_\xi\left[\int_{\mathbb{R}}\mathrm{d}y\,\mathcal{Z}_0\left(y,\sqrt{\eta}\xi,\rho-\eta\right)\partial_\omega f_g\left(y,\sqrt{q}\xi,V\right)\right] \\ \hat{q} = \alpha\mathbb{E}_\xi\left[\int_{\mathbb{R}}\mathrm{d}y\,\mathcal{Z}_0\left(y,\sqrt{\eta}\xi,\rho-\eta\right) f_g^2\left(y,\sqrt{q}\xi,V\right)\right] \\ \hat{m} = \alpha\mathbb{E}_\xi\left[\int_{\mathbb{R}}\mathrm{d}y\,\partial_\omega\mathcal{Z}_0\left(y,\sqrt{\eta}\xi,\rho-\eta\right) f_g\left(y,\sqrt{q}\xi,V\right)\right] \\ \hat{A} = -\alpha\frac{\varepsilon_t}{\sqrt{N}}\mathbb{E}_\xi\left[\int_{\mathbb{R}}\mathrm{d}y\,\mathcal{Z}_0\left(y,\sqrt{\eta}\xi,\rho-\eta\right) g'\left(y\mathcal{P}_{Vg}(...)-\varepsilon_t\frac{A}{N}\right)\right] \\ \hat{N} = \alpha\frac{\varepsilon_t}{2}\frac{A}{N^{3/2}}\mathbb{E}_\xi\left[\int_{\mathbb{R}}\mathrm{d}y\,\mathcal{Z}_0\left(y,\sqrt{\eta}\xi,\rho-\eta\right) g'\left(y\mathcal{P}_{Vg}(...)-\varepsilon_t\frac{A}{N}\right)\right]\end{cases} \tag{A45}$$

K: fg also depends on epsilon, A and N no? M: Yes, because $\mathcal{Z}_y$ does

$$\begin{cases} V = \frac{1}{d}\operatorname{tr}\boldsymbol{\Lambda}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{x}} \\ q = \frac{1}{d}\operatorname{tr}\left[\left(\hat{q}\boldsymbol{\Sigma}_{\boldsymbol{x}} + \hat{m}^2\boldsymbol{\Sigma}_{\boldsymbol{x}}^\top\boldsymbol{\theta}_0\boldsymbol{\theta}_0^\top\boldsymbol{\Sigma}_{\boldsymbol{x}}\right)\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{\Lambda}^{-2}\right] \\ m = \frac{\hat{m}}{d}\operatorname{tr}\boldsymbol{\Sigma}_{\boldsymbol{x}}^\top\boldsymbol{\theta}_0\boldsymbol{\theta}_0^\top\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{\Lambda}^{-1} \\ A = \frac{1}{d}\operatorname{tr}\left[\left(\hat{q}\boldsymbol{\Sigma}_{\boldsymbol{x}} + \hat{m}^2\boldsymbol{\Sigma}_{\boldsymbol{x}}^\top\boldsymbol{\theta}_0\boldsymbol{\theta}_0^\top\boldsymbol{\Sigma}_{\boldsymbol{x}}\right)\boldsymbol{\Sigma}_{\boldsymbol{\delta}}\boldsymbol{\Lambda}^{-2}\right] \\ N = \frac{1}{d}\operatorname{tr}\left[\left(\hat{q}\boldsymbol{\Sigma}_{\boldsymbol{x}} + \hat{m}^2\boldsymbol{\Sigma}_{\boldsymbol{x}}^\top\boldsymbol{\theta}_0\boldsymbol{\theta}_0^\top\boldsymbol{\Sigma}_{\boldsymbol{x}}\right)\boldsymbol{\Lambda}^{-2}\right] \end{cases} \tag{A46}$$

where we remember that $\boldsymbol{\Lambda} = \lambda\boldsymbol{\Sigma}_{\boldsymbol{w}} + \hat{V}\boldsymbol{\Sigma}_{\boldsymbol{x}} + \hat{A}\boldsymbol{\Sigma}_{\boldsymbol{\delta}} + \hat{N}\boldsymbol{I}$ and we defined $\eta = m^2/q$. Also the we have that $\mathcal{P}_{Vg}(...)$ indicates the proximal opertator of the loss function in the case of the adversarial attack.

5. Errors as a function of the overlaps

K: It would be cool to come up with a robustness measure in terms of overlaps. M: Indeed

a. *Generalisation Error*

b. *Training Error*

c. *Adversarial Training Error*

d. *Calibration*

K: If we will look at adversarial training at optimal regularization, it might be interesting to consider the calibration as well. M: Yes, calibration could be a spot-on measure of how much the network thinks he has been fooled in a sense, it measures how much the network is confindent in the prediction

K: Could we try to analyze some ERM or similar method to come up with good priors? I.e. the teacher prior has some covariance and we want to reasonably guess a prior as a student.ß

Appendix B: Numerical Resolution of the Fixed-Point equations

Finding a set of overlaps that satisfy eqs. (A45) and (A46) can be done with the help of a numerical fixed point iteration.

K: Currently the situation is this:

1. For the gaussian vanilla case we had in the TP, everything works fine using the magic we did with the Moreau-derivative
2. If we use the non-magic derivative for the channel as is currently described in this document, we have the correct relationship between the overlaps, as the generalization and training error is correct. The actual scaling of the overlaps is only wrong in the epsilon non-zero case.
3. if we move to a data-covariance all epsilon cases break down for both derivatives, only for the non-zero epsilon case. Further complicating the model does not help.
4. The ERM-code will only need slight adaptation for the $\boldsymbol{\Sigma}_{\boldsymbol{\delta}}$. For now it looks like there should be no need to rewrite the Cython code.

Hence, what we want to do next is

1. Understand the relationship between the two derivatives and why they might or might not break down
2. Look for some scaling issues, it might be that there are numerical advantages of normalizing everything to $\rho = 1$? It is sometimes considered a trick of the trade in nn's to keep the argument of an activation function standardized. Might apply here as well.

3. Double check our implementation of the hat equations

4. Double check out ERM implementation.

Appendix C: Numerical Empirical Risk Minimisation

Computing the loss, gradient and hessian of the logistic loss, let alone the adversarially perturbed logistic loss pose well-known numerical challenges [6]. In this section we explain in detail how to numerically evaluate the ERM-estimator. To begin, here is the loss function over a data set $(X_\mu, y_\mu)\mu = 1, ..., N$.

$$\mathcal{L} = \sum_{\mu=1}^{N} \log\left(1 + \exp\left(-y_\mu \frac{X_\mu^T \boldsymbol{w}}{\sqrt{d}} + \frac{\varepsilon_t \boldsymbol{w}^T \boldsymbol{\Sigma_\delta} \boldsymbol{w}}{\sqrt{d}\sqrt{\boldsymbol{w}^T \boldsymbol{w}}}\right)\right) \tag{C1}$$

To compute the gradient, we switch from a -1/1 labelling to a 0/1 labelling. We can write the following:

$$\begin{aligned}
\mathcal{L} &= \sum_{\mu=1}^{N} \log\left(1 + \exp\left(-y_\mu \frac{X_\mu^T \boldsymbol{w}}{\sqrt{d}} + \frac{\varepsilon_t \boldsymbol{w}^T \boldsymbol{\Sigma_\delta} \boldsymbol{w}}{\sqrt{d}\sqrt{\boldsymbol{w}^T \boldsymbol{w}}}\right)\right) \\
&= \sum_{\mu=1}^{N} \delta(y_\mu = +1) \log\left(1 + \exp\left(-\frac{X_\mu^T \boldsymbol{w}}{\sqrt{d}} + \frac{\varepsilon_t \boldsymbol{w}^T \boldsymbol{\Sigma_\delta} \boldsymbol{w}}{\sqrt{d}\sqrt{\boldsymbol{w}^T \boldsymbol{w}}}\right)\right) + \delta(y_\mu = -1) \log\left(1 + \exp\left(\frac{X_\mu^T \boldsymbol{w}}{\sqrt{d}} + \frac{\varepsilon_t \boldsymbol{w}^T \boldsymbol{\Sigma_\delta} \boldsymbol{w}}{\sqrt{d}\sqrt{\boldsymbol{w}^T \boldsymbol{w}}}\right)\right) \\
&= \sum_{\mu=1}^{N} y_\mu \log\left(1 + \exp\left(-\frac{X_\mu^T \boldsymbol{w}}{\sqrt{d}} + \frac{\varepsilon_t \boldsymbol{w}^T \boldsymbol{\Sigma_\delta} \boldsymbol{w}}{\sqrt{d}\sqrt{\boldsymbol{w}^T \boldsymbol{w}}}\right)\right) + (1 - y_\mu) \log\left(1 + \exp\left(\frac{X_\mu^T \boldsymbol{w}}{\sqrt{d}} + \frac{\varepsilon_t \boldsymbol{w}^T \boldsymbol{\Sigma_\delta} \boldsymbol{w}}{\sqrt{d}\sqrt{\boldsymbol{w}^T \boldsymbol{w}}}\right)\right) \\
&= \sum_{\mu=1}^{N} -y_\mu \log\left(\frac{\exp\left(\frac{X_\mu^T \boldsymbol{w}}{\sqrt{d}} - \frac{\varepsilon_t \boldsymbol{w}^T \boldsymbol{\Sigma_\delta} \boldsymbol{w}}{\sqrt{d}\sqrt{\boldsymbol{w}^T \boldsymbol{w}}}\right)}{1 + \exp\left(\frac{X_\mu^T \boldsymbol{w}}{\sqrt{d}} - \frac{\varepsilon_t \boldsymbol{w}^T \boldsymbol{\Sigma_\delta} \boldsymbol{w}}{\sqrt{d}\sqrt{\boldsymbol{w}^T \boldsymbol{w}}}\right)}\right) + (1 - y_\mu) \log\left(1 + \exp\left(\frac{X_\mu^T \boldsymbol{w}}{\sqrt{d}} + \frac{\varepsilon_t \boldsymbol{w}^T \boldsymbol{\Sigma_\delta} \boldsymbol{w}}{\sqrt{d}\sqrt{\boldsymbol{w}^T \boldsymbol{w}}}\right)\right)
\end{aligned} \tag{C2}$$

In the second step we are free to choose the labelling to be 0/1 as there is no more explicit dependence on the y_μ instance.

1. Computing the Loss

To accurately compute the loss we base our computation on the work of [4] who among other showed in 2012 how to accurately compute $\log(1 + \exp(x))$ in R. The idea is to choose the most suitable approximation depending on the argument of the function.

Finally, we can write the loss with 0/1 labels as

$$\mathcal{L} = \sum_{\mu=1}^{N} -y_\mu \frac{X_\mu^T \boldsymbol{w}}{\sqrt{d}} + y_\mu \frac{\varepsilon_t \boldsymbol{w}^T \boldsymbol{\Sigma_\delta} \boldsymbol{w}}{\sqrt{d}\sqrt{\boldsymbol{w}^T \boldsymbol{w}}} + (1 - y_\mu) \log\left(1 + \exp\left(\frac{X_\mu^T \boldsymbol{w}}{\sqrt{d}} + \frac{\varepsilon_t \boldsymbol{w}^T \boldsymbol{\Sigma_\delta} \boldsymbol{w}}{\sqrt{d}\sqrt{\boldsymbol{w}^T \boldsymbol{w}}}\right)\right) + y_\mu \log\left(1 + \exp\left(\frac{X_\mu^T \boldsymbol{w}}{\sqrt{d}} - \frac{\varepsilon_t \boldsymbol{w}^T \boldsymbol{\Sigma_\delta} \boldsymbol{w}}{\sqrt{d}\sqrt{\boldsymbol{w}^T \boldsymbol{w}}}\right)\right) \tag{C3}$$

[4] showed for the R language how to accurately compute $\log(1 + \exp(x))$, [5] inspire us to achieve faster computation by extending the case distinction by [4] by the ≤ 2 and using directly the log function.

$$\log(1 + \exp(x)) := \begin{cases} \exp(x) & x \leq -37 \\ \log 1\mathrm{p}(\exp(x)) & -37 < x \leq x_0 := -2 \\ \log(1. + \exp(x)) & x_0 < x \leq x_1 := 18 \\ x + \exp(-x) & x_1 < x \leq x_2 := 33.3 \\ x & x > x_2 \end{cases} \tag{C4}$$

2. Computing the Gradient

Call the derivative of the optimal attack H:

$$H = \partial_{\boldsymbol{w}} B = \partial_{\boldsymbol{w}} \frac{\varepsilon_t \boldsymbol{w}^T \boldsymbol{\Sigma}_{\boldsymbol{\delta}} \boldsymbol{w}}{\sqrt{d}\sqrt{\boldsymbol{w}^T \boldsymbol{w}}} = \frac{\varepsilon_t}{\sqrt{d}} \frac{\boldsymbol{w}}{(\boldsymbol{w}^T \boldsymbol{w})^{3/2}} \cdot \left[\left((\boldsymbol{\Sigma}_{\boldsymbol{\delta}} + \boldsymbol{\Sigma}_{\boldsymbol{\delta}}) \boldsymbol{w}^T \boldsymbol{w} \right) - \boldsymbol{w}^T \boldsymbol{\Sigma}_{\boldsymbol{\delta}} \boldsymbol{w} \right] \tag{C5}$$

We define the arguments of the sigmoid activation C and $\overline{C}$

$$C = \frac{X_\mu^T \boldsymbol{w}}{\sqrt{d}} + \frac{\varepsilon_t \boldsymbol{w}^T \boldsymbol{\Sigma}_{\boldsymbol{\delta}} \boldsymbol{w}}{\sqrt{d}\sqrt{\boldsymbol{w}^T \boldsymbol{w}}} \text{ and } \quad \overline{C} = \frac{X_\mu^T \boldsymbol{w}}{\sqrt{d}} - \frac{\varepsilon_t \boldsymbol{w}^T \boldsymbol{\Sigma}_{\boldsymbol{\delta}} \boldsymbol{w}}{\sqrt{d}\sqrt{\boldsymbol{w}^T \boldsymbol{w}}} \tag{C6}$$

Then we can write the derivative of the loss as

$$\begin{aligned} \partial_{\boldsymbol{w}} \mathcal{L} &= \sum_{\mu=1}^{N} -y_\mu \frac{X_\mu}{\sqrt{d}} + y_\mu H + \frac{(1-y_\mu)}{1+\exp\left(-C\right)} \left(\frac{X_\mu}{\sqrt{d}} + H \right) + \frac{y_\mu}{1+\exp\left(-\overline{C}\right)} \left(\frac{X_\mu}{\sqrt{d}} - H \right) \\ &= \sum_{\mu=1}^{N} \left[H \cdot \left(\frac{(1-y_\mu)}{1+\exp\left(-C\right)} + \frac{y_\mu \exp\left(-\overline{C}\right)}{1+\exp\left(-\overline{C}\right)} \right) + \frac{X_\mu}{\sqrt{d}} \cdot \left(\frac{(1-y_\mu)}{1+\exp\left(-C\right)} - \frac{y_\mu \exp\left(-\overline{C}\right)}{1+\exp\left(-\overline{C}\right)} \right) \right] \end{aligned} \tag{C7}$$

With this it is easy to see what the loss per sample is when factoring out the data X and the derivative of the optimal attack H. We see that their respective contributions are similar. One can avoid overflows by placing the exponential part of the sigmoid carefully depending on the argument.

3. Computing the Hessian

$$\begin{aligned} \partial_{\boldsymbol{w}^2} \mathcal{L} &= \sum_{\mu=1}^{N} y_\mu \partial_{\boldsymbol{w}} H + \frac{(1-y_\mu)}{1+\exp\left(-C\right)} \left(\partial_{\boldsymbol{w}} H \right) + \frac{y_\mu}{1+\exp\left(-\overline{C}\right)} \left(-\partial_{\boldsymbol{w}} H \right) \\ &\quad + (1-y_\mu) \left(\frac{X_\mu}{\sqrt{d}} + H \right) \partial_{\boldsymbol{w}} \frac{1}{1+\exp\left(-C\right)} + y_\mu \left(\frac{X_\mu}{\sqrt{d}} - H \right) \partial_{\boldsymbol{w}} \frac{1}{1+\exp\left(-\overline{C}\right)} \end{aligned} \tag{C8}$$

Let's compute the individual terms

K: Without loss of generality, we choose an easier optimal attack to compute the Hessian. We let $H = \frac{\varepsilon_t}{\sqrt{d}} \frac{\boldsymbol{w}}{\sqrt{\boldsymbol{w}^T \boldsymbol{w}}}$. Thus, the derivative is

$$\partial_{\boldsymbol{w}} H = \frac{\varepsilon_t}{\sqrt{d}\sqrt{\boldsymbol{w}^T \boldsymbol{w}}} \mathbb{I} - \frac{\varepsilon_t}{\sqrt{d}} \frac{\boldsymbol{w}\boldsymbol{w}^T}{(\boldsymbol{w}^T \boldsymbol{w})^{3/2}} \tag{C9}$$

$$\partial_{\boldsymbol{w}} C_\mu = \frac{X_\mu^T}{\sqrt{d}} + \varepsilon_t \frac{\boldsymbol{w}^T}{\sqrt{d}\sqrt{\boldsymbol{w}^T \boldsymbol{w}}} \quad \partial_{\boldsymbol{w}} \overline{C}_\mu = \frac{X_\mu^T}{\sqrt{d}} - \varepsilon_t \frac{\boldsymbol{w}^T}{\sqrt{d}\sqrt{\boldsymbol{w}^T \boldsymbol{w}}} \tag{C10}$$

$$\partial_{\boldsymbol{w}} \frac{1}{1+\exp(-C)} = \frac{\partial_{\boldsymbol{w}} C}{2\cosh C + 2} \tag{C11}$$

$$\begin{aligned} \partial_{\boldsymbol{w}^2} \mathcal{L} &= \sum_{\mu=1}^{N} \left[\frac{(1-y_\mu)}{1+\exp\left(-C_\mu\right)} + \frac{y_\mu}{1+\exp\left(\overline{C}_\mu\right)} \right] \left(\partial_{\boldsymbol{w}} H \right) \\ &\quad + \frac{(1-y_\mu)}{2\cosh C_\mu + 2} \left(\frac{X_\mu}{\sqrt{d}} + H \right) \left(\frac{X_\mu}{\sqrt{d}} + H \right)^T + \frac{y_\mu}{2\cosh \overline{C}_\mu + 2} \left(\frac{X_\mu}{\sqrt{d}} - H \right) \left(\frac{X_\mu}{\sqrt{d}} - H \right)^T \end{aligned} \tag{C12}$$

K: Double check implementation, it looks like that for some datasets, stuff breaks down!

K: Show that it fulfills a good precision by doing something similar as in the blog post... Either by reporting values or finding some nice plots...