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! Does it actually matter?I would expect so, it looks as if ERM finds different minima depending on it, double check though.. K: If you look at Brunos work, theorem 1.0 you can see that the regularization they choose has lambda/2 explicitly. I don't see however how this is translated into the replica computation.

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] 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`.

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

[4] 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.

[5] 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 [2]. 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. (A25), 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 [2]. 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.

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

explicitly we have that the term in P_g is

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

K: Why is this the correct beta scaling and why is there a 1/sqrt(2pi) ? K: It is somehow associated to the zero temperature scaling of the partition function, look at bruno A.31 the last part it is equal to:

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

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

$$\begin{aligned}1 &\propto \int\prod_{1\le a\le b\le r}\mathrm{d}B^{ab}\,\delta\left(dB^{ab} - \boldsymbol{w}^a\boldsymbol{\Sigma_\delta}\boldsymbol{w}^b\right)\int\prod_{1\le a\le b\le r}\mathrm{d}L^{ab}\,\delta\left(dL^{ab} - \boldsymbol{w}^a\cdot\boldsymbol{w}^a\right)\\ &= \int\prod_{1\le a\le b\le r}\frac{\mathrm{d}B^{ab}\,\mathrm{d}\hat{B}^{ab}}{2\pi}e^{-i\hat{B}^{ab}\left(dB^{ab} - \boldsymbol{w}^a\boldsymbol{\Sigma_\delta}\boldsymbol{w}^a\right)}\int\prod_{1\le a\le b\le r}\frac{\mathrm{d}L^{ab}\,\mathrm{d}\hat{L}^{ab}}{2\pi}e^{-i\hat{L}^{ab}\left(dL^{ab} - \boldsymbol{w}^a\cdot\boldsymbol{w}^a\right)}\end{aligned} \tag{A11}$$

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

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

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

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

we will refer to the elements in the first line of eq. (A13) as the trace term. Note in equation A12 we factored out d. K: Comment a bit more: We factor d out such that we can later evaluate the partition function in the thermodynamic limit using Laplace's method. Or is it the Saddle-point method here? We also have defined the prior part of the free energy Ψ_w to be

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

and the channel part of the free energy Ψ_y as

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

where we have used 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

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

Before we take the replica zero limit, let's check that our Ansatz above is well-defined and does not have a an order one term in $\Phi^{(r)}$ that diverges. For this, we need to ensure that $\lim_{r\to 0^+} \Phi^{(r)} = 0$. The trace terms depends on r except for $\rho\hat{\rho}$, $\lim_{r\to 0^+} \Psi_y^{(r)} = \lim_{r\to 0^+} \Psi_w^{(r)} = 0$ hold. Note that $\lim_{r\to 0^+} \log A^r = 0$ but $\lim_{r\to 0^+} \partial_r \log A^r = \log A$, and thus all we need to check is the prior part of the free energy in the zero replica limit.

$$\lim_{r\to 0^+} \Phi^{(r)} = -\rho\hat{\rho} \tag{A18}$$

For this limit to be zero, we must fix $\hat{\rho} = 0$ and note that ρ is a constant we fixed earlier.

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 :) K: I put them in to be explicit, and I'll add computation to show they cancel as the channel has no dependence on them.

Plugging in the Ansatz, the trace term becomes

$$-\rho\hat{\rho} - rm\hat{m} - \frac{r(r-1)}{2} q\hat{q} + \frac{r}{2} Q\hat{Q} + \frac{r}{2} A\hat{A} + \frac{r}{2} N\hat{N} - \frac{r(r-1)}{2} a\hat{a} - \frac{r(r-1)}{2} n\hat{n} \tag{A19}$$

Now we take the limit $r \to 0$ after dividing the trace term T (which is no longer an actual trace as we introduced overlaps beyond the traditional replica matrix ansatz) by r

$$T = \frac{1}{2}(q\hat{q} + Q\hat{Q}) + \frac{1}{2}(a\hat{a} + A\hat{A}) + \frac{1}{2}(n\hat{n} + N\hat{N}) - m\hat{m} \tag{A20}$$

a. *Prior Replica Zero Limit*

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

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

to perform in the following the $r \to 0^+$ limit we can change a bit the integral by factoring out all the terms.

In what follows, define $V = Q - q$ and $\hat{V} = \hat{Q} + \hat{q}$, $C = A - a$ and $\hat{C} = \hat{A} + \hat{a}$ and $M = N - n$ and $\hat{M} = \hat{N} + \hat{n}$ and note the identity

$$\begin{aligned}
&\sum_{1\le a<b\le r} \left(\hat{q}\boldsymbol{w}^{a\top}\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{w}^{b} + \hat{a}\boldsymbol{w}^{a\top}\boldsymbol{\Sigma}_{\boldsymbol{\delta}}\boldsymbol{w}^{b} + \hat{n}\boldsymbol{w}^{a}\cdot\boldsymbol{w}^{b}\right) - \frac{1}{2}\sum_{a=1}^{r}\left(\hat{Q}\boldsymbol{w}^{a\top}\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{w}^{a} + \hat{A}\boldsymbol{w}^{a\top}\boldsymbol{\Sigma}_{\boldsymbol{\delta}}\boldsymbol{w}^{a} + \hat{N}\boldsymbol{w}^{a}\cdot\boldsymbol{w}^{a}\right)\\
=&\sum_{1\le a<b\le r} \left(\hat{q}\boldsymbol{w}^{a\top}\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{w}^{b} + \hat{a}\boldsymbol{w}^{a\top}\boldsymbol{\Sigma}_{\boldsymbol{\delta}}\boldsymbol{w}^{b} + \hat{n}\boldsymbol{w}^{a}\cdot\boldsymbol{w}^{b}\right) - \frac{1}{2}\sum_{a=1}^{r}\left(\hat{V}\boldsymbol{w}^{a\top}\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{w}^{a} + \hat{C}\boldsymbol{w}^{a\top}\boldsymbol{\Sigma}_{\boldsymbol{\delta}}\boldsymbol{w}^{a} + \hat{M}\boldsymbol{w}^{a}\cdot\boldsymbol{w}^{a}\right)\\
&+\frac{1}{2}\sum_{a=1}^{r}\left(\hat{q}\boldsymbol{w}^{a\top}\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{w}^{a} + \hat{a}\boldsymbol{w}^{a\top}\boldsymbol{\Sigma}_{\boldsymbol{\delta}}\boldsymbol{w}^{a} + \hat{n}\boldsymbol{w}^{a}\cdot\boldsymbol{w}^{a}\right)\\
=&\frac{1}{2}\sum_{1\le a,b\le r} \left(\hat{q}\boldsymbol{w}^{a\top}\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{w}^{b} + \hat{a}\boldsymbol{w}^{a\top}\boldsymbol{\Sigma}_{\boldsymbol{\delta}}\boldsymbol{w}^{b} + \hat{n}\boldsymbol{w}^{a}\cdot\boldsymbol{w}^{b}\right) - \frac{1}{2}\sum_{a=1}^{r}\left(\hat{V}\boldsymbol{w}^{a\top}\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{w}^{a} + \hat{C}\boldsymbol{w}^{a\top}\boldsymbol{\Sigma}_{\boldsymbol{\delta}}\boldsymbol{w}^{a} + \hat{M}\boldsymbol{w}^{a}\cdot\boldsymbol{w}^{a}\right)
\end{aligned} \tag{A22}$$

To perform this simplification we will use the multidimensional Hubbard-Stratonovic identity which reads

$$e^{\frac{1}{2}\sum_{a,b=1}^{r}\boldsymbol{w}^{a\top}[\hat{q}\boldsymbol{\Sigma}_{\boldsymbol{x}}+\hat{a}\boldsymbol{\Sigma}_{\boldsymbol{\delta}}+\hat{n}\boldsymbol{I}]\boldsymbol{w}^{b}} = \mathbb{E}_{\boldsymbol{\xi}}\left[e^{\boldsymbol{\xi}^{\top}\sqrt{\hat{q}\boldsymbol{\Sigma}_{\boldsymbol{x}}+\hat{a}\boldsymbol{\Sigma}_{\boldsymbol{\delta}}+\hat{n}\boldsymbol{I}}\sum_{a=1}^{r}\boldsymbol{w}^{a}}\right] \tag{A23}$$

where $\boldsymbol{\xi} \sim \mathcal{N}(\boldsymbol{0}, 1_d)$.

Thus by calling the part inside the log in eq. (A21) with the letter $\mathcal{A}$ we have that (putting in $\hat{\rho} = 0$)

$$\begin{aligned}
\mathcal{A} &= \mathbb{E}_{\boldsymbol{\theta}_0}\int_{\mathbb{R}^{d\times r}}\prod_{a=1}^{r}P_w\left(\mathrm{d}\boldsymbol{w}^{a}\right)e^{-\hat{m}\sum_{a=1}^{r}\boldsymbol{\theta}_0^{\top}\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{w}^{a}+\frac{1}{2}\sum_{1\le a,b\le r}\left(\hat{q}\boldsymbol{w}^{a\top}\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{w}^{b}+\hat{a}\boldsymbol{w}^{a\top}\boldsymbol{\Sigma}_{\boldsymbol{\delta}}\boldsymbol{w}^{b}+\hat{n}\boldsymbol{w}^{a}\cdot\boldsymbol{w}^{b}\right)-\frac{1}{2}\sum_{a=1}^{r}\left(\hat{V}\boldsymbol{w}^{a\top}\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{w}^{a}+\hat{C}\boldsymbol{w}^{a\top}\boldsymbol{\Sigma}_{\boldsymbol{\delta}}\boldsymbol{w}^{a}+\hat{M}\boldsymbol{w}^{a}\cdot\boldsymbol{w}^{a}\right)}\\
&= \mathbb{E}_{\boldsymbol{\theta}_0}\int_{\mathbb{R}^{d\times r}}\prod_{a=1}^{r}P_w\left(\mathrm{d}\boldsymbol{w}^{a}\right)e^{-\sum_{a=1}^{r}\left(\frac{\hat{V}}{2}\boldsymbol{w}^{a}\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{w}^{a}+\frac{\hat{C}}{2}\boldsymbol{w}^{a}\boldsymbol{\Sigma}_{\boldsymbol{\delta}}\boldsymbol{w}^{a}+\frac{\hat{M}}{2}\boldsymbol{w}^{a}\boldsymbol{w}^{a}+\hat{m}\boldsymbol{\theta}_0\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{w}^{a}\right)+\frac{1}{2}\sum_{1\le a,b\le r}\left(\hat{q}\boldsymbol{w}^{a\top}\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{w}^{b}+\hat{a}\boldsymbol{w}^{a\top}\boldsymbol{\Sigma}_{\boldsymbol{\delta}}\boldsymbol{w}^{b}+\hat{n}\boldsymbol{w}^{a}\cdot\boldsymbol{w}^{b}\right)}\\
&= \mathbb{E}_{\boldsymbol{\theta}_0}\int_{\mathbb{R}^{d\times r}}\prod_{a=1}^{r}P_w\left(\mathrm{d}\boldsymbol{w}^{a}\right)\mathbb{E}_{\boldsymbol{\xi}}\left[e^{-\frac{\hat{V}}{2}\boldsymbol{w}^{a}\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{w}^{a}-\frac{\hat{C}}{2}\boldsymbol{w}^{a}\boldsymbol{\Sigma}_{\boldsymbol{\delta}}\boldsymbol{w}^{a}-\frac{\hat{M}}{2}\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}}+\hat{a}\boldsymbol{\Sigma}_{\boldsymbol{\delta}}+\hat{n}\boldsymbol{I}}\boldsymbol{\xi}\right)}\right]\\
&= \mathbb{E}_{\boldsymbol{\xi}\boldsymbol{\theta}_0}\left[\left[\int_{\mathbb{R}^{d}}P_w\left(\mathrm{d}\boldsymbol{w}\right)e^{-\frac{\hat{V}}{2}\boldsymbol{w}\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{w}-\frac{\hat{C}}{2}\boldsymbol{w}\boldsymbol{\Sigma}_{\boldsymbol{\delta}}\boldsymbol{w}-\frac{\hat{M}}{2}\boldsymbol{w}\boldsymbol{w}-\boldsymbol{w}\left(\hat{m}\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{\theta}_0-\sqrt{\hat{q}\boldsymbol{\Sigma}_{\boldsymbol{x}}+\hat{a}\boldsymbol{\Sigma}_{\boldsymbol{\delta}}+\hat{n}\boldsymbol{I}}\boldsymbol{\xi}\right)}\right]^{r}\right]
\end{aligned} \tag{A24}$$

Then we can take the derivative and limit and obtain

$$\Psi_w = \lim_{r\to 0^+}\partial_r\Psi_w^{(r)} = \frac{1}{d}\mathbb{E}_{\boldsymbol{\xi},\boldsymbol{\theta}_0}\left[\log\int_{\mathbb{R}^{d}}P_w\left(\mathrm{d}\boldsymbol{w}\right)e^{-\frac{\hat{V}}{2}\boldsymbol{w}\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{w}-\frac{\hat{C}}{2}\boldsymbol{w}\boldsymbol{\Sigma}_{\boldsymbol{\delta}}\boldsymbol{w}-\frac{\hat{M}}{2}\boldsymbol{w}\boldsymbol{w}-\boldsymbol{w}\left(\hat{m}\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{\theta}_0-\sqrt{\hat{q}\boldsymbol{\Sigma}_{\boldsymbol{x}}+\hat{a}\boldsymbol{\Sigma}_{\boldsymbol{\delta}}+\hat{n}\boldsymbol{I}}\boldsymbol{\xi}\right)}\right] \tag{A25}$$

where we still need to take the limit $d \to \infty$. K: Do we actually take the limit though?

b. *Channel Replica Zero Limit*

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

$$\Psi_y^{(r)} = \log\left[\int_{\mathbb{R}}\mathrm{d}y\int_{\mathbb{R}}\mathrm{d}\nu\,P_0(y\mid\nu)\int\prod_{a=1}^{r}\mathrm{d}\lambda^{a}\,P_g\left(y\mid\lambda^{a},A,N,\varepsilon_t\right)\mathcal{N}\left(\nu,\lambda^{a};\boldsymbol{0},\Sigma^{ab}\right)\right] \tag{A26}$$

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

$$\Sigma = \begin{pmatrix} \rho & m & m & \dots & m\\ m & Q & q & \dots & q\\ m & q & Q & \dots & q\\ \vdots & \vdots & \vdots & \ddots & \vdots\\ m & q & q & \dots & Q \end{pmatrix} \tag{A27}$$

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

$$\begin{aligned}
\left(\Sigma^{-1}\right)_{00} &\equiv \tilde{\rho} = \frac{Q+(r-1)q}{\rho(Q+(r-1)q)-rm^2}\,, & \left(\Sigma^{-1}\right)_{0a} &\equiv \tilde{m} = \frac{m}{rm^2-\rho(Q+(r-1)q)}\,,\\
\left(\Sigma^{-1}\right)_{aa} &\equiv \tilde{Q} = \frac{\rho(Q+(r-2)q)-(r-1)m^2}{(Q-q)\left(\rho(Q+(r-1)q)-rm^2\right)}\,, & \left(\Sigma^{-1}\right)_{ab} &\equiv \tilde{q} = \frac{m^2-\rho q}{(Q-q)\left(\rho(Q+(r-1)q)-rm^2\right)}
\end{aligned} \tag{A28}$$

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

$$\begin{aligned}
&\lambda_1 = Q-q\,, \qquad \lambda_2 = \frac{1}{2}\Big(-Q-q(r-1)-\rho-\tilde{\Delta}\Big)\,, \qquad \lambda_3 = \frac{1}{2}\Big(-Q-q(r-1)-\rho+\tilde{\Delta}\Big)\,,\\
&d_1 = r-1\,, \qquad d_2 = 1\,, \qquad\qquad\qquad d_3 = 1\,,
\end{aligned} \tag{A29}$$

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

$$\begin{aligned}
\det\left(2\pi\Sigma\right) &= (2\pi)^{r+1}(Q-q)^{r-1}\frac{1}{4}(-Q-q(r-1)-\rho-\tilde{\Delta})(-Q-q(r-1)-\rho+\tilde{\Delta})\\
&= (2\pi)^{r+1}(Q-q)^{r-1}(\rho(Q+(r-1)q)-rm^2)
\end{aligned} \tag{A30}$$

Thus we have that

$$\begin{aligned}
\mathcal{B} &= \int_{\mathbb{R}}\mathrm{d}y\int_{\mathbb{R}}\mathrm{d}\nu\,P_0(y\mid\nu)e^{-\frac{1}{2}\tilde{\rho}\nu^2}\int\prod_{a=1}^{r}\mathrm{d}\lambda^a\,P_g\left(y\mid\lambda^a,A,N,\varepsilon_t\right)e^{-\tilde{m}\nu\sum_{a=1}^{r}\lambda^a-\frac{1}{2}\tilde{Q}\sum_{a=1}^{r}(\lambda^a)^2-\frac{1}{2}\tilde{q}\sum_{1\le a,b\le r,a\neq b}\lambda^a\lambda^b-\frac{1}{2}\log\det(2\pi\Sigma)}\\
&= \mathbb{E}_\xi\int_{\mathbb{R}}\mathrm{d}y\,e^{-\frac{1}{2}\log\det(2\pi\Sigma)}\int_{\mathbb{R}}\mathrm{d}\nu\,P_0(y\mid\nu)e^{-\frac{1}{2}\tilde{\rho}\nu^2}\left[\int\mathrm{d}\lambda\,P_g\left(y\mid\lambda,A,N,\varepsilon_t\right)e^{-\frac{\tilde{Q}-\tilde{q}}{2}\lambda^2+(\sqrt{-\tilde{q}}\xi-\tilde{m}\nu)\lambda}\right]^r
\end{aligned} \tag{A31}$$

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

$$\begin{aligned}
\Psi_y = \lim_{r\to0^+}\partial_r\Psi_y^{(r)} &= \mathbb{E}_\xi\left[\int_{\mathbb{R}}\mathrm{d}y\int\frac{\mathrm{d}\nu}{\sqrt{2\pi\rho}}P_0(y\mid\nu)e^{-\frac{1}{2\rho}\nu^2}\log\left[\int\frac{\mathrm{d}\lambda}{\sqrt{2\pi}}P_y(y\mid\lambda,A,N,\varepsilon_t)e^{-\frac{1}{2}\frac{\lambda^2}{V}+\left(\frac{\sqrt{q-m^2/\rho}}{V}\xi+\frac{m/\rho}{V}\nu\right)\lambda}\right]\right]\\
&\quad-\frac{1}{2}\log V-\frac{1}{2}\frac{q}{V}
\end{aligned} \tag{A32}$$

K: fun little exercise the term outside with a log gives the normalisation constant and the other term K: what does it do?

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

$$\mathcal{Z}_0(y,\omega,V) = \int\frac{\mathrm{d}x}{\sqrt{2\pi V}}e^{-\frac{1}{2V}(x-\omega)^2}P_0(y\mid x)\,, \quad \mathcal{Z}_y(y,\omega,V,a,n) = \int\frac{\mathrm{d}x}{\sqrt{2\pi V}}e^{-\frac{1}{2V}(x-\omega)^2}P_y(y\mid x,a,n,\varepsilon_t) \tag{A33}$$

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. K: Boy I am naive sometimes

$$\mathbb{E}_\xi\left[\int_{\mathbb{R}}\mathrm{d}y\,\mathcal{Z}_0\left(y,\frac{m}{\sqrt{q}}\xi,\rho-\frac{m^2}{q}\right)\log\mathcal{Z}_y(y,\sqrt{q}\xi,V,A,N)\right] \tag{A34}$$

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

$$P_w(\mathrm{d}\boldsymbol{w}) = \frac{1}{(2\pi)^{d/2}}\exp\left(-\frac{\beta\lambda}{2}\boldsymbol{w}\boldsymbol{\Sigma}_{\boldsymbol{w}}\boldsymbol{w}\right)\mathrm{d}\boldsymbol{w} \tag{A35}$$

We want to calculate the term inside the log in eq. (A25)

$$\begin{aligned}&\int_{\mathbb{R}^d} P_w(\mathrm{d}\boldsymbol{w}) e^{-\frac{\hat{V}}{2}\boldsymbol{w}\boldsymbol{\Sigma_x}\boldsymbol{w}-\frac{\hat{C}}{2}\boldsymbol{w}\boldsymbol{\Sigma_\delta}\boldsymbol{w}-\frac{\hat{M}}{2}\boldsymbol{w}\boldsymbol{w}-\boldsymbol{w}\left(\hat{m}\boldsymbol{\Sigma_x}\boldsymbol{\theta}_0-\sqrt{\hat{q}\boldsymbol{\Sigma_x}+\hat{a}\boldsymbol{\Sigma_\delta}+\hat{n}\boldsymbol{I}}\boldsymbol{\xi}\right)}\\ &= \frac{\exp\left(\frac{1}{2}\left(-\hat{m}\boldsymbol{\Sigma_x}^\top\boldsymbol{\theta}_0+\sqrt{\hat{q}\boldsymbol{\Sigma_x}+\hat{a}\boldsymbol{\Sigma_\delta}+\hat{n}\boldsymbol{I}}\boldsymbol{\xi}\right)^\top\boldsymbol{\Lambda}^{-1}\left(-\hat{m}\boldsymbol{\Sigma_x}^\top\boldsymbol{\theta}_0+\sqrt{\hat{q}\boldsymbol{\Sigma_x}+\hat{a}\boldsymbol{\Sigma_\delta}+\hat{n}\boldsymbol{I}}\boldsymbol{\xi}\right)^\top\right)}{\sqrt{\det\boldsymbol{\Lambda}}}\end{aligned} \tag{A36}$$

where we defined $\boldsymbol{\Lambda}=\beta\lambda\boldsymbol{\Sigma_w}+\hat{V}\boldsymbol{\Sigma_x}+\hat{C}\boldsymbol{\Sigma_\delta}+\hat{M}\boldsymbol{I}$. Now the prior term becomes after taking the log and using the identity $\log\det=\operatorname{tr}\log$

$$\begin{aligned}\Psi_w &= \frac{1}{d}\mathbb{E}_{\boldsymbol{\xi},\boldsymbol{\theta}_0}\left[\frac{1}{2}\left(-\hat{m}\boldsymbol{\Sigma_x}^\top\boldsymbol{\theta}_0+\sqrt{\hat{q}\boldsymbol{\Sigma_x}+\hat{a}\boldsymbol{\Sigma_\delta}+\hat{n}\boldsymbol{I}}\boldsymbol{\xi}\right)^\top\boldsymbol{\Lambda}^{-1}\left(-\hat{m}\boldsymbol{\Sigma_x}^\top\boldsymbol{\theta}_0+\sqrt{\hat{q}\boldsymbol{\Sigma_x}+\hat{a}\boldsymbol{\Sigma_\delta}+\hat{n}\boldsymbol{I}}\boldsymbol{\xi}\right)^\top\right]-\frac{1}{2d}\operatorname{tr}\log\boldsymbol{\Lambda}\\ &= \frac{1}{2d}\operatorname{tr}\left[\left(\hat{m}^2\boldsymbol{\Sigma_x}^\top\boldsymbol{\theta}_0\boldsymbol{\theta}_0^\top\boldsymbol{\Sigma_x}+\hat{q}\boldsymbol{\Sigma_x}+\hat{a}\boldsymbol{\Sigma_\delta}+\hat{n}\boldsymbol{I}\right)\boldsymbol{\Lambda}^{-1}\right]-\frac{1}{2d}\operatorname{tr}\log\boldsymbol{\Lambda}\end{aligned} \tag{A37}$$

The factor $\frac{1}{d}$ comes from the required scaling on d for the free entropy and the expectation from our replica zero limit of the prior term.

K: An interesting thing to do would be to compute the prior free entropy term in terms of the partifions functions as in aubin equation 99. How would then the new overlaps come into play? Is it actually doable in the covariate case?

3. Saddle-point equations

Recall our free-entropy where we extremize over the overlaps

$$\Phi(\alpha)=\operatorname{extr}\left[T+\alpha\Psi_y+\Psi_w\right] \tag{A38}$$

From this we get our set of fixed point equations

$$\begin{aligned}\hat{Q}&=-2\alpha\partial_Q\Psi_y, & Q&=-2\partial_{\hat{Q}}\Psi_w\\ \hat{q}&=-2\alpha\partial_q\Psi_y, & q&=-2\partial_{\hat{q}}\Psi_w,\\ \hat{N}&=-2\alpha\partial_N\Psi_y, & N&=-2\partial_{\hat{N}}\Psi_w\\ \hat{n}&=-2\alpha\partial_n\Psi_y, & n&=-2\partial_{\hat{n}}\Psi_w,\\ \hat{A}&=-2\alpha\partial_A\Psi_y, & A&=-2\partial_{\hat{A}}\Psi_w\\ \hat{a}&=-2\alpha\partial_a\Psi_y, & a&=-2\partial_{\hat{a}}\Psi_w,\\ \hat{m}&=\alpha\partial_m\Psi_y, & m&=\partial_{\hat{m}}\Psi_w.\end{aligned} \tag{A39}$$

K: Describe next how they simplify, because there are some dependencies cancelling as Ψ_y does not depend on all the overlaps. We introduced all the overlaps, such that we can do the hubbard transformation for the prior term

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

a. *The Channel Saddle-Point Equations*

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 [2]. 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 $\mathcal{Z}$

$$\begin{aligned}\partial_A\mathcal{Z}_y(y,\omega,V) &= \frac{\partial}{\partial A}\sqrt{\beta}\int\frac{\mathrm{d}x}{\sqrt{2\pi V}}e^{-\frac{1}{2V}(x-\omega)^2}\frac{e^{-\beta g\left(yx-\varepsilon_t A/\sqrt{N}\right)}}{\sqrt{2\pi}}\\ &= \sqrt{\beta}\int\frac{\mathrm{d}x}{\sqrt{2\pi V}}e^{-\frac{1}{2V}(x-\omega)^2}e^{-\beta g\left(yx-\varepsilon_t A/\sqrt{N}\right)}\left(\beta\frac{\varepsilon_t}{\sqrt{N}}g'\left(yx-\varepsilon_t\frac{A}{\sqrt{N}}\right)\right)\end{aligned} \tag{A40}$$

We also have for the derivative with respect to N is

$$\begin{aligned}\partial_N \mathcal{Z}_y(y,\omega,V) &= \frac{\partial}{\partial N}\sqrt{\beta}\int \frac{\mathrm{d}x}{\sqrt{2\pi V}} e^{-\frac{1}{2V}(x-\omega)^2} \frac{e^{-\beta g\left(yx-\varepsilon_t A/\sqrt{N}\right)}}{\sqrt{2\pi}} \\ &= \sqrt{\beta}\int \frac{\mathrm{d}x}{\sqrt{2\pi V}} e^{-\frac{1}{2V}(x-\omega)^2} e^{-\beta g\left(yx-\varepsilon_t A/\sqrt{N}\right)} \left(-\beta \frac{\varepsilon_t A}{2N^{3/2}} g'\left(yx-\varepsilon_t \frac{A}{\sqrt{N}}\right)\right)\end{aligned} \tag{A41}$$

The factor $\sqrt{\beta}$ in front of everything is taken care with the temperature scalings of the order parameter V chosen in eq. (A64).

Another way to obtain the derivative w.r.t a is this:

$$\begin{aligned}\partial_a \Psi_y &= \partial_a \mathbb{E}_{y,\xi}\left[\mathcal{Z}_0\left(y, \frac{m}{\sqrt{q}}\xi, \rho - \frac{m^2}{q}\right)\log \mathcal{Z}_y(y,\sqrt{q}\xi,V,a,n)\right] \\ &= \mathbb{E}_{y,\xi}\left[\mathcal{Z}_0\left(y, \frac{m}{\sqrt{q}}\xi, \rho - \frac{m^2}{q}\right)\frac{\partial_a \mathcal{Z}_y}{\mathcal{Z}_y}\right]\end{aligned} \tag{A42}$$

So the question is, what is the derivative of the partition function w.r.t. a. To answer this question, let us introduce the following alternative representation of the partition function in the zero temperature limit.

$$\begin{aligned}\mathcal{Z}_y(y,\omega,V,a,n) &= \int \frac{\mathrm{d}x}{\sqrt{2\pi V}} e^{-\frac{\beta}{2V}(x-\omega)^2} P_y(y \mid x,a,n,\varepsilon_t) \\ &= \int \frac{\mathrm{d}x}{\sqrt{2\pi V}} e^{-\frac{\beta}{2V}(x-\omega)^2} \sqrt{\beta} e^{(-\beta g(yx-\frac{\varepsilon_t}{\sqrt{d}}\frac{a}{\sqrt{n}}))} \\ &\underset{\beta\to\infty}{=} e^{-\beta \mathcal{M}_{Vg(y,\cdot)}(\omega)}\end{aligned} \tag{A43}$$

K: Note I didn't take the 1/2pi with from the definition earlier K: Still the beta scaling confuses me...

Where we introduced the Moreau-envelope

$$\mathcal{M}_{Vg(y,\cdot)}(\omega,a,n,\varepsilon_t) = \inf_{x\in\mathbb{R}}\left[\frac{(x-\omega)^2}{2V} + g(y,x,a,n,\varepsilon_t)\right] \tag{A44}$$

It is possible to describe the channel overlap equations in this formalism. As in [2] we get

$$\begin{aligned}\hat{V} &= -\alpha \mathbb{E}_\xi\left[\int_{\mathbb{R}} \mathrm{d}y \mathcal{Z}_0 \partial_\omega f_g\right] \\ \hat{q} &= \alpha \mathbb{E}_\xi\left[\int_{\mathbb{R}} \mathrm{d}y \mathcal{Z}_0 f_g^2\right] \\ \hat{m} &= \alpha \mathbb{E}_\xi\left[\int_{\mathbb{R}} \mathrm{d}y \partial_\omega \mathcal{Z}_0 f_g\right]\end{aligned} \tag{A45}$$

where

$$f_g(y,\omega,V,a,n,\varepsilon_t) = -\partial_\omega \mathcal{M}_{Vg(y,\cdot)}(\omega,a,n,\varepsilon_t) \tag{A46}$$

One can also obtain this derivative using the proximal operator

$$\mathcal{P}_{Vg(y,\cdot)}(\omega,a,n,\varepsilon_t) = \arg\min_{x\in\mathbb{R}}\left[\frac{(x-\omega)^2}{2V} + g(y,x,a,n,\varepsilon_t)\right] \tag{A47}$$

and the envelope theorem

$$\partial_\omega \mathcal{M}_{Vg(y,\cdot)}(\omega,a,n,\varepsilon_t) = V^{-1}\left(\omega - \mathcal{P}_{Vg(y,\cdot)}(\omega,a,n,\varepsilon_t)\right) \tag{A48}$$

For the new equations for both $\hat{a}$ and $\hat{n}$ we need new denoising functions than f_a and f_n.

$$\begin{aligned}f_a &= \partial_a \mathcal{M}_{Vg(y,\cdot)}(\omega,a,n,\varepsilon_t) \\ f_n &= \partial_n \mathcal{M}_{Vg(y,\cdot)}(\omega,a,n,\varepsilon_t)\end{aligned} \tag{A49}$$

How does one compute these derivatives? Let us write the Moreau-envelope exlicitly:

$$\mathcal{M}_{Vg(y,\cdot)}(\omega, a, n, \varepsilon_t) = \inf_{x\in\mathbb{R}} \left[\frac{(x-\omega)^2}{2V} + \log\left(1 + \exp\left(-yx + \varepsilon_t \frac{a}{\sqrt{n}}\right)\right)\right] \tag{A50}$$

In order to obtain the derivative we now define the function $\varepsilon(a,n) = \varepsilon_t \frac{a}{\sqrt{n}}$. We derive the derivative as a function of an arbitrary variable k. Then we have

$$\begin{aligned}
\partial_k \mathcal{M}_{Vg(y,\cdot)}(\omega,\varepsilon) &= \lim_{h\to 0} \frac{1}{h}\left[\mathcal{M}_{Vg(y,\cdot)}(\omega,\varepsilon+h) - \mathcal{M}_{Vg(y,\cdot)}(\omega,\varepsilon)\right] \\
&= \lim_{h\to 0} \frac{1}{h}\left[\log\left(1+e^{-yx+\varepsilon(k)+h}\right) - \log\left(1+e^{-yx+\varepsilon(k)+h}\right)\right] \\
&= \lim_{h\to 0} \frac{1}{h}\left[\log\left(1+e^{-yx+\varepsilon(k)}\right) + h\frac{\partial_k\varepsilon(k)}{1+e^{yx-\varepsilon(k)}} - \log\left(1+e^{-yx+\varepsilon(k)}\right)\right] \\
&= \frac{\partial_k\varepsilon(k)}{1+e^{yx-\varepsilon(k)}}
\end{aligned} \tag{A51}$$

where x is the value at the given $\varepsilon(k)$ that is given by the proximal operator. With this, we can write the new equations over a and n as

$$\begin{aligned}
\hat{a} &= \alpha \mathbb{E}_\xi \left[\int_{\mathbb{R}} \mathrm{d}y \mathcal{Z}_0 f_a\right] \\
\hat{n} &= \alpha \mathbb{E}_\xi \left[\int_{\mathbb{R}} \mathrm{d}y \mathcal{Z}_0 f_n\right]
\end{aligned} \tag{A52}$$

K: Should we get even more explicit, i.e. that somebody could look at an equation and type it in code right away without needing to compute anything?

b. *The Prior Saddle-Point Equations*

For the Prior saddle-point equations our starting point is

$$\Psi_w = \frac{1}{2d} \operatorname{tr}\left[\left(\hat{m}^2 \boldsymbol{\Sigma}_{\boldsymbol{x}}^\top \boldsymbol{\theta}_0 \boldsymbol{\theta}_0^\top \boldsymbol{\Sigma}_{\boldsymbol{x}} + \hat{q}\boldsymbol{\Sigma}_{\boldsymbol{x}} + \hat{a}\boldsymbol{\Sigma}_{\boldsymbol{\delta}} + \hat{n}\boldsymbol{I}\right)\boldsymbol{\Lambda}^{-1}\right] - \frac{1}{2d}\operatorname{tr}\log\boldsymbol{\Lambda} \tag{A53}$$

where $\boldsymbol{\Lambda} = \beta\lambda\boldsymbol{\Sigma}_{\boldsymbol{w}} + \hat{V}\boldsymbol{\Sigma}_{\boldsymbol{x}} + \hat{C}\boldsymbol{\Sigma}_{\boldsymbol{\delta}} + \hat{M}\boldsymbol{I}$ and we will use $\boldsymbol{H} = \hat{m}^2\boldsymbol{\Sigma}_{\boldsymbol{x}}^\top\boldsymbol{\theta}_0\boldsymbol{\theta}_0^\top\boldsymbol{\Sigma}_{\boldsymbol{x}} + \hat{q}\boldsymbol{\Sigma}_{\boldsymbol{x}} + \hat{a}\boldsymbol{\Sigma}_{\boldsymbol{\delta}} + \hat{n}\boldsymbol{I}$. K: Turns out the version where the new overlaps don't show up in H was right after all. we'll need to fix this... We want to compute a few derivatives of the term Ψ_w to obtain equations for the overlaps A, a, N, n.

We begin with the hat-variables $\hat{A}$ and $\hat{N}$

$$\partial_{\hat{A}}\Psi_w = \frac{1}{d}\operatorname{tr}\left[\boldsymbol{H}\boldsymbol{\Sigma}_{\boldsymbol{\delta}}\boldsymbol{\Lambda}^{-2}\right] - \frac{1}{d}\operatorname{tr}\left[\boldsymbol{\Sigma}_{\boldsymbol{\delta}}\boldsymbol{\Lambda}^{-1}\right] \tag{A54}$$

and

$$\partial_{\hat{N}}\Psi_w = \frac{1}{d}\operatorname{tr}\left[\boldsymbol{H}\boldsymbol{I}\boldsymbol{\Lambda}^{-2}\right] - \frac{1}{d}\operatorname{tr}\left[\boldsymbol{I}\boldsymbol{\Lambda}^{-1}\right] \tag{A55}$$

For $\hat{Q}$ it is similar

$$\partial_{\hat{Q}}\Psi_w = \frac{1}{d}\operatorname{tr}\left[\boldsymbol{H}\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{\Lambda}^{-2}\right] - \frac{1}{d}\operatorname{tr}\left[\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{\Lambda}^{-1}\right] \tag{A56}$$

Let's look at the equations for $\hat{q}, \hat{a}, \hat{n}$:

$$\partial_{\hat{q}}\Psi_w = -\frac{1}{d}\operatorname{tr}\left[\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{\Lambda}^{-1}\right] + \frac{1}{d}\operatorname{tr}\left[\boldsymbol{H}\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{\Lambda}^{-2}\right] + \frac{1}{d}\operatorname{tr}\left[\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{\Lambda}^{-1}\right] \tag{A57}$$

$$\partial_{\hat{a}}\Psi_w = -\frac{1}{d}\operatorname{tr}\left[\boldsymbol{\Sigma}_{\boldsymbol{\delta}}\boldsymbol{\Lambda}^{-1}\right] + \frac{1}{d}\operatorname{tr}\left[\boldsymbol{H}\boldsymbol{\Sigma}_{\boldsymbol{\delta}}\boldsymbol{\Lambda}^{-2}\right] + \frac{1}{d}\operatorname{tr}\left[\boldsymbol{\Sigma}_{\boldsymbol{\delta}}\boldsymbol{\Lambda}^{-1}\right] \tag{A58}$$

$$\partial_{\hat{n}}\Psi_w = -\frac{1}{d}\operatorname{tr}\left[\boldsymbol{I}\boldsymbol{\Lambda}^{-1}\right] + \frac{1}{d}\operatorname{tr}\left[\boldsymbol{H}\boldsymbol{I}\boldsymbol{\Lambda}^{-2}\right] + \frac{1}{d}\operatorname{tr}\left[\boldsymbol{I}\boldsymbol{\Lambda}^{-1}\right] \tag{A59}$$

Remember that $V = Q - q$, so the equation for $V = -2(\partial_Q - \partial_q)\Psi_w$ and hence it is given by

$$\partial_{\hat{V}}\Psi_w = \frac{1}{d}\operatorname{tr}\left[\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{\Lambda}^{-1}\right] \tag{A60}$$

The equation for $\hat{m}$ is given by

Note that for the numerical evaluation $\frac{1}{d}\operatorname{tr}$ is just the mean of the eigenspectrum.

K: Motivate why we iterate V instead of Q (yes, there is this AMP relation, but we don't do amp and even if, you could just get V back by Q-q) a K: Note that I don't believe that the log term is supposed to vanish in the zero temperature limit, it's derivatives simply cancel e.g. for Lambda. K: Maybe it is also easier to track the corresponding Vhat instead of Qhat, I'll look into the channel part again anyways... K: Something I wonder about the channel part is if we scale epsilon with sqrt(d) properly

$$\partial_{\hat{A}}\Psi_w = \frac{1}{2d}\operatorname{tr}\left[\left(\hat{m}^2\Phi^\top\boldsymbol{\theta}_0\boldsymbol{\theta}_0^\top\Phi + \hat{q}\Omega\right)\boldsymbol{\Sigma}_{\boldsymbol{\delta}}\boldsymbol{\Lambda}^{-2}\right] \tag{A61}$$

and

$$\partial_{\hat{N}}\Psi_w = \frac{1}{2d}\operatorname{tr}\left[\left(\hat{m}^2\Phi^\top\boldsymbol{\theta}_0\boldsymbol{\theta}_0^\top\Phi + \hat{q}\Omega\right)\boldsymbol{\Lambda}^{-2}\right] \tag{A62}$$

with the same definition of $\boldsymbol{\Lambda}$ as in the previous section.

$$\begin{aligned} \hat{A} &= -\alpha\partial_A\Psi_y \qquad A = -\partial_{\hat{A}}\Psi_w \\ \hat{N} &= -\alpha\partial_N\Psi_y \qquad N = -\partial_{\hat{N}}\Psi_w \end{aligned} \tag{A63}$$

4. Zero temperature limit

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

$$\begin{aligned} V &\to \beta^{-1}V \quad & q &\to q \quad & m &\to m \quad & A &\to \beta^2 A \quad & N &\to \beta^2 N \\ \hat{V} &\to \beta\hat{V} \quad & \hat{q} &\to \beta^2\hat{q} \quad & \hat{m} &\to \beta\hat{m} \quad & \hat{A} &\to \beta\hat{A} \quad & \hat{N} &\to \beta\hat{N} \end{aligned} \tag{A64}$$

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

$$\Psi_w = \lim_{\beta\to\infty}\frac{1}{\beta}\Psi_w = \frac{1}{2d}\operatorname{tr}\left[\left(\hat{m}^2\Phi^\top\boldsymbol{\theta}_0\boldsymbol{\theta}_0^\top\Phi + \hat{q}\Omega\right)\boldsymbol{\Lambda}^{-1}\right] \tag{A65}$$

and then the limit of the channel term becomes

$$\Psi_y = \lim_{\beta\to\infty}\frac{1}{\beta}\Psi_y = -\mathbb{E}_{\xi\sim\mathcal{N}(0,1)}\left[\int \mathrm{d}y\mathcal{Z}_0\left(y, \frac{m}{\sqrt{q}}\xi, \rho - \frac{m^2}{q}\right)\mathcal{M}_{Vg(y,\cdot;A,N,\varepsilon_t)}(\sqrt{q}\xi)\right] \tag{A66}$$

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. (A40) and (A41) 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

$$\begin{cases} \hat{V} = -\alpha \mathbb{E}_\xi \left[\int_{\mathbb{R}} \mathrm{d}y \, \mathcal{Z}_0(y, \sqrt{\eta}\xi, \rho - \eta) \partial_\omega f_g(y, \sqrt{q}\xi, V) \right] \\ \hat{q} = \alpha \mathbb{E}_\xi \left[\int_{\mathbb{R}} \mathrm{d}y \, \mathcal{Z}_0(y, \sqrt{\eta}\xi, \rho - \eta) f_g^2(y, \sqrt{q}\xi, V) \right] \\ \hat{m} = \alpha \mathbb{E}_\xi \left[\int_{\mathbb{R}} \mathrm{d}y \, \partial_\omega \mathcal{Z}_0(y, \sqrt{\eta}\xi, \rho - \eta) f_g(y, \sqrt{q}\xi, V) \right] \\ \hat{A} = -\alpha \frac{\varepsilon_t}{\sqrt{N}} \mathbb{E}_\xi \left[\int_{\mathbb{R}} \mathrm{d}y \, \mathcal{Z}_0(y, \sqrt{\eta}\xi, \rho - \eta) g'\left(y \mathcal{P}_{Vg}(...) - \varepsilon_t \frac{A}{N}\right) \right] \\ \hat{N} = \alpha \frac{\varepsilon_t}{2} \frac{A}{N^{3/2}} \mathbb{E}_\xi \left[\int_{\mathbb{R}} \mathrm{d}y \, \mathcal{Z}_0(y, \sqrt{\eta}\xi, \rho - \eta) g'\left(y \mathcal{P}_{Vg}(...) - \varepsilon_t \frac{A}{N}\right) \right] \end{cases} \tag{A67}$$

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

$$\begin{cases} V = \frac{1}{d} \operatorname{tr} \boldsymbol{\Lambda}^{-1} \boldsymbol{\Sigma}_{\boldsymbol{x}} \\ q = \frac{1}{d} \operatorname{tr} \left[\left(\hat{q} \boldsymbol{\Sigma}_{\boldsymbol{x}} + \hat{m}^2 \boldsymbol{\Sigma}_{\boldsymbol{x}}^\top \boldsymbol{\theta}_0 \boldsymbol{\theta}_0^\top \boldsymbol{\Sigma}_{\boldsymbol{x}} \right) \boldsymbol{\Sigma}_{\boldsymbol{x}} \boldsymbol{\Lambda}^{-2} \right] \\ m = \frac{\hat{m}}{d} \operatorname{tr} \boldsymbol{\Sigma}_{\boldsymbol{x}}^\top \boldsymbol{\theta}_0 \boldsymbol{\theta}_0^\top \boldsymbol{\Sigma}_{\boldsymbol{x}} \boldsymbol{\Lambda}^{-1} \\ A = \frac{1}{d} \operatorname{tr} \left[\left(\hat{q} \boldsymbol{\Sigma}_{\boldsymbol{x}} + \hat{m}^2 \boldsymbol{\Sigma}_{\boldsymbol{x}}^\top \boldsymbol{\theta}_0 \boldsymbol{\theta}_0^\top \boldsymbol{\Sigma}_{\boldsymbol{x}} \right) \boldsymbol{\Sigma}_{\boldsymbol{\delta}} \boldsymbol{\Lambda}^{-2} \right] \\ N = \frac{1}{d} \operatorname{tr} \left[\left(\hat{q} \boldsymbol{\Sigma}_{\boldsymbol{x}} + \hat{m}^2 \boldsymbol{\Sigma}_{\boldsymbol{x}}^\top \boldsymbol{\theta}_0 \boldsymbol{\theta}_0^\top \boldsymbol{\Sigma}_{\boldsymbol{x}} \right) \boldsymbol{\Lambda}^{-2} \right] \end{cases} \tag{A68}$$

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. (A67) and (A68) can be done with the help of a numerical fixed point iteration.

K: Currently the situation is this:

1. 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

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 [5]. 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 $(\boldsymbol{x}_\mu, y_\mu)\ \mu = 1, ..., N$.

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

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{\boldsymbol{x}_\mu^\top \boldsymbol{w}}{\sqrt{d}} + \frac{\varepsilon_t \boldsymbol{w}^\top \boldsymbol{\Sigma}_{\boldsymbol{\delta}} \boldsymbol{w}}{\sqrt{d}\sqrt{\boldsymbol{w}^\top \boldsymbol{w}}}\right)\right) \\
&= \sum_{\mu=1}^{N} \delta_{y_\mu,+1} \log\left(1 + \exp\left(-\frac{\boldsymbol{x}_\mu^\top \boldsymbol{w}}{\sqrt{d}} + \frac{\varepsilon_t \boldsymbol{w}^\top \boldsymbol{\Sigma}_{\boldsymbol{\delta}} \boldsymbol{w}}{\sqrt{d}\sqrt{\boldsymbol{w}^\top \boldsymbol{w}}}\right)\right) + \delta_{y_\mu,-1} \log\left(1 + \exp\left(\frac{\boldsymbol{x}_\mu^\top \boldsymbol{w}}{\sqrt{d}} + \frac{\varepsilon_t \boldsymbol{w}^\top \boldsymbol{\Sigma}_{\boldsymbol{\delta}} \boldsymbol{w}}{\sqrt{d}\sqrt{\boldsymbol{w}^\top \boldsymbol{w}}}\right)\right) \\
&= \sum_{\mu=1}^{N} y_\mu \log\left(1 + \exp\left(-\frac{\boldsymbol{x}_\mu^\top \boldsymbol{w}}{\sqrt{d}} + \frac{\varepsilon_t \boldsymbol{w}^\top \boldsymbol{\Sigma}_{\boldsymbol{\delta}} \boldsymbol{w}}{\sqrt{d}\sqrt{\boldsymbol{w}^\top \boldsymbol{w}}}\right)\right) + (1 - y_\mu) \log\left(1 + \exp\left(\frac{\boldsymbol{x}_\mu^\top \boldsymbol{w}}{\sqrt{d}} + \frac{\varepsilon_t \boldsymbol{w}^\top \boldsymbol{\Sigma}_{\boldsymbol{\delta}} \boldsymbol{w}}{\sqrt{d}\sqrt{\boldsymbol{w}^\top \boldsymbol{w}}}\right)\right) \\
&= \sum_{\mu=1}^{N} -y_\mu \log\left(\frac{\exp\left(\frac{\boldsymbol{x}_\mu^\top \boldsymbol{w}}{\sqrt{d}} - \frac{\varepsilon_t \boldsymbol{w}^\top \boldsymbol{\Sigma}_{\boldsymbol{\delta}} \boldsymbol{w}}{\sqrt{d}\sqrt{\boldsymbol{w}^\top \boldsymbol{w}}}\right)}{1 + \exp\left(\frac{\boldsymbol{x}_\mu^\top \boldsymbol{w}}{\sqrt{d}} - \frac{\varepsilon_t \boldsymbol{w}^\top \boldsymbol{\Sigma}_{\boldsymbol{\delta}} \boldsymbol{w}}{\sqrt{d}\sqrt{\boldsymbol{w}^\top \boldsymbol{w}}}\right)}\right) + (1 - y_\mu) \log\left(1 + \exp\left(\frac{\boldsymbol{x}_\mu^\top \boldsymbol{w}}{\sqrt{d}} + \frac{\varepsilon_t \boldsymbol{w}^\top \boldsymbol{\Sigma}_{\boldsymbol{\delta}} \boldsymbol{w}}{\sqrt{d}\sqrt{\boldsymbol{w}^\top \boldsymbol{w}}}\right)\right) \\
&= \sum_{\mu=1}^{N} -y_\mu \left(\frac{\boldsymbol{x}_\mu^\top \boldsymbol{w}}{\sqrt{d}} - \frac{\varepsilon_t \boldsymbol{w}^\top \boldsymbol{\Sigma}_{\boldsymbol{\delta}} \boldsymbol{w}}{\sqrt{d}\sqrt{\boldsymbol{w}^\top \boldsymbol{w}}}\right) + y_\mu \log\left(1 + \exp\left(\frac{\boldsymbol{x}_\mu^\top \boldsymbol{w}}{\sqrt{d}} - \frac{\varepsilon_t \boldsymbol{w}^\top \boldsymbol{\Sigma}_{\boldsymbol{\delta}} \boldsymbol{w}}{\sqrt{d}\sqrt{\boldsymbol{w}^\top \boldsymbol{w}}}\right)\right) \\
&\quad + (1 - y_\mu) \log\left(1 + \exp\left(\frac{\boldsymbol{x}_\mu^\top \boldsymbol{w}}{\sqrt{d}} + \frac{\varepsilon_t \boldsymbol{w}^\top \boldsymbol{\Sigma}_{\boldsymbol{\delta}} \boldsymbol{w}}{\sqrt{d}\sqrt{\boldsymbol{w}^\top \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 [3] 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

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

[3] showed for the R language how to accurately compute $\log(1 + \exp(x))$, [4] inspire us to achieve faster computation by extending the case distinction by [3] 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

First we notice that

$$\frac{\mathrm{d}}{\mathrm{d}z}[\log(1+e^z)] = \frac{1}{1+e^{-z}}\,, \quad \frac{\mathrm{d}}{\mathrm{d}z}\left[\frac{1}{1+e^{-z}}\right] = \frac{e^z}{(1+e^z)^2}\,, \quad \frac{\mathrm{d}}{\mathrm{d}z}\left[\frac{1}{1+e^z}\right] = -\frac{1}{2\cosh(z)+2}\,. \tag{C5}$$

also we remember the following derivatives

$$\boldsymbol{\nabla}_{\boldsymbol{w}}\left[\boldsymbol{w}^\top \boldsymbol{\Sigma}_{\boldsymbol{\delta}} \boldsymbol{w}\right] = 2\boldsymbol{\Sigma}_{\boldsymbol{\delta}}\boldsymbol{w}\,, \quad \mathrm{Hess}\left(\boldsymbol{w}^\top \boldsymbol{\Sigma}_{\boldsymbol{\delta}} \boldsymbol{w}\right) = 2\boldsymbol{\Sigma}_{\boldsymbol{\delta}}\,, \quad \boldsymbol{\nabla}_{\boldsymbol{w}}[\|\boldsymbol{w}\|_2] = \frac{\boldsymbol{w}}{\|\boldsymbol{w}\|_2}\,, \quad \mathrm{Hess}\left(\|\boldsymbol{w}\|_2\right) = \frac{\mathbf{1}}{\|\boldsymbol{w}\|_2} - \frac{\boldsymbol{w}\boldsymbol{w}^\top}{\|\boldsymbol{w}\|_2^3}\,, \tag{C6}$$

since we suppose that $\boldsymbol{\Sigma}_{\boldsymbol{\delta}}$ is a symmetric matrix.

Call the derivative of the optimal attack $\boldsymbol{h}$

$$\begin{aligned} \boldsymbol{h} = \boldsymbol{\nabla}_{\boldsymbol{w}} B = \boldsymbol{\nabla}_{\boldsymbol{w}}\left[\frac{\varepsilon_t \boldsymbol{w}^\top \boldsymbol{\Sigma}_{\boldsymbol{\delta}} \boldsymbol{w}}{\sqrt{d}\sqrt{\boldsymbol{w}^\top \boldsymbol{w}}}\right] &= \frac{\varepsilon_t}{\sqrt{d}} \frac{\boldsymbol{w}}{(\boldsymbol{w}^T\boldsymbol{w})^{3/2}} \cdot \left[\left((\boldsymbol{\Sigma}_{\boldsymbol{\delta}} + \boldsymbol{\Sigma}_{\boldsymbol{\delta}})\boldsymbol{w}^T\boldsymbol{w}\right) - \boldsymbol{w}^T\boldsymbol{\Sigma}_{\boldsymbol{\delta}}\boldsymbol{w}\right] \\ M := \frac{2\varepsilon_t}{\sqrt{d}\|\boldsymbol{w}\|_2}\boldsymbol{\Sigma}_{\boldsymbol{\delta}}\boldsymbol{w} &- \frac{\varepsilon_t}{\sqrt{d}\|\boldsymbol{w}\|_2^3}\boldsymbol{w}\boldsymbol{\Sigma}_{\boldsymbol{\delta}}\boldsymbol{w}\,\boldsymbol{w} \end{aligned} \tag{C7}$$

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

$$C_\mu := \frac{\boldsymbol{x}_\mu^\top \boldsymbol{w}}{\sqrt{d}} + \frac{\varepsilon_t \boldsymbol{w}^\top \boldsymbol{\Sigma}_{\boldsymbol{\delta}} \boldsymbol{w}}{\sqrt{d}\sqrt{\boldsymbol{w}^\top \boldsymbol{w}}}\,, \qquad \overline{C}_\mu = \frac{\boldsymbol{x}_\mu^\top \boldsymbol{w}}{\sqrt{d}} - \frac{\varepsilon_t \boldsymbol{w}^\top \boldsymbol{\Sigma}_{\boldsymbol{\delta}} \boldsymbol{w}}{\sqrt{d}\sqrt{\boldsymbol{w}^\top \boldsymbol{w}}}\,, \tag{C8}$$

Then we can write the derivative of the loss as

$$\begin{aligned} \boldsymbol{\nabla}_{\boldsymbol{w}}\mathcal{L} &= \sum_{\mu=1}^{N}\left[-y_\mu\frac{\boldsymbol{x}_\mu}{\sqrt{d}} + y_\mu \boldsymbol{h} + \frac{(1-y_\mu)}{1+\exp(-C_\mu)}\left(\frac{\boldsymbol{x}_\mu}{\sqrt{d}} + \boldsymbol{h}\right) + \frac{y_\mu}{1+\exp\left(-\overline{C}_\mu\right)}\left(\frac{\boldsymbol{x}_\mu}{\sqrt{d}} - \boldsymbol{h}\right)\right] \\ &= \sum_{\mu=1}^{N}\left[\boldsymbol{h}\left(\frac{(1-y_\mu)}{1+\exp(-C_\mu)} + \frac{y_\mu}{1+\exp\left(\overline{C}_\mu\right)}\right) + \frac{\boldsymbol{x}_\mu}{\sqrt{d}}\left(\frac{(1-y_\mu)}{1+\exp(-C_\mu)} - \frac{y_\mu}{1+\exp\left(\overline{C}_\mu\right)}\right)\right] \end{aligned} \tag{C9}$$

With this it is easy to see what the loss per sample is when factoring out the data $\boldsymbol{x}$ and the derivative of the optimal attack $\boldsymbol{h}$. 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

To produce this derivative it is easier to derive in index notation with respect to $\boldsymbol{w}_i$. We start from

$$\frac{\partial^2 B}{\partial \boldsymbol{w}_j \partial \boldsymbol{w}_i} = \frac{\partial \boldsymbol{h}_i}{\partial \boldsymbol{w}_j} = \frac{\varepsilon_t}{\sqrt{d}}\left[\frac{2(\boldsymbol{\Sigma}_{\boldsymbol{\delta}})_{ij}}{\|\boldsymbol{w}\|_2} - \frac{2}{\|\boldsymbol{w}\|_2^3}\left[(\boldsymbol{\Sigma}_{\boldsymbol{\delta}}\boldsymbol{w})_i \boldsymbol{w}_j + (\boldsymbol{\Sigma}_{\boldsymbol{\delta}}\boldsymbol{w})_j \boldsymbol{w}_i + \left(\boldsymbol{w}^\top \boldsymbol{\Sigma}_{\boldsymbol{\delta}} \boldsymbol{w}\right)\delta_{ij}\right] + \frac{3}{\|\boldsymbol{w}\|_2^5}\left(\boldsymbol{w}^\top \boldsymbol{\Sigma}_{\boldsymbol{\delta}} \boldsymbol{w}\right)\boldsymbol{w}_i \boldsymbol{w}_j\right] \tag{C10}$$

Now we can continue by considering the second term in eq. (C9) and differentiating it

$$\frac{\boldsymbol{x}_i^\mu}{\sqrt{d}}\left[\frac{e^{C_\mu}}{1+e^{C_\mu}}\left(\frac{\boldsymbol{x}_j^\mu}{\sqrt{d}} + \boldsymbol{h}_j\right) + \frac{1}{2(\cosh(\overline{C})+1)}\left(\frac{\boldsymbol{x}_j^\mu}{\sqrt{d}} - \boldsymbol{h}_j\right)\right] \tag{C11}$$

then we get from the derivative of the first piece a term similar to the previous one with $\boldsymbol{h}_i$ instead of $\boldsymbol{x}_i^\mu/\sqrt{d}$ which is

$$\boldsymbol{h}_i\left[\frac{e^{C_\mu}}{1+e^{C_\mu}}\left(\frac{\boldsymbol{x}_j^\mu}{\sqrt{d}} + \boldsymbol{h}_j\right) - \frac{1}{2(\cosh(\overline{C})+1)}\left(\frac{\boldsymbol{x}_j^\mu}{\sqrt{d}} - \boldsymbol{h}_j\right)\right] \tag{C12}$$

and the last term

$$\frac{\partial^2 B}{\partial \boldsymbol{w}_j \partial \boldsymbol{w}_i}\left(\frac{(1-y_\mu)}{1+\exp(-C_\mu)} + \frac{y_\mu}{1+\exp\left(\overline{C}_\mu\right)}\right) \tag{C13}$$

where the value in front is given by eq. (C10) and to complete it one should sum over all the data points to obtain the hessian.

$$\begin{aligned}\partial_{\boldsymbol{w}^2}\mathcal{L} = &\sum_{\mu=1}^{N} y_\mu \partial_{\boldsymbol{w}} H + \frac{(1-y_\mu)}{1+\exp\left(-C_\mu\right)}\left(\partial_{\boldsymbol{w}} H\right) + \frac{y_\mu}{1+\exp\left(-\overline{C}_\mu\right)}\left(-\partial_{\boldsymbol{w}} H\right) \\ &+ (1-y_\mu)\left(\frac{X_\mu}{\sqrt{d}} + H\right)\partial_{\boldsymbol{w}}\frac{1}{1+\exp\left(-C_\mu\right)} + y_\mu\left(\frac{X_\mu}{\sqrt{d}} - H\right)\partial_{\boldsymbol{w}}\frac{1}{1+\exp\left(-\overline{C}_\mu\right)}\end{aligned} \tag{C14}$$

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

$$\partial_{\boldsymbol{w}} H = \frac{\varepsilon_t}{\sqrt{d}\sqrt{\boldsymbol{w}^T\boldsymbol{w}}}\mathbb{I} - \frac{\varepsilon_t}{\sqrt{d}}\frac{\boldsymbol{w}\boldsymbol{w}^T}{(\boldsymbol{w}^T\boldsymbol{w})^{3/2}} \tag{C15}$$

$$\partial_{\boldsymbol{w}} C_\mu = \frac{X_\mu^T}{\sqrt{d}} + \varepsilon_t\frac{\boldsymbol{w}^T}{\sqrt{d}\sqrt{\boldsymbol{w}^T\boldsymbol{w}}} \quad \partial_{\boldsymbol{w}}\overline{C}_\mu = \frac{X_\mu^T}{\sqrt{d}} - \varepsilon_t\frac{\boldsymbol{w}^T}{\sqrt{d}\sqrt{\boldsymbol{w}^T\boldsymbol{w}}} \tag{C16}$$

$$\partial_{\boldsymbol{w}}\frac{1}{1+\exp(-C)} = \frac{\partial_{\boldsymbol{w}} C}{2\cosh C + 2} \tag{C17}$$

$$\begin{aligned}\partial_{\boldsymbol{w}^2}\mathcal{L} = &\sum_{\mu=1}^{N}\left[\frac{(1-y_\mu)}{1+\exp\left(-C_\mu\right)} + \frac{y_\mu}{1+\exp\left(\overline{C}_\mu\right)}\right]\left(\partial_{\boldsymbol{w}} H\right) \\ &+ \frac{(1-y_\mu)}{2\cosh C_\mu + 2}\left(\frac{X_\mu}{\sqrt{d}} + H\right)\left(\frac{X_\mu}{\sqrt{d}} + H\right)^T + \frac{y_\mu}{2\cosh\overline{C}_\mu + 2}\left(\frac{X_\mu}{\sqrt{d}} - H\right)\left(\frac{X_\mu}{\sqrt{d}} - H\right)^T\end{aligned} \tag{C18}$$

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... K: Try hard finding a dataset with nonconvex hessian. Random Covariate is an example, but the assumption on Schur complement fails and the matrix is not positive definite...