Statistical Physics of Computation 2025 - Exercises

Vittorio Erba, Emanuele Troiani

Week 12

Disclaimer: this is a **huge** exercise, and we debated quite a lot whether it was of an appropriate length. We decided to stick to it, and **we expect you to solve this exercise across both weeks 12 and 13**. The material was difficult to split into two parts. Please, already consider this as active preparation for the exam: there is a sizable amount of references to previous material that should help you revising.
Some questions are marked **(Recap)**: these questions involve explicitly previous content or previously developed skills, and you should be able to answer them quite confidently. If you do not, please spend time on those questions!
Some questions are marked as **(Comp)**: these questions are mostly heavy on computations. We invite you to try to solve them if you want, but it is fine if you just go through the solutions provided in detail and make sure all steps make sense. This is still exam material.
One question is marked as **(Bonus)**: this question covers material that is related and nice to explore, but not included in the exam program.

12.1 A model of linear regression

This week we analyze the most prototypical ML problem there is: regularized linear regression. The model we will study is one where you are given n input-output pairs $\{x_\mu, y_\mu\}_{\mu=1}^n$, with the input $x_\mu \in \mathbb{R}^d$ and the output $y_\mu \in \mathbb{R}$. We assume that each x_μ is independently drawn from $N(0, \mathbb{I}_d)$, while the outputs are generated by a noisy linear model

$$y_\mu = \frac{w^{\star\top} x_\mu}{\sqrt{d}} + \sqrt{\Delta}\xi_\mu \tag{1}$$

with $\Delta \geq 0$ the noise strength, $\xi_\mu \sim N(0,1)$ independently for each sample, and with true weights $w^\star$ generated by a prior distribution $w_i^\star \sim P_*$.

1. **(Recap)** Write the Bayes optimal posterior distribution for $w^\star$ given the dataset $\{x_\mu, y_\mu\}_{\mu=1}^n$. Here you can think as the dataset to be the observation, and the hidden weights to be the signal to infer. What is the BO estimator w.r.t. the mean squared error?

   By Bayes theorem we have

$$\begin{aligned} P_{\text{post}}(w|\{x_\mu, y_\mu\}_{\mu=1}^n) &\propto P_{\text{prior}}(w) P_{\text{out}}(\{x_\mu, y_\mu\}_{\mu=1}^n|w) \\ &= \prod_{i=1}^d P_*(w_i) \prod_{\mu=1}^n N(y_\mu; \frac{w^\top x_\mu}{\sqrt{d}}; \Delta) N(x_\mu, 0, \mathbb{I}_d) \\ &\propto \prod_{i=1}^d P_*(w_i) \prod_{\mu=1}^n \exp\left(-\frac{(y_\mu - w^\top x_\mu/\sqrt{d})^2}{2\Delta}\right) \end{aligned} \tag{2}$$

where as usual we kept explicit only terms that depend on w. The BO estimator is the mean of the posterior.

As an alternative to Bayes optimal learning, which is not always algorithmically easy to perform, we may try to learn the association from the input to the output with a linear model model of the form

$$\hat{f}(x; w) = \frac{\hat{w}^\top x}{\sqrt{d}} \tag{3}$$

where $\hat{w}$ is a d-dimensional weights vector we will have to find. As discussed in the lecture, we usually learn $\hat{w}$ by empirical risk minimization

$$\hat{w} = \arg\min_w \mathcal{R}(w)\,, \qquad \mathcal{R}(w) = \sum_\mu \left[y_\mu - \frac{w^\top x_\mu}{\sqrt{d}} \right]^2 + \lambda \sum_{i=1}^d \rho(w_i) \tag{4}$$

The first piece is the data-part of the loss, and it expresses how well the weights fit the training dataset. The second piece is called the regularization, and it's used to constraint w (for e.g. avoiding that it goes to infinity) and to avoid the problem being ill-defined (for e.g. if the data part of the loss has multiple global minima). The strength of the regularization compared with the data part of the loss is given by λ. Some special forms include $\rho(w) = w^2/2$, which is typically referred to as ridge regression, and $\rho(w) = |w|$, typically called LASSO. Notice that both choices of regularization make the overall risk function strictly convex, i.e. they guarantee that there exists a unique global minimum (no spurious local ones!).

1. **(Recap)** We can think of the risk as an energy functional, which we will then want to minimize. Assume that the risk (4) has a single global minimum. Argue that, for a given dataset and set of weights $w^\star$, we have

$$\frac{1}{d} \min_w \mathcal{R}(w) = -\lim_{\beta \to +\infty} \frac{1}{d} \partial_\beta \log \mathcal{Z} = -\lim_{\beta \to +\infty} \frac{1}{d\beta} \log \mathcal{Z} \tag{5}$$

and

$$\lim_{\beta \to +\infty} \frac{e^{-\beta \mathcal{R}(w)}}{\mathcal{Z}} = \delta(w - \arg\min_{w'} \mathcal{R}(w')) \tag{6}$$

where

$$\mathcal{Z} = \int \mathrm{d}w\, e^{-\beta\, \mathcal{R}(w)}\,. \tag{7}$$

This means that the order parameters we may find studying $\mathcal{Z}$ will describe the behaviour of the global minimum in the limit $\beta \to \infty$.

Recall Exercise 1 and 2. There we argued that a Gibbs measure with Hamiltonian H converges to the uniform measure over global minima, which here by assumption is a single set of weights. Additionally, we showed that $-\partial_\beta \log Z / d$ computes the average energy of the system, hence its limit computes the energy of the global minimum. The final equality of the energy relationship is a consequence of L'Hopital's Rule.

2. **(Recap)** Consider the Gibbs distribution

$$p_\beta(w) \propto e^{-\beta \mathcal{R}(w)} \tag{8}$$

for the risk in (4). For which value of β and $\rho(w)$ does the Gibbs distribution coincide with the posterior for the inference problem for $w^\star$?

We compare the two distributions. The Gibbs distribution is

$$\begin{aligned} p_\beta(w) &\propto \exp\left(-\beta \sum_\mu \left[y_\mu - \frac{w^\top x_\mu}{\sqrt{d}}\right]^2 - \beta\lambda \sum_{i=1}^d \rho(w_i)\right) \\ &\propto \prod_{i=1}^d \exp\left(-\beta\lambda\rho(w_i)\right) \prod_{\mu=1}^n \exp\left(-\beta\left[y_\mu - \frac{w^\top x_\mu}{\sqrt{d}}\right]^2\right) \end{aligned} \tag{9}$$

and the posterior you computed above is

$$P_{\text{post}}(w|\{x_\mu, y_\mu\}_{\mu=1}^n) \propto \prod_{i=1}^d P_*(w_i) \prod_{\mu=1}^n \exp\left(-\frac{1}{2\Delta}\left[y_\mu - \frac{w^\top x_\mu}{\sqrt{d}}\right]^2\right) \tag{10}$$

We see that the two distributions coincide for $\beta = 1/(2\Delta)$ and choosing

$$\rho(w) = -\frac{1}{\beta\lambda} \log P_*(w)\,. \tag{11}$$

Thus, up to here we learn that:

- We can study the minimization problem (4) by writing a Gibbs distribution and studying its low-temperature limit $\beta \gg 1$.
- We could actually reinterpret the Gibbs distribution at finite temperature as the posterior distribution, if we choose properly the temperature and the regularization.

Our (ambitious) aim today is to set up a single replica computation that, for generic prior $P_\star$, can be adapted to study both the Bayes optimal estimator (by setting β, ρ appropriately) and the empirical risk minimizer (by setting $\beta \to \infty$ and choosing an appropriate regularization ρ). This is the power of replicas: when you work hard to solve a model, if you work in a generic enough setting, you get many different computations for free!

12.2 Replica computation for generic prior and regularization

Let's study the Gibbs measure (8) using the replica method. We will assume that either ρ is a strictly convex function, guaranteeing that $\mathcal{R}(w)$ has a unique global minimum, or that we are in the Bayes optimal setting (above choices for β, ρ). In both cases, the Replica Symmetric ansatz will be correct. Finally, we keep $\beta > 0$ and $P_\star$ generic for the moment and consider the high-dimensional limit $d, n \gg 1$.

1. **(Recap)** By self-averaging (review lecture 4), we will consider the averaged version of the problem, averaging over the dataset $\{x_\mu, y_\mu\}_{\mu=1}^n$ (thus including the label noise ξ_μ) and the teacher weights $w^\star$. Let's setup the associated replica computation. Replicate the partition function and use the explicit expression of y_μ to write

$$\mathbb{E}_{x,\xi,w^\star}[\mathcal{Z}^r] = \mathbb{E}_{x,\xi,w^\star}\left[\int \mathrm{d}w_a e^{-\beta\sum_{a=1}^r \left[\sum_{\mu=1}^n [(w_0 - w_a)x_\mu^\top/\sqrt{d} + \sqrt{\Delta}\xi_\mu]^2 + \lambda \sum_{i=1}^d \rho(w_a^i)\right]}\right] \tag{12}$$

where we identified with the zeroth replica $w^\star = w_0$ for notational simplicity, and the integration is over $w_a \in \mathbb{R}^d$ for $a = 1, \dots, r$.

One starts by replicating

$$\mathbb{E}_{x,\xi,w^\star}[\mathcal{Z}^r] = \mathbb{E}_{x,\xi,w^\star}\left[\int \mathrm{d}w_a e^{-\beta\sum_{a=1}^r \mathcal{R}(w_a)}\right] \tag{13}$$

where the in the integration measure we integrate over all the w_a for $a = 1, ..., n$ (not on $a = 0$, this is done in the disorder average). Now we use the explicit form of $\mathcal{R}(w)$

$$\mathbb{E}_{x,\xi,w^\star}[\mathcal{Z}^r] = \mathbb{E}_{x,\xi,w^\star}\left[\int \mathrm{d}w_a e^{-\beta\sum_{a=1}^r\left[\sum_{\mu=1}^n (y_\mu - w_a^\top x_\mu/\sqrt{d})^2 + \lambda\sum_{i=1}^d \rho(w_a^i)\right]}\right] \tag{14}$$

Now we use the explicit expression for y_μ (1) to get the result.

2. Notice how the integrand depends on the input data x_μ (over which we want to average) only through the scalar pre-activations h^a for $a = 0, 1, ..., n$

$$h_\mu^a = \frac{w_a^\top x_\mu}{\sqrt{d}}\,. \tag{15}$$

Assuming all w_a are fixed, compute the distribution over the preactivations $\{h_a^\mu\}_{\substack{\mu=1,\dots,n\\ a=0,\dots,r}}$ induced by the randomness on $\{x_\mu\}_{\mu=1,\dots,n}$. Hint: recall that $x_\mu \sim N(0, \mathbb{I}_d)$, and that linear combinations of Gaussians are Gaussians.

Since x_μ are all independent standard Gaussian variables, the h_μ^a are also Gaussian variables as they are just linear combinations of the x_μ. They are independent for each index μ as each μ has a different realization of x_μ, but not in the index a, as all index a for a given μ use the same realization of x_μ. To characterize a Gaussian distribution, we need to compute the first two moments. For the mean we have

$$\mathbb{E}_{x^\mu} h_\mu^a = \frac{w_a^\top \mathbb{E}_{x^\mu} x_\mu}{\sqrt{d}} = 0 \tag{16}$$

while for the covariance we have

$$\begin{aligned}\mathbb{E}_{x^\mu,x^\nu} h_\mu^a h_\nu^b &= \mathbb{E}_{x^\mu,x^\nu} \frac{w_a^\top x_\mu}{\sqrt{d}}\frac{w_b^\top x_\nu}{\sqrt{d}} = \sum_{ij=1}^d \frac{w_{a,i}w_{b,j}}{d}\mathbb{E}_{x^\mu,x^\nu} x_i^\mu x_j^\nu = \sum_{ij=1}^d \frac{w_{a,i}w_{b,j}}{d}\delta_{\mu\nu}\delta_{ij}\\ &= \delta_{\mu\nu}\frac{1}{d}\sum_{i=1}^d w_{a,i}w_{b,i} = \delta_{\mu\nu}\frac{w_a^\top w_b}{d}\,.\end{aligned} \tag{17}$$

In formulas, calling μ the distribution of the preactivations, we have

$$\mu(\{h_a^\mu\}_{\substack{\mu=1,\dots,n\\ a=0,\dots,r}}) = \mathbb{E}_x \prod_{\substack{\mu=1,\dots,n\\ a=0,\dots,r}} \delta\left(h_\mu^a - \frac{w_a^\top x_\mu}{\sqrt{d}}\right) = N\left(\{h_a^\mu\}_{\substack{\mu=1,\dots,n\\ a=0,\dots,r}}, 0, \delta_{\mu\nu}\frac{w_a^\top w_b}{d}\right)\,. \tag{18}$$

3. From here onwards let's call $q_{ab}(w)$ the covariance of h^a_μ for fixed values of w_a

$$h^a_\mu \sim \mathcal{N}(0, q_{ab}(w)\,\delta_{\mu\nu})\,, \qquad\qquad q_{ab}(w) = \frac{w_a^\top w_b}{d}\,. \tag{19}$$

Notice that $q_{ab}(w)$ is nothing else than the overlap between weights of replicas a and b, the same overlap we found in all previous replica computations! Use the previous point to perform the average over the input data x_μ, finding that

$$\begin{aligned}&\mathbb{E}_{x,\xi,w_0}[\mathcal{Z}^r]\\ &= \mathbb{E}_{w_0}\int \mathrm{d}w_a e^{-\beta\lambda\sum_{a=1}^r\sum_{i=1}^d \rho(w_a^i)}\mathbb{E}_{\substack{h^a_\mu\sim\mathcal{N}(0,q_{ab}(w_a)\,\delta_{\mu\nu})\\ \xi_\mu\sim\mathcal{N}(0,1)}}\prod_{\mu=1}^n e^{-\beta\sum_{a=1}^r(h^0_\mu - h^a_\mu+\sqrt{\Delta}\xi_\mu)^2}\end{aligned} \tag{20}$$

First we "decouple" all the samples by moving the $\sum_{\mu=1}^n$ out of the exponential. Then we use that the distribution over h induced by x at fixed w is the Gaussian found in the previous point to convert the average over x to an average over h. In equations, one has

$$\begin{aligned}\mathbb{E}_x f\left(\{h^\mu_a(x,w)\}_{\substack{\mu=1,\dots,n\\ a=0,\dots,r}}\right) &= \int dh^\mu_a f\left(\{h^\mu_a\}_{\substack{\mu=1,\dots,n\\ a=0,\dots,r}}\right)\mathbb{E}_x\prod_{\substack{\mu=1,\dots,n\\ a=0,\dots,r}}\delta\left(h^a_\mu - \frac{w_a^\top x_\mu}{\sqrt{d}}\right)\\ &= \int dh^\mu_a f\left(\{h^\mu_a\}_{\substack{\mu=1,\dots,n\\ a=0,\dots,r}}\right) N\left(\{h^\mu_a\}_{\substack{\mu=1,\dots,n,\\ a=0,\dots,r}}, 0, \delta_{\mu\nu}\frac{w_a^\top w_b}{d}\right)\end{aligned} \tag{21}$$

and get the result. In this expression, the integral over h^μ_a is over $\mu = 1,\dots,n$ and $a = 0,\dots,r$.

4. **(Recap)** Argue that we can drop the index μ and write

$$\begin{aligned}&\mathbb{E}_{x,\xi,w_0}[\mathcal{Z}^r]\\ &= \mathbb{E}_{w_0}\int \mathrm{d}w_a e^{-\beta\lambda\sum_{a=1}^r\sum_{i=1}^d \rho(w_a^i)}\left[\mathbb{E}_{\substack{h\sim\mathcal{N}(0,q_{ab}(w_a))\\ \xi\sim\mathcal{N}(0,1)}} e^{-\beta\sum_{a=1}^r(h^0 - h^a+\sqrt{\Delta}\xi)^2}\right]^n.\end{aligned} \tag{22}$$

Since the h^a_μ and the noise ξ_μ are independent for different indices μ (the h have a term $\delta_{\mu\nu}$ in their covariance!) and the integrand is factorized over samples μ, the index μ is mute and the product factorizes as shown in the equation.

5. **(Recap)** We are now ready to decouple the regularization from the loss part by introducing the usual order parameters, the overlaps. Introduce a Dirac delta with the definition of the overlap q, as well as its Fourier conjugate $\hat{q}$, to write

$$\mathbb{E}_{x,\xi,w_0}[\mathcal{Z}^r] = \int \mathrm{d}q_{ab}\,\mathrm{d}\hat{q}_{ab}\,\exp\left\{-d\sum_{0\le a\le b\le n} q_{ab}\hat{q}_{ab} + d\log I_{\mathrm{reg}}(\hat{q}_{ab}) + n\log I_{\mathrm{loss}}(q_{ab})\right\}. \tag{23}$$

where the integration over $q_{ab}, \hat{q}_{ab}$ is over all indices $0 \le a \le b \le r$, and where

$$I_{\mathrm{loss}}(\hat{q}) = \mathbb{E}_{\substack{h\sim\mathcal{N}(0,q_{ab})\\ \xi\sim\mathcal{N}(0,1)}} e^{-\beta\sum_{a=1}^r(h^0 - h^a+\sqrt{\Delta}\xi)^2}\,, \tag{24}$$

and

$$I_{\rm reg}(q) = \int_{\mathbb{R}} dw_0\, P_*(w_0) \int_{\mathbb{R}} \mathrm{d}w_a e^{-\beta\lambda \sum_{a=1}^r \rho(w_a) + \sum_{0\le a\le b\le r} \hat{q}_{ab} w_a w_b}\,, \tag{25}$$

where we stressed that the integrals over w_a are not anymore in $\mathbb{R}^d$, but just in $\mathbb{R}$.

We introduce the definition of the overlap to get

$$\begin{aligned}\mathbb{E}_{x,\xi,w_0}[\mathcal{Z}^r] = \int \mathrm{d}q_{ab} \left[\mathbb{E}_{w_0} \int \mathrm{d}w_a e^{-\beta\lambda \sum_{a=1}^r \sum_{i=1}^d \rho(w_a^i)} \prod_{0\le a\le b\le r} \delta(d\, q_{ab} - w_a^\top w_b)\right] \\ \times \left[\mathop{\mathbb{E}_{h\sim\mathcal{N}(0,q_{ab})}}_{\xi\sim\mathcal{N}(0,1)} e^{-\beta \sum_{a=1}^r (h^0 - h^a + \sqrt{\Delta}\xi)^2}\right]^n,\end{aligned} \tag{26}$$

$I_{\rm loss}$ is then directly the quantity in the second bracket. Then, we introduce the Fourier representation of the Dirac's delta (plus the usual Wick rotation)

$$\prod_{a,b} \delta(d\, q_{ab} - w_a^\top w_b) = \int \mathrm{d}\hat{q}_{ab}\, e^{\sum_{a,b} \left(-d\, \hat{q}_{ab} q_{ab} + \sum_{i=1}^d \hat{q}_{ab} w_a^i w_b^i\right)}\,, \tag{27}$$

to get

$$\begin{aligned}\mathbb{E}_{x,\xi,w_0}[\mathcal{Z}^r] = \int \mathrm{d}q_{ab}\, \mathrm{d}\hat{q}_{ab} e^{-d \sum_{0\le a\le b\le r} q_{ab}\hat{q}_{ab}} I_{\rm loss}^n \\ \mathbb{E}_{w_0} \int \mathrm{d}w_a e^{-\beta\lambda \sum_{a=1}^r \sum_{i=1}^d \rho(w_a^i) + \sum_{0\le a\le b\le r} \sum_{i=1}^d \hat{q}_{ab} w_a^i w_b^i}\,,\end{aligned} \tag{28}$$

and finally we notice that the whole part of the integrand depending on w factorizes over the dimension index $i = 1, \dots, d$ (the integrand explicitly, while the average over w_0 by the assumption that its distribution is factorized over dimension indices). Thus, we get

$$\begin{aligned}\mathbb{E}_{x,\xi,w_0}[\mathcal{Z}^r] = \int \mathrm{d}q_{ab}\, \mathrm{d}\hat{q}_{ab} e^{-d \sum_{0\le a\le b\le r} q_{ab}\hat{q}_{ab}} I_{\rm loss}^n \\ \left[\mathbb{E}_{w_0} \int \mathrm{d}w_a e^{-\beta\lambda \sum_{a=1}^r \rho(w_a) + \sum_{0\le a\le b\le r} \hat{q}_{ab} w_a w_b}\right]^d,\end{aligned} \tag{29}$$

where crucially all integrals over w, w_0 ar now over $\mathbb{R}$, and not anymore over $\mathbb{R}^d$. The expression in the first bracket is then $I_{\rm reg}$.

6. We are now at a common step of replica computations. We decoupled all high-dimensional integrals, meaning that both n and d enter only parametrically (and not anymore structurally) in our partition function. On the other hand, we have the usual dependence on the $r(r+1)$ order parameter q and $\hat{q}$, which makes it difficult to take the limit $r \to 0$. To continue, we impose a Replica Symmetric ansatz. In the context of empirical risk minimization, we do not have Nishimori's identities, so that we need to keep track of the magnetization m between replicas w_a and ground truth w_0 separate from the replica-replica overlap q. We have the following ansatz then

Entry	Value	Entry	Value
q_{00}	$Q^\star$	$\hat{q}_{00}$	$\hat{Q}^\star$
q_{0a} $(0 < a < r)$	m	$\hat{q}_{0a}$ $(0 < a < r)$	$\hat{m}$
q_{ab} $(0 < a < b < r)$	q	$\hat{q}_{ab}$ $(0 < a < b < r)$	$\hat{q}$
q_{aa} $(0 < a < r)$	Q	$\hat{q}_{aa}$ $(0 < a < r)$	$-\hat{Q}/2$

(30)

where the sign and constant of $\hat{Q}$ are taken for later convenience. What we are asking here is that at the saddle point, all replicas are equivalent apart from the ground truth one w_0. Show that one has at first order in $r \ll 1$

$$\sum_{0\le a\le b\le r} q_{ab}\hat{q}_{ab} = Q^\star\hat{Q}^\star + r\left[m\hat{m} - \frac{Q\hat{Q}}{2} - \frac{q\hat{q}}{2}\right] + \mathcal{O}(r^2)\,. \tag{31}$$

One has to treat each piece in the sum differently depending on whether $a = b$ or not, and whether $a = 0$ or not, obtaining for cases, one per RS overlap ($Q^\star$,Q,m,q). The factors in front count how many addends have the same RS overlaps: recall that a matrix of size r has r^2 entries and r entries on the diagonal, and above the diagonal one has $r(r-1)/2$. This gives

$$\begin{aligned}\sum_{0\le a\le b\le r} q_{ab}\hat{q}_{ab} &= q_{00}\hat{q}_{00} + \sum_{1\le b\le r} q_{0b}\hat{q}_{0b} + \sum_{1\le a\le r} q_{aa}\hat{q}_{aa} + \sum_{1\le a<b\le r} q_{ab}\hat{q}_{ab} \\ &= Q^\star\hat{Q}^\star + rm\hat{m} - \frac{r}{2}Q\hat{Q} + \frac{r(r-1)}{2}q\hat{q}\,,\end{aligned} \tag{32}$$

which in turns gives the result with the approximation at leading order $(r^2 - r)/2 \approx -r/2$.

7. Notice that in (31) we have a term of order $O(1)$ instead of $O(r)$, which may give us problems when doing the replica trick. We should check that this term vanishes at the saddle point. Consider for a moment the case $r = 0$, where $\mathbb{E}[\mathcal{Z}^r] = 1$. Argue that $\hat{Q}^\star = 0$ at the saddle point. What is the value of $Q^\star$ at the saddle point?

   For $r = 0$ we have

$$\mathbb{E}_{x,\xi,w_0}[\mathcal{Z}^r] = \int \mathrm{d}Q^\star\,\mathrm{d}\hat{Q}^\star e^{-dQ^\star\hat{Q}^\star + d\log\mathbb{E}_{w_0}e^{\hat{Q}^\star w_0^2}}\,, \tag{33}$$

   where we used that $\log I_{\text{loss}} = 0$ for $r = 0$ as its integrand becomes equal to 1. At the saddle point we get

$$\partial_{Q^\star}\left[-Q^\star\hat{Q}^\star + \log\mathbb{E}_{w_0}e^{\hat{Q}^\star w_0^2}\right] = 0 \implies \hat{Q}^\star = 0 \tag{34}$$

   and

$$\partial_{\hat{Q}^\star}\left[-Q^\star\hat{Q}^\star + \log\mathbb{E}_{w_0}e^{-\hat{Q}^\star w_0^2}\right] = 0 \implies Q^\star = \frac{\mathbb{E}_{w_0}e^{\hat{Q}^\star w_0^2}w_0^2}{\mathbb{E}_{w_0}e^{\hat{Q}^\star w_0^2}} = \mathbb{E}_{w_0}w_0^2 \tag{35}$$

   where in the last step we used $\hat{Q}^\star = 0$. The interpretation is that the norm squared of the teacher weights $w^\star = w_0$ is concentrating for large d to a deterministic value $Q^\star$ by the law of large numbers. Notice that at the saddle point

$$\mathbb{E}_{x,\xi,w_0}[\mathcal{Z}^{r=0}] = \int \mathrm{d}Q^\star\,\mathrm{d}\hat{Q}^\star e^{-dQ^\star\hat{Q}^\star + d\log\mathbb{E}_{w_0}e^{\hat{Q}^\star w_0^2}} = e^0 = 1 \tag{36}$$

   as expected.

Computing I_{loss} explicitly in the RS ansatz involves an $r+1$-dimensional Gaussian integral (h, ξ are Gaussian, and the integrand is the exponential of a quadratic function), but the algebra is painfully involved, so we give you directly the result

$$\log I_{\text{loss}}(q_{ab}) = -\frac{r}{2}\left[\log(1+\beta(Q-q)) + \frac{2\beta(Q^\star - 2m + q + \Delta)}{1+2\beta(Q-q)}\right] + \mathcal{O}(r^2) \tag{37}$$

If you are really interested in the derivation, check for e.g. `https://arxiv.org/pdf/2006.06560` Eq.91 and following (there the derivation is performed in a more general setting).

8. **(Comp)** Let's now compute the regularization/prior term in the RS ansatz. Use the RS ansatz, the fact that $\hat{Q}^\star = 0$, the Hubbard-Stratonovich trick to decouple replicas, and the limit $r \to 0$ to get, at leading order

$$\log(I_{\text{reg}}) = r \int Dz \int dw_0 P_*(w_0) \log \int \mathrm{d}w e^{-\beta\lambda\rho(w) - \frac{\hat{Q}+\hat{q}}{2}w^2 + (\hat{m}w_0 + z\sqrt{\hat{q}})w} + \mathcal{O}(r^2)\,, \quad (38)$$

where Dz is integration over a standard Gaussian measure.

The procedure is very similar to Exercise 5 (point 8), with slightly more terms due to the lack of Nishimori identities. We have

$$\begin{aligned}
&\mathbb{E}_{w^\star} \int \mathrm{d}w_a e^{-\beta\lambda\sum_{a=1}^r \rho(w_a) - \sum_{0\le a\le b\le r}\hat{q}_{ab}w_a w_b} \\
&\overset{(a)}{=} \int dw_0 P_*(w_0) \int \mathrm{d}w_a e^{-\beta\lambda\sum_{a=1}^r \rho(w_a)} e^{-\frac{\hat{Q}}{2}\sum_{a=1}^r w_a^2 + \hat{m}\sum_{a=1}^r w_a w_0 + \hat{q}\sum_{1\le a<b\le r} w_a w_b} \\
&\overset{(b)}{=} \int dw_0 P_*(w_0) \int \mathrm{d}w_a e^{-\beta\lambda\sum_{a=1}^r \rho(w_a)} e^{-\frac{\hat{Q}}{2}\sum_{a=1}^r w_a^2 + \hat{m}\sum_{a=1}^r w_a w_0 - \frac{\hat{q}}{2}\sum_{a=1}^r w_a^2 + \frac{\hat{q}}{2}\sum_{a,b=1}^r w_a w_b} \\
&= \int dw_0 P_*(w_0) \int \mathrm{d}w_a e^{-\beta\lambda\sum_{a=1}^r \rho(w_a)} e^{-\frac{\hat{Q}+\hat{q}}{2}\sum_{a=1}^r w_a^2 + \hat{m}\sum_{a=1}^r w_a w_0 + \frac{\hat{q}}{2}(\sum_{a=1}^r w_a)^2} \\
&\overset{(c)}{=} \int Dz \int dw_0 P_*(w_0) \int \mathrm{d}w_a e^{-\beta\lambda\sum_{a=1}^r \rho(w_a)} e^{-\frac{\hat{Q}+\hat{q}}{2}\sum_{a=1}^r w_a^2 + \hat{m}\sum_{a=1}^r w_a w_0 + z\sqrt{\hat{q}}\sum_{a=1}^r w_a} \\
&\overset{(d)}{=} \int Dz \int dw_0 P_*(w_0) \left[\int \mathrm{d}w e^{-\beta\lambda\rho(w)} e^{-\frac{\hat{Q}+\hat{q}}{2}w^2 + (\hat{m}w_0 + z\sqrt{\hat{q}})w}\right]^r \\
&\overset{(e)}{=} \int Dz \int dw_0 P_*(w_0) \exp\left(r \log \int \mathrm{d}w e^{-\beta\lambda\rho(w) - \frac{\hat{Q}+\hat{q}}{2}w^2 + (\hat{m}w_0 + z\sqrt{\hat{q}})w}\right) \\
&\overset{(f)}{=} \int Dz \int dw_0 P_*(w_0) \left(1 + r \log \int \mathrm{d}w e^{-\beta\lambda\rho(w) - \frac{\hat{Q}+\hat{q}}{2}w^2 + (\hat{m}w_0 + z\sqrt{\hat{q}})w} + \mathcal{O}(r^2)\right) \\
&\overset{(g)}{=} 1 + r \int Dz \int dw_0 P_*(w_0) \log \int \mathrm{d}w e^{-\beta\lambda\rho(w) - \frac{\hat{Q}+\hat{q}}{2}w^2 + (\hat{m}w_0 + z\sqrt{\hat{q}})w} + \mathcal{O}(r^2)
\end{aligned} \quad (39)$$

where

- in (a) we inserted the RS ansatz
- in (b) we completed the last double sum to highlight that it is a square of a single sum
- in (c) we used Hubbard-Stratonovich to convert the square of the sum to a single sum, decoupling replicas (meaning that now the replica index is a mute index!)
- in (d) we write explicitly that replicas decoupled
- in (e) we use $x^r = e^{r\log x}$
- in (f) we expand the exponential for small r
- in (g) we use that

$$\int Dz \int dw_0 P_*(w_0) = 1 \quad (40)$$

by normalization

Applying the final log and expanding again for small r gives the result.

9. **(Recap)** Use the saddle-point and the replica trick to finally find that

$$\begin{aligned}
&\frac{1}{d}\log \mathbb{E}_{x,\xi,w_0}[\mathcal{Z}] \to \Phi \\
&\text{where} \quad \Phi = \text{extr}_{Q,m,q,\hat{Q},\hat{m},\hat{q}}\Bigg[\frac{Q\hat{Q}}{2} - m\hat{m} + \frac{q\hat{q}}{2} \\
&\qquad + \int Dz \int dw_0 P_*(w_0) \log \int \mathrm{d}w e^{-\beta\lambda\rho(w) - \frac{\hat{Q}+\hat{q}}{2}w^2 + (\hat{m}w_0 + z\sqrt{\hat{q}})w} \\
&\qquad - \frac{n}{d}\left[\frac{1}{2}\log(1+\beta(Q-q)) + \frac{\beta(Q^\star - 2m + q + \Delta)}{1+2\beta(Q-q)}\right]\Bigg]
\end{aligned} \tag{41}$$

By the replica trick we have

$$\frac{1}{d}\log \mathbb{E}_{x,\xi,w_0}[\mathcal{Z}] = \frac{1}{d}\lim_{r\to 0}\frac{\mathbb{E}[\mathcal{Z}^r]-1}{r}\,. \tag{42}$$

By saddle point we have

$$\mathbb{E}[\mathcal{Z}^r] = \exp\left(rd\Phi\right)\quad, \tag{43}$$

with Φ defined as above. Expanding for small r we then get

$$\frac{1}{d}\log \mathbb{E}_{x,\xi,w_0}[\mathcal{Z}] = \frac{1}{d}\lim_{r\to 0}\frac{\exp\left(rd\Phi\right)-1}{r} = \Phi\,. \tag{44}$$

10. What happens if $n \ll d$? Argue that sending $n, d \to \infty$ with fixed ratio α is the most interesting regime to look at.

    If $n \ll d$ then the loss part of the free entropy, the only thing in the free entropy that had a dependence on the training set, is subleading. This implies that we would have obtained the same free entropy if we had no training data at all ($n = 0$). Thus, there is not enough data to learn anything meaningful about $w^\star$. The Gibbs distribution of the problem reduces to just the regularization part.
    If $n \sim d$ then the loss part of the free entropy, depending on the data, is of the same order as the regularization part. It means that we have enough data to start learning something. We thus expect this to be the most interesting regime to study.

The expression (41) we found is valid for all values of β, P^* and ρ. We stress that replica symmetry holds for convex ρ, which is a very non-trivial result to show, or if we are in the Bayesian choice for β, ρ. From now on, we will consider $n = \alpha d$ for fixed α.

12.3 Bayes optimal estimation

Let us consider the Bayes optimal estimator for $w^\star$. In this case, we saw above that the posterior corresponds to the choices $\beta = 1/(2\Delta)$ and

$$\rho(w) = -\frac{1}{\beta\lambda}\log P_*(w)\,. \tag{45}$$

1. Show that in this case

$$\Phi = \mathrm{extr}_{q,\hat{q}} \Big[-\frac{q\hat{q}}{2} - \frac{\alpha}{2}\log\left(2\Delta + Q^\star - q\right) + \int Dz \int dw_0 P_*(w_0) \log \int \mathrm{d}w P_*(w) e^{-\frac{\hat{q}}{2}w^2 + (\hat{q}w_0 + z\sqrt{\hat{q}})w} \Big] \tag{46}$$

Hint: use Nishimori's identites, both for the normal overlaps and for the hat overlaps, and discard constant additive terms.

Nishimori's identities imply

$$Q = Q^\star\,, \quad \hat{Q} = \hat{Q}^\star = 0\,, \quad m = q\,, \quad \hat{m} = \hat{q}\,, \tag{47}$$

and additionally we substitute the values of β, ρ.

2. **(Recap)** The prior term

$$\int Dz \int dw_0 P_*(w_0) \log \int \mathrm{d}w P_*(w) e^{-\frac{\hat{q}}{2}w^2 + (\hat{q}w_0 + z\sqrt{\hat{q}})w} \tag{48}$$

should be familiar. In which problems did we find this term already?

This term is the Bayes optimal free entropy of a scalar denoising problem (Exercise 3), which already appeared in a replica computation in the case of the spiked-Wigner model (Exercise 5, point 10).

3. **(Comp)** Compare back to (Exercise 5, point 11/12) to show that the state equations for our model are given by

$$\hat{q} = \frac{\alpha}{2\Delta + Q^\star - q} \tag{49}$$

and

$$q = \int Dz \int dw_0 P_*(w_0) \frac{\int \mathrm{d}w P_*(w) e^{-\frac{\hat{q}}{2}w^2 + (\hat{q}w_0 + z\sqrt{\hat{q}})w} w w_0}{\int \mathrm{d}w P_*(w) e^{-\frac{\hat{q}}{2}w^2 + (\hat{q}w_0 + z\sqrt{\hat{q}})w}} \tag{50}$$

The state equations are given by the zero derivative condition of Φ. The first is simple, it is just a derivative. For the second, one needs to go back to Exercise 5, point 11 and 12, and compare with what we found there in order to "skip" the computation of the derivative of the integral.

4. **(Recap)** Argue that the Bayes optimal estimation error

$$e_{\mathrm{est}} = \frac{1}{d}\|w^\star - \hat{w}_{\mathrm{BO}}\|^2 \tag{51}$$

satisfies $e_{\mathrm{est}} = Q^\star - q$ at the saddle point.

See Exercise 3!

5. Argue that the generalization error

$$e_{\mathrm{gen}} = \mathbb{E}_{x_{\mathrm{test}}} \| \frac{w_\star^\top x_{\mathrm{test}}}{\sqrt{d}} - \frac{\hat{w}_{\mathrm{BO}}^\top x_{\mathrm{test}}}{\sqrt{d}} \|^2 \tag{52}$$

satisfies $e_{\mathrm{gen}} = e_{\mathrm{est}}$, where x_{test} is a new test input drawn from $N(0, \mathbb{I}_d)$.

We have

$$
\begin{aligned}
e_{\text{gen}} &= \mathbb{E}_{x_{\text{test}}} \| \frac{w_\star^\top x_{\text{new}}}{\sqrt{d}} - \frac{\hat{w}^\top x_{\text{new}}}{\sqrt{d}} \|^2 \\
&= \mathbb{E}_{x_{\text{test}}} \sum_{i,j=1}^{d} \frac{(w_\star - \hat{w})_i x_i^{\text{test}}}{\sqrt{d}} \frac{(w_\star - \hat{w})_j x_j^{\text{test}}}{\sqrt{d}} \\
&= \sum_{i=1}^{d} \frac{(w_\star - \hat{w})_i^2}{d} = e_{\text{est}} \, .
\end{aligned}
\tag{53}
$$

Thus, we find a complete characterization of the Bayes optimal estimator ant its generalization error for generic P_*!

12.4 Empirical risk minimization: large β limit

Let's now consider the empirical risk minimizer $\hat{w} = \arg\min \mathcal{R}(w)$. To analyze it, we need to compute the $\beta \to \infty$ limit of (41).
We consider the following change of variable for large β. We substitute $(Q, \hat{Q})$ with $(\Sigma, \hat{\Sigma})$ using

$$
Q = q + \Sigma/\beta \quad \text{and} \quad \hat{q} + \hat{Q} = \beta \hat{\Sigma} \tag{54}
$$

The first change of variable is intuitive: as β grows, the measure concentrates more and more around the unique global minimum of the risk, implying that the norm squared of a sample Q becomes closer and closer to the overlap between two samples q. The same argument would not apply if the minimum were not unique. Additionally, we rescale $\hat{q} \to \beta^2 \hat{q}$ and $\hat{m} \to \beta \hat{m}$ (without these rescalings, the saddle point equations would not lead to a consistent result for large β, in the sense that some order parameters would diverge to infinity).

1. Show that at leading order in $\beta \gg 1$ we have

$$
\begin{aligned}
\frac{1}{d\beta} \log \mathbb{E}_{x,\xi,w_0}[\mathcal{Z}] \to \text{extr}_{\Sigma,m,q,\hat{\Sigma},\hat{m},\hat{q}} \Bigg[& \frac{q\hat{\Sigma}}{2} - \frac{\Sigma\hat{q}}{2} - m\hat{m} \\
& + \frac{1}{\beta} \int Dz \int dw_0 P_*(w_0) \log \int \mathrm{d}w e^{\beta\left[-\lambda\rho(w) - \frac{\hat{\Sigma}}{2}w^2 + (\hat{m}w_0 + z\sqrt{\hat{q}})w\right]} \\
& - \frac{n}{d} \frac{Q^\star - 2m + q + \Delta}{1 + 2\Sigma} \Bigg]
\end{aligned}
\tag{55}
$$

Just substitute the change of variable and the rescaling, simplify all terms, and then notice that there are some terms of order $1/\beta$ that can be neglected.

2. **(Recap)** The prior/regularization term is still dependent on β and β is large. Show that at leading order

$$
\begin{aligned}
& \frac{1}{\beta} \int Dz \int dw_0 P_*(w_0) \log \int \mathrm{d}w e^{\beta\left[-\lambda\rho(w) - \frac{\hat{\Sigma}}{2}w^2 + (\hat{m}w_0 + z\sqrt{\hat{q}})w\right]} \\
& \qquad = -\int Dz \int dw_0 P_*(w_0) \min_w \left[\lambda\rho(w) + \frac{\hat{\Sigma}}{2}w^2 - (\hat{m}w_0 + z\sqrt{\hat{q}})w \right]
\end{aligned}
\tag{56}
$$

This is a direct application of the saddle point method in the form

$$\int dx e^{-\beta f(x)} \approx e^{-\beta \min_x f(x)} . \tag{57}$$

Notice that (56) is the average value of the minimum of a scalar minimization problem with risk

$$f(w) = \lambda\rho(w) + \frac{\hat{\Sigma}}{2}w^2 - (\hat{m}w_0 + z\sqrt{\hat{q}})w \tag{58}$$

which is very close to a mean square error penalty $[w - (\hat{m}w_0 + z\sqrt{\hat{q}})]^2$ regularized by $\lambda\rho(w)$. This is reminiscent of the Bayes optimal scalar denoising problem appearing in the prior term, for e.g. in (48). Thus, again, we have a prior term that does some denoising! In this case, it performs denoising by empirical risk minimization on f.

3. Consider the generalization error

$$e_{gen} = \mathbb{E}_{x_{\text{test}}} \| \frac{w_\star^\top}{\sqrt{d}} - \frac{\hat{w}^\top}{\sqrt{d}} \|^2 \tag{59}$$

where $\hat{w} = \hat{w}(\{x, y\})$ is the global minimum of the risk. Compute the test error as a function of the order parameters $Q^\star, m, q$. Here x_{test} is a new test input drawn from $N(0, \mathbb{I}_d)$.

$$\begin{aligned} e_{\text{test}} &= \mathbb{E}_{x_{\text{test}}} \| \frac{w_\star^\top x_{\text{new}}}{\sqrt{d}} - \frac{\hat{w}^\top x_{\text{new}}}{\sqrt{d}} \|^2 \\ &= \mathbb{E}_{x_{\text{test}}} \sum_{i,j=1}^{d} \frac{(w_\star - \hat{w})_i x_i^{\text{test}}}{\sqrt{d}} \frac{(w_\star - \hat{w})_j x_j^{\text{test}}}{\sqrt{d}} \\ &= \sum_{i=1}^{d} \frac{(w_\star - \hat{w})_i^2}{d} \\ &= Q^\star - 2m + Q \\ &= Q^\star - 2m + q \end{aligned} \tag{60}$$

where in the last step we used $Q = q + \Sigma/\beta = q + o_\beta(1)$ for large β.

4. **(Bonus)** The order parameter Σ appearing here has a nice geometric interpretation. Show that

$$\Sigma = \frac{1}{d}\text{Tr}\mathcal{H}[\mathcal{R}](\hat{w})^{-1} \tag{61}$$

where $\hat{w}$ is the global minimum of the risk, $\mathcal{H}[\mathcal{R}]$ is the Hessian of the risk and Tr denotes the trace. This means that Σ is related to the inverse curvature at the global minimum of the loss: if $\Sigma \to \infty$, then it means that the loss develops flat directions around its global minimum, i.e. the global minimum is not unique anymore.

$$Q - q = \frac{\mathbb{E}_{\text{Gibbs}}\|w\|^2 - \|\mathbb{E}_{\text{Gibbs}}w\|^2}{d} = \frac{1}{d}\sum_{i=1}^{d}\left[\mathbb{E}_{\text{Gibbs}}w_i^2 - (\mathbb{E}_{\text{Gibbs}}w_i)^2\right] \tag{62}$$

Now assume that the risk around the global minimum has expansion

$$\mathcal{R}(w) = \mathcal{R}(\hat{w}) + \frac{1}{2}(w-\hat{w})^\top \mathcal{H}[\mathcal{R}](\hat{w})(w-\hat{w}) + \dots \tag{63}$$

where $\mathcal{H}[\mathcal{R}]$ is the Hessian of $\mathcal{R}$. Then the Gibbs measure reads

$$p(w) = \frac{e^{-\frac{\beta}{2}(w-\hat{w})^\top \mathcal{H}[\mathcal{R}](\hat{w})(w-\hat{w})}}{Z} \tag{64}$$

Notice that the Gibbs measure at inverse temperature β will have its mass concentrated on a region around $\hat{w}$ or radius $|w-\hat{w}| = \mathcal{O}(\beta^{-1/2})$ (you see this by imposing that the quadratic term of the loss expansion is of order one when $\beta \gg 1$), implying in particular that all higher order terms of the expansion are subleading in β. This implies that the Gibbs measure is Gaussian for $\beta \gg 1$, and that

$$\frac{\mathbb{E}_{\text{Gibbs}}\|w\|^2 - \|\mathbb{E}_{\text{Gibbs}} w\|^2}{d} = \frac{1}{\beta d}\sum_{i=1}^{d} \mathcal{H}[\mathcal{R}](\hat{w})^{-1}_{ii} = \frac{1}{\beta d}\text{Tr}\mathcal{H}[\mathcal{R}](\hat{w})^{-1} \tag{65}$$

giving that

$$\Sigma = \beta(Q-q) = \frac{1}{d}\text{Tr}\mathcal{H}[\mathcal{R}](\hat{w})^{-1}\,. \tag{66}$$

Thus, we find a description of our system directly at zero temperature as

$$\begin{aligned}\lim_{\beta\to\infty}\lim_{d\to\infty}\frac{1}{d\beta}\log \mathbb{E}_{x,\xi,w_0}[\mathcal{Z}] \to \text{extr}_{\Sigma,m,q,\hat{\Sigma},\hat{m},\hat{q}}\Bigg[\frac{q\hat{\Sigma}}{2} - \frac{\Sigma\hat{q}}{2} - m\hat{m} - \alpha\frac{Q^\star - 2m + q + \Delta}{1+2\Sigma} \\ - \int Dz \int dw_0 P_*(w_0)\min_w\left[\lambda\rho(w) + \frac{\hat{\Sigma}}{2}w^2 - (\hat{m}w_0 + z\sqrt{\hat{q}})w\right]\Bigg]\,.\end{aligned} \tag{67}$$

At this point we could derive directly the state equations for generic ρ and $P_\star$, but that is a bit of a painful computation, so we are not doing it. The state equations obtained deriving w.r.t. Σ, m, q are simple and read:

$$\begin{aligned}\hat{q} &= \frac{4\alpha(\Delta - 2m + q + Q^\star)}{(2\Sigma+1)^2} \\ \hat{\Sigma} &= \frac{2\alpha}{2\Sigma+1} \\ \hat{m} &= \frac{2\alpha}{2\Sigma+1}\end{aligned} \tag{68}$$

.

while the other three involve less immediate expressions and we do not give them here.

We are now ready to study some specific examples.

12.5 Empirical risk minimization: ridge regression

Consider the case of ridge regression $\rho(w) = w^2/2$ for generic prior $P_\star$.

1. Show that

$$\min_w\left[\lambda\rho(w) + \frac{\hat{\Sigma}}{2}w^2 - (\hat{m}w_0 + z\sqrt{\hat{q}})w\right] = -\frac{(\hat{m}w_0 + z\sqrt{\hat{q}})^2}{2(\hat{\Sigma}+\lambda)} \tag{69}$$

Call

$$f(w) = \frac{\hat{\Sigma} + \lambda}{2} w^2 - (\hat{m} w_0 + z\sqrt{\hat{q}}) w \,. \tag{70}$$

The stationarity condition for the minimum reads

$$w = \frac{\hat{m} w_0 + z\sqrt{\hat{q}}}{\hat{\Sigma} + \lambda} \tag{71}$$

giving

$$\min_w \left[\lambda \rho(w) + \frac{\hat{\Sigma}}{2} w^2 - (\hat{m} w_0 + z\sqrt{\hat{q}}) w \right] = f\left(\frac{\hat{m} w_0 + z\sqrt{\hat{q}}}{\hat{\Sigma} + \lambda} \right) = -\frac{(\hat{m} w_0 + z\sqrt{\hat{q}})^2}{2(\hat{\Sigma} + \lambda)} \tag{72}$$

2. Show that

$$-\int Dz \int dw_0 P_*(w_0) \min_w \left[\lambda \rho(w) + \frac{\hat{\Sigma}}{2} w^2 - (\hat{m} w_0 + z\sqrt{\hat{q}}) w \right] = \frac{\hat{m}^2 Q_* + \hat{q}}{2(\hat{\Sigma} + \lambda)} \tag{73}$$

Using the previous point we have

$$\begin{aligned} \int Dz \int dw_0 P_*(w_0) \frac{(\hat{m} w_0 + z\sqrt{\hat{q}})^2}{2(\hat{\Sigma} + \lambda)} &= \int Dz \int dw_0 P_*(w_0) \frac{\hat{m}^2 w_0^2 + z^2 \hat{q} + 2\hat{m} w_0 z\sqrt{\hat{q}}}{2(\hat{\Sigma} + \lambda)} \\ &= \frac{\hat{m}^2 Q_* + \hat{q}}{2(\hat{\Sigma} + \lambda)} \end{aligned} \tag{74}$$

where we used that

$$\int Dz\, z = 0 \qquad \int Dz\, z^2 = 1 \qquad \int dw_0 P_*(w_0) w_0^2 = Q_* \,. \tag{75}$$

3. How will the properties of the global minimum of the risk $\mathcal{R}$ depend on the prior P_*?

   They will depend only on the second moment of the prior $Q^\star$! No other structure of the prior will matter for the ridge regression estimator.

Thus, we get the final from of the free entropy for ridge regression

$$\begin{aligned} &\lim_{\beta\to\infty} \lim_{d\to\infty} \frac{1}{d\beta} \log \mathbb{E}_{x,\xi,w_0}[\mathcal{Z}] \to \\ &\qquad \mathrm{extr}_{\Sigma,m,q,\hat{\Sigma},\hat{m},\hat{q}} \left[\frac{q\hat{\Sigma}}{2} - \frac{\Sigma\hat{q}}{2} - m\hat{m} + \frac{\hat{m}^2 Q_* + \hat{q}}{2(\hat{\Sigma} + \lambda)} - \alpha \frac{Q^\star - 2m + q + \Delta}{1 + 2\Sigma} \right] \end{aligned} \tag{76}$$

3. **(Recap)** Show that the state equations obtained deriving w.r.t. $\hat{\Sigma}, \hat{m}, \hat{q}$ are given by

$$\begin{aligned} \Sigma &= \frac{1}{\hat{\Sigma} + \lambda} \\ m &= \frac{\hat{m} Q^\star}{\hat{\Sigma} + \lambda} \\ q &= \frac{\hat{m}^2 Q^\star + \hat{q}}{(\hat{\Sigma} + \lambda)^2} \end{aligned} \tag{77}$$

while the other three are the same as in (68).

One just computes the gradient equal to zero condition for the free entropy.

4. It turns out that the state equations can be solved explicitly, obtaining the following expression for the test error as a function of $Q^\star, \lambda, \Delta, \alpha$:

$$e_{\text{test}}(Q^\star, \Delta, \lambda, \alpha) = Q^\star \left[\frac{1-\alpha-\Delta}{2} + \frac{2+2\Delta+2\alpha(\alpha+\Delta-2)+\lambda(1+\alpha+\Delta)}{2\sqrt{(2+\lambda)^2+4\alpha(\lambda-2)+4\alpha^2}} \right] \tag{78}$$

where

$$C = 4(\alpha+1)\lambda + 4(\alpha-1)^2 + \lambda^2 \tag{79}$$

Ugly, but explicit! Plot the test error as a function of $\alpha \in (0,3)$ for $Q^\star = \Delta = 1$ and several value of the regularization (down to $\lambda = 10^{-3}$). What do you observe? What happens at $\alpha = 1$ for $\lambda \ll 1$?

Here is the plot: We see that as the regularization gets lowered, the test error around $\alpha \approx 1$

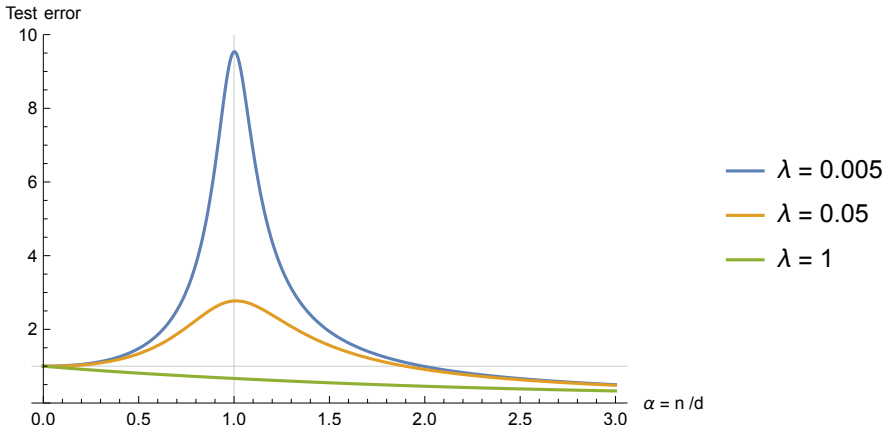

Figure 1: Test error of ridge regression with $Q^\star = \Delta = 1$

degrades visibly. This is a typical case of overfitting. When $\lambda \ll 1$, the ridge regression problem becomes just a regression problem (no regularization), and its global minimum is given by any solution of the linear system $y_\mu = w_\star^\top x_\mu/\sqrt{d}$ for $\mu = 1, \dots, n$ (actually, at leading order in $\lambda \ll 1$, one will have that the global minimum is the solution of the linear system with smallest L2 norm due to the regularization). Thus, the problem at $\lambda \ll 1$ will have $\mathcal{R} = O(\lambda)$ for $\alpha < 1$, as the solution fits the dataset, and $\mathcal{R} = O(1)$ for $\alpha > 1$, as the solution cannot fit perfectly anymore the dataset. At $\alpha = 1$, there is only one solution of the linear system, which is noisy ($\Delta = 1$)! Thus, the solution of the linear system of equation will fit perfectly the label noise, leading to a degradation in test error (as the label noise gives no information on $w^\star$). Only at much larger values of $\alpha > n/d$ the test error decreases again. Larger regularization λ prevents this effect!

5. Show that for $Q^\star = 1$, $\Delta = 0$ and $\lambda \to 0^+$ the test error reduces to $e_{\text{test}} = \max(0, 1-\alpha)$. What happens at $\alpha = 1$?

The test error for $Q^\star = 1$ and $\Delta = \lambda = 0$ gives

$$e_{\text{test}} = \frac{1}{2}\left(1 - \alpha + |1-\alpha|\right) \tag{80}$$

giving the result. At $\alpha = 1$, i.e. $n = d$, the global minimum of the regression problem is just the solution of the linear system $y_\mu = w_\star^\top x_\mu/\sqrt{d}$ for $\mu = 1, \dots, n$ (as $\lambda \to 0^+$). For

$n = d$, there is a unique solution (with high probability the matrix being inverted to solve the linear system is full rank as it is given by i.i.d. entries), so that $\hat{w} = w^\star$ (as there is no label noise). Thus at $\alpha = 1$ we have perfect recovery of the hidden weights, hence perfect generalization, as expected from linear algebra. What replicas gives us additionally is the whole test error curve for $0 < \alpha < 1$, as well as the non-trivial generalization for $\Delta, \lambda > 0$.

6. Fix $\Delta, \alpha > 0$ and a generic prior P_* with $Q* = 1$. What is the value of the regularization that minimizes the test error?

   We know that for a specific prior, i.e. $P_*(w) = N(w, 0, 1)$ and $Q^* = 1$, ridge regression is Bayes optimal for $\lambda = 2\Delta$ (see lecture, it comes from the fact that the posterior and the Gibbs measure are then the same, and are Gaussian). This implies that for $\lambda = 2\Delta$ the test error is always the lowest for this specific prior by Bayes optimality. Now, we remark that for any prior with $Q^* = 1$, ridge regression performs the same, and thus $\lambda = 2\Delta$ will be optimal among all choices of λ (but not necessarily matching the Bayes optimal error for that specific prior!).

12.6 Noiseless sparse priors and LASSO

We now consider a sparse Gaussian prior $P_\star(w) = (1-\epsilon)\delta(w) + \epsilon N(w, 0, 1)$ for sparsity $\epsilon \in (0, 1)$ and noiseless data $\Delta = 0$. For both the BO and ERM cases, the free entropy and the state equations will not be solvable analytically, and one would need to resort to numerical simulations. We will not do that, but given that we will discuss these estimators in Lecture 13, let's at least see how to reduce (46) and (67) for this specific case.

1. Apply the results above to show that the BO free entropy for this problem reads

$$\begin{aligned}\Phi_{\rm BO} = \mathrm{extr}_{q,\hat{q}}\Bigg[&- \frac{q\hat{q}}{2} - \frac{\alpha}{2}\log\left(Q^\star - q\right) \\ &+ \int Dz \int dw_0 P_*(w_0) \log \int \mathrm{d}w P_*(w) e^{-\frac{\hat{q}}{2}w^2 + (\hat{q}w_0 + z\sqrt{\hat{q}})w} \Bigg]\end{aligned} \tag{81}$$

   We then could use the explicit form of $P^\star$ to fully solve the integrals in w and w_0, but the result is not super clean, thus we avoid deriving this.

   Just set $\Delta = 0$ in (46).

2. Apply the results above to show that the ERM free entropy for this problem reads

$$\begin{aligned}\Phi_{\rm ERM} = \mathrm{extr}_{\Sigma,m,q,\hat{\Sigma},\hat{m},\hat{q}}\Bigg[&\frac{q\hat{\Sigma}}{2} - \frac{\Sigma\hat{q}}{2} - m\hat{m} - \alpha\frac{Q^\star - 2m + q}{1 + 2\Sigma} \\ &- \int Dz \int dw_0 P_*(w_0) \min_w \left[\lambda|w| + \frac{\hat{\Sigma}}{2}w^2 - (\hat{m}w_0 + z\sqrt{\hat{q}})w\right]\Bigg].\end{aligned} \tag{82}$$

   Additionally, argue that the minimization problem $\min_w$ in the ERM free entropy achieves its minimum at

$$w = \hat{\Sigma}^{-1}\phi_\lambda(\hat{m}w_0 + z\sqrt{\hat{q}})\,, \tag{83}$$

   where

$$\phi_a(x) = \begin{cases} x + a & \text{if } x < a \\ 0 & \text{if } |x| \le a \\ x - a & \text{if } x > a \end{cases}. \tag{84}$$

Just set $\Delta = 0$ and $\rho(w) = |w|$ in (67). For the minimization problem, the stationarity condition reads

$$\lambda \text{sign}(w) + \hat{\Sigma} w - (\hat{m} w_0 + z\sqrt{\hat{q}}) = 0 \implies w + \frac{\lambda}{\hat{\Sigma}} \text{sign}(w) = \frac{\hat{m} w_0 + z\sqrt{\hat{q}}}{\hat{\Sigma}} \tag{85}$$

translating to the following conditions

$$\begin{cases} w = \frac{\hat{m} w_0 + z\sqrt{\hat{q}} - \lambda}{\hat{\Sigma}} \\ w > 0 \end{cases} \quad \text{and} \quad \begin{cases} w = \frac{\hat{m} w_0 + z\sqrt{\hat{q}} + \lambda}{\hat{\Sigma}} \\ w < 0 \end{cases} \tag{86}$$

and if none of these conditions are satisfied, it means that the minimum is at the angular point of the original function, i.e. $w = 0$. This gives

$$w = \hat{\Sigma}^{-1} \phi_\lambda (\hat{m} w_0 + z\sqrt{\hat{q}}) \tag{87}$$

where

$$\phi_a(x) = \begin{cases} x + a & \text{if } x < a \\ 0 & \text{if } |x| \leq a \\ x - a & \text{if } x > a \end{cases} \tag{88}$$