

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

$$y_\mu = \frac{w^* \top x_\mu}{\sqrt{d}} + \sqrt{\Delta} \xi_\mu \quad (1)$$

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

1. **(Recap)** Write the Bayes optimal posterior distribution for w^* 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?

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

$$\hat{f}(x; w) = \frac{\hat{w} \top x}{\sqrt{d}} \quad (2)$$

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

$$\hat{w} = \arg \min_w \mathcal{R}(w), \quad \mathcal{R}(w) = \sum_{\mu} \left[y_{\mu} - \frac{w^{\top} x_{\mu}}{\sqrt{d}} \right]^2 + \lambda \sum_{i=1}^d \rho(w_i) \quad (3)$$

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 (3) has a single global minimum. Argue that, for a given dataset and set of weights w^* , we have

$$\frac{1}{d} \min_w \mathcal{R}(w) = - \lim_{\beta \rightarrow +\infty} \frac{1}{d} \partial_{\beta} \log \mathcal{Z} = - \lim_{\beta \rightarrow +\infty} \frac{1}{d\beta} \log \mathcal{Z} \quad (4)$$

and

$$\lim_{\beta \rightarrow +\infty} \frac{e^{-\beta \mathcal{R}(w)}}{\mathcal{Z}} = \delta(w - \arg \min_{w'} \mathcal{R}(w')) \quad (5)$$

where

$$\mathcal{Z} = \int dw e^{-\beta \mathcal{R}(w)}. \quad (6)$$

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

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

$$p_{\beta}(w) \propto e^{-\beta \mathcal{R}(w)} \quad (7)$$

for the risk in (3). For which value of β and $\rho(w)$ does the Gibbs distribution coincide with the posterior for the inference problem for w^* ?

Thus, up to here we learn that:

- We can study the minimization problem (3) 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 $\beta \rightarrow \infty$ 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 (7) 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

$$\mathbb{E}_{x, \xi, w^\star}[\mathcal{Z}^r] = \mathbb{E}_{x, \xi, w^\star} \left[\int dw_a e^{-\beta \sum_{a=1}^r \left[\sum_{\mu=1}^n [(w_0 - w_a) x_\mu^\top / \sqrt{d} + \sqrt{\Delta} \xi_\mu]^2 + \lambda \sum_{i=1}^d \rho(w_a^i) \right]} \right] \quad (8)$$

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 $a = 1, \dots, r$.

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 $a = 0, 1, \dots, n$

$$h_\mu^a = \frac{w_a^\top x_\mu}{\sqrt{d}}. \quad (9)$$

Assuming all w_a are fixed, compute the distribution over the preactivations $\{h_\mu^a\}_{\substack{\mu=1, \dots, n \\ a=0, \dots, r}}$ 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.

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

$$h_\mu^a \sim \mathcal{N}(0, q_{ab}(w) \delta_{\mu\nu}), \quad q_{ab}(w) = \frac{w_a^\top w_b}{d}. \quad (10)$$

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

$$\begin{aligned} & \mathbb{E}_{x, \xi, w_0}[\mathcal{Z}^r] \\ &= \mathbb{E}_{w_0} \int dw_a e^{-\beta \lambda \sum_{a=1}^r \sum_{i=1}^d \rho(w_a^i)} \mathbb{E}_{\substack{h_\mu^a \sim \mathcal{N}(0, q_{ab}(w_a) \delta_{\mu\nu}) \\ \xi_\mu \sim \mathcal{N}(0, 1)}} \prod_{\mu=1}^n e^{-\beta \sum_{a=1}^r (h_\mu^0 - h_\mu^a + \sqrt{\Delta} \xi_\mu)^2} \end{aligned} \quad (11)$$

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

$$\begin{aligned} & \mathbb{E}_{x, \xi, w_0}[\mathcal{Z}^r] \\ &= \mathbb{E}_{w_0} \int dw_a e^{-\beta \lambda \sum_{a=1}^r \sum_{i=1}^d \rho(w_a^i)} \left[\mathbb{E}_{\substack{h \sim \mathcal{N}(0, q_{ab}(w_a)) \\ \xi \sim \mathcal{N}(0, 1)}} e^{-\beta \sum_{a=1}^r (h^0 - h^a + \sqrt{\Delta} \xi)^2} \right]^n. \end{aligned} \quad (12)$$

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

$$\mathbb{E}_{x,\xi,w_0}[\mathcal{Z}^r] = \int dq_{ab} d\hat{q}_{ab} \exp \left\{ -d \sum_{0 \leq a \leq b \leq n} q_{ab} \hat{q}_{ab} + d \log I_{\text{reg}}(\hat{q}_{ab}) + n \log I_{\text{loss}}(q_{ab}) \right\}. \quad (13)$$

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

$$I_{\text{loss}}(\hat{q}) = \mathbb{E}_{\substack{h \sim \mathcal{N}(0, q_{ab}) \\ \xi \sim \mathcal{N}(0, 1)}} e^{-\beta \sum_{a=1}^r (h^0 - h^a + \sqrt{\Delta} \xi)^2}, \quad (14)$$

and

$$I_{\text{reg}}(q) = \int_{\mathbb{R}} dw_0 P_*(w_0) \int_{\mathbb{R}} dw_a e^{-\beta \lambda \sum_{a=1}^r \rho(w_a) - \sum_{0 \leq a \leq b \leq r} \hat{q}_{ab} w_a w_b}, \quad (15)$$

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

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 $r \rightarrow 0$. 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^*	\hat{q}_{00}	\hat{Q}^*
$q_{0a} \ (0 < a < r)$	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	$\hat{q}_{aa} \ (0 < a < r)$	$-\hat{Q}/2$

(16)

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 $r \ll 1$

$$\sum_{0 \leq a \leq b \leq r} q_{ab} \hat{q}_{ab} = Q^* \hat{Q}^* + r \left[m \hat{m} - \frac{Q \hat{Q}}{2} - \frac{q \hat{q}}{2} \right] + \mathcal{O}(r^2). \quad (17)$$

7. Notice that in (17) 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 $\hat{Q}^* = 0$ at the saddle point. What is the value of Q^* at the saddle point?

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

$$\log I_{\text{loss}}(q_{ab}) = -\frac{r}{2} \left[\log(1 + \beta(Q - q)) + \frac{2\beta(Q^* - 2m + q + \Delta)}{1 + 2\beta(Q - q)} \right] + \mathcal{O}(r^2) \quad (18)$$

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 $\hat{Q}^* = 0$, the Hubbard-Stratonovich trick to decouple replicas, and the limit $r \rightarrow 0$ to get, at leading order

$$\log(I_{\text{reg}}) = r \int Dz \int dw_0 P_*(w_0) \log \int dw e^{-\beta \lambda \rho(w) - \frac{\hat{Q} + \hat{q}}{2} w^2 + (\hat{m} w_0 + z \sqrt{\hat{q}}) w} + \mathcal{O}(r^2), \quad (19)$$

where Dz is integration over a standard Gaussian measure.

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

$$\begin{aligned} \frac{1}{d} \log \mathbb{E}_{x, \xi, w_0} [Z] &\rightarrow \Phi \\ \text{where } \Phi &= \text{extr}_{Q, m, q, \hat{Q}, \hat{m}, \hat{q}} \left[\frac{Q\hat{Q}}{2} - m\hat{m} + \frac{q\hat{q}}{2} \right. \\ &\quad \left. + \int Dz \int dw_0 P_*(w_0) \log \int dw e^{-\beta \lambda \rho(w) - \frac{\hat{Q} + \hat{q}}{2} w^2 + (\hat{m} w_0 + z \sqrt{\hat{q}}) w} \right. \\ &\quad \left. - \frac{n}{d} \left[\frac{1}{2} \log(1 + \beta(Q - q)) + \frac{\beta(Q^* - 2m + q + \Delta)}{1 + 2\beta(Q - q)} \right] \right] \end{aligned} \quad (20)$$

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

The expression (20) 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^* . In this case, we saw above that the posterior corresponds to the choices $\beta = 1/(2\Delta)$ and

$$\rho(w) = -\frac{1}{\beta \lambda} \log P_*(w). \quad (21)$$

1. Show that in this case

$$\begin{aligned} \Phi &= \text{extr}_{q, \hat{q}} \left[-\frac{q\hat{q}}{2} - \frac{\alpha}{2} \log(2\Delta + Q^* - q) \right. \\ &\quad \left. + \int Dz \int dw_0 P_*(w_0) \log \int dw P_*(w) e^{-\frac{\hat{q}}{2} w^2 + (\hat{q} w_0 + z \sqrt{\hat{q}}) w} \right] \end{aligned} \quad (22)$$

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

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

$$\int Dz \int dw_0 P_*(w_0) \log \int dw P_*(w) e^{-\frac{\hat{q}}{2} w^2 + (\hat{q} w_0 + z \sqrt{\hat{q}}) w} \quad (23)$$

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

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

$$\hat{q} = \frac{\alpha}{2\Delta + Q^* - q} \quad (24)$$

and

$$q = \int Dz \int dw_0 P_*(w_0) \frac{\int dw P_*(w) e^{-\frac{\hat{q}}{2}w^2 + (\hat{q}w_0 + z\sqrt{\hat{q}})w} w w_0}{\int dw P_*(w) e^{-\frac{\hat{q}}{2}w^2 + (\hat{q}w_0 + z\sqrt{\hat{q}})w}} \quad (25)$$

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

$$e_{\text{est}} = \frac{1}{d} \|w^* - \hat{w}_{\text{BO}}\|^2 \quad (26)$$

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

5. Argue that the generalization error

$$e_{\text{gen}} = \mathbb{E}_{x_{\text{test}}} \left\| \frac{w_*^\top x_{\text{test}}}{\sqrt{d}} - \frac{\hat{w}_{\text{BO}}^\top x_{\text{test}}}{\sqrt{d}} \right\|^2 \quad (27)$$

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

Thus, we find a complete characterization of the Bayes optimal estimator and 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 $\beta \rightarrow \infty$ limit of (20).

We consider the following change of variable for large β . We substitute (Q, \hat{Q}) with $(\Sigma, \hat{\Sigma})$ using

$$Q = q + \Sigma/\beta \quad \text{and} \quad \hat{q} + \hat{Q} = \beta \hat{\Sigma} \quad (28)$$

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 $\hat{q} \rightarrow \beta^2 \hat{q}$ and $\hat{m} \rightarrow \beta \hat{m}$ (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

$$\begin{aligned} \frac{1}{d\beta} \log \mathbb{E}_{x, \xi, w_0} [\mathcal{Z}] &\rightarrow \text{extr}_{\Sigma, m, q, \hat{\Sigma}, \hat{m}, \hat{q}} \left[\frac{q\hat{\Sigma}}{2} - \frac{\Sigma\hat{q}}{2} - m\hat{m} \right. \\ &\quad \left. + \frac{1}{\beta} \int Dz \int dw_0 P_*(w_0) \log \int dw e^{\beta[-\lambda\rho(w) - \frac{\Sigma}{2}w^2 + (\hat{m}w_0 + z\sqrt{\hat{q}})w]} \right. \\ &\quad \left. - \frac{n}{d} \frac{Q^* - 2m + q + \Delta}{1 + 2\Sigma} \right] \end{aligned} \quad (29)$$

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

$$\begin{aligned} & \frac{1}{\beta} \int Dz \int dw_0 P_*(w_0) \log \int dw e^{\beta[-\lambda\rho(w) - \frac{\hat{\Sigma}}{2}w^2 + (\hat{m}w_0 + z\sqrt{\hat{q}})w]} \\ &= - \int Dz \int dw_0 P_*(w_0) \min_w \left[\lambda\rho(w) + \frac{\hat{\Sigma}}{2}w^2 - (\hat{m}w_0 + z\sqrt{\hat{q}})w \right] \end{aligned} \quad (30)$$

Notice that (30) is the average value of the minimum of a scalar minimization problem with risk

$$f(w) = \lambda\rho(w) + \frac{\hat{\Sigma}}{2}w^2 - (\hat{m}w_0 + z\sqrt{\hat{q}})w \quad (31)$$

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 (23). 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

$$e_{gen} = \mathbb{E}_{x_{test}} \left\| \frac{w_*^\top}{\sqrt{d}} - \frac{\hat{w}^\top}{\sqrt{d}} \right\|^2 \quad (32)$$

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^*, m, q . Here x_{test} is a new test input drawn from $N(0, \mathbb{I}_d)$.

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

$$\Sigma = \frac{1}{d} \text{Tr} \mathcal{H}[\mathcal{R}](\hat{w})^{-1} \quad (33)$$

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 $\Sigma \rightarrow \infty$, then it means that the loss develops flat directions around its global minimum, i.e. the global minimum is not unique anymore.

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

$$\begin{aligned} \lim_{\beta \rightarrow \infty} \lim_{d \rightarrow \infty} \frac{1}{d\beta} \log \mathbb{E}_{x, \xi, w_0} [\mathcal{Z}] &\rightarrow \text{extr}_{\Sigma, m, q, \hat{\Sigma}, \hat{m}, \hat{q}} \left[\frac{q\hat{\Sigma}}{2} - \frac{\Sigma\hat{q}}{2} - m\hat{m} - \alpha \frac{Q^* - 2m + q + \Delta}{1 + 2\Sigma} \right. \\ &\quad \left. - \int Dz \int dw_0 P_*(w_0) \min_w \left[\lambda\rho(w) + \frac{\hat{\Sigma}}{2}w^2 - (\hat{m}w_0 + z\sqrt{\hat{q}})w \right] \right]. \end{aligned} \quad (34)$$

At this point we could derive directly the state equations for generic ρ and P_* , 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:

$$\begin{aligned} \hat{q} &= \frac{4\alpha(\Delta - 2m + q + Q^*)}{(2\Sigma + 1)^2} \\ \hat{\Sigma} &= \frac{2\alpha}{2\Sigma + 1} \\ \hat{m} &= \frac{2\alpha}{2\Sigma + 1} \end{aligned} \quad (35)$$

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_* .

1. Show that

$$\min_w \left[\lambda \rho(w) + \frac{\hat{\Sigma}}{2} w^2 - (\hat{m} w_0 + z \sqrt{\hat{q}}) w \right] = -\frac{(\hat{m} w_0 + z \sqrt{\hat{q}})^2}{2(\hat{\Sigma} + \lambda)} \quad (36)$$

2. Show that

$$- \int Dz \int dw_0 P_*(w_0) \min_w \left[\lambda \rho(w) + \frac{\hat{\Sigma}}{2} w^2 - (\hat{m} w_0 + z \sqrt{\hat{q}}) w \right] = \frac{\hat{m}^2 Q_* + \hat{q}}{2(\hat{\Sigma} + \lambda)} \quad (37)$$

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

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

$$\lim_{\beta \rightarrow \infty} \lim_{d \rightarrow \infty} \frac{1}{d\beta} \log \mathbb{E}_{x, \xi, w_0} [\mathcal{Z}] \rightarrow \text{extr}_{\Sigma, m, q, \hat{\Sigma}, \hat{m}, \hat{q}} \left[\frac{q \hat{\Sigma}}{2} - \frac{\Sigma \hat{q}}{2} - m \hat{m} + \frac{\hat{m}^2 Q_* + \hat{q}}{2(\hat{\Sigma} + \lambda)} - \alpha \frac{Q^* - 2m + q + \Delta}{1 + 2\Sigma} \right] \quad (38)$$

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

$$\begin{aligned} \Sigma &= \frac{1}{\hat{\Sigma} + \lambda} \\ m &= \frac{\hat{m} Q^*}{\hat{\Sigma} + \lambda} \\ q &= \frac{\hat{m}^2 Q^* + \hat{q}}{(\hat{\Sigma} + \lambda)^2} \end{aligned} \quad (39)$$

while the other three are the same as in (35).

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

$$e_{\text{test}}(Q^*, \Delta, \lambda, \alpha) = Q^* \left[\frac{1 - \alpha - \Delta}{2} + \frac{2 + 2\Delta + 2\alpha(\alpha + \Delta - 2) + \lambda(1 + \alpha + \Delta)}{2\sqrt{(2 + \lambda)^2 + 4\alpha(\lambda - 2) + 4\alpha^2}} \right] \quad (40)$$

where

$$C = 4(\alpha + 1)\lambda + 4(\alpha - 1)^2 + \lambda^2 \quad (41)$$

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

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

12.6 Noiseless sparse priors and LASSO

We now consider a sparse Gaussian prior $P_*(w) = (1 - \epsilon)\delta(w) + \epsilon N(w, 0, 1)$ 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 (22) and (34) for this specific case.

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

$$\begin{aligned} \Phi_{\text{BO}} = \text{extr}_{q, \hat{q}} & \left[-\frac{q\hat{q}}{2} - \frac{\alpha}{2} \log(Q^* - q) \right. \\ & \left. + \int Dz \int dw_0 P_*(w_0) \log \int dw P_*(w) e^{-\frac{\hat{q}}{2}w^2 + (\hat{q}w_0 + z\sqrt{\hat{q}})w} \right] \end{aligned} \quad (42)$$

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

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

$$\begin{aligned} \Phi_{\text{ERM}} = \text{extr}_{\Sigma, m, q, \hat{\Sigma}, \hat{m}, \hat{q}} & \left[\frac{q\hat{\Sigma}}{2} - \frac{\Sigma\hat{q}}{2} - m\hat{m} - \alpha \frac{Q^* - 2m + q}{1 + 2\Sigma} \right. \\ & \left. - \int Dz \int dw_0 P_*(w_0) \min_w \left[\lambda|w| + \frac{\hat{\Sigma}}{2}w^2 - (\hat{m}w_0 + z\sqrt{\hat{q}})w \right] \right]. \end{aligned} \quad (43)$$

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

$$w = \hat{\Sigma}^{-1} \phi_\lambda(\hat{m}w_0 + z\sqrt{\hat{q}}), \quad (44)$$

where

$$\phi_a(x) = \begin{cases} x + a & \text{if } x < -a \\ 0 & \text{if } |x| \leq a \\ x - a & \text{if } x > a \end{cases}. \quad (45)$$