Statistical Physics of Computation - Exercises

Emanuele Troiani, Vittorio Erba, Yizhou Xu

September 2024

Week 7

7.1 Empirical risk minimisation for the binary perceptron

A simple and algorithm to train a classifier that separates two sets of labelled points is to choose a student vector minimizing the risk (or energy, or cost) $E(J)$:

$$\hat{J} = \text{argmin}_{\|J\|^2=N} E(J) = \text{argmin}_{\|J\|^2=N} \sum_{\mu=1}^{P} v\left(\frac{\sigma^\mu}{\sqrt{N}} J^T \xi^\mu - \kappa\right) . \tag{1}$$

The training dataset is composed of $P = \alpha N$ points $\{\xi^\mu\}_{\mu=1}^P$ of dimension N, each with a label $\sigma^\mu = \pm 1$ taken at random.

The random labels can be modified into teacher-student generated labels as we saw in the previous lectures, allowing us to discuss the generalization and estimation tasks. This modifies the replica computation making it a bit more cumbersome, adding more order parameters, so we don't do it here. Instead, we focus our attention on the Empirical Risk Minimisation task, i.e. the fact that our student will be trained by minimizing a cost function.

We will take $v(x) = 0$ for $x \geq 0$, and $v(x) > 0$ for $x < 0$ in all following points.

1. Intuitively, why is it a good idea to pick $v(x)$ to be convex and differentiable when on $x < 0$?

   Convexity implies uniqueness of the minimum, so that any algorithms minimizing the function will land in the global minimum, and not in a spurious local minimum, as long as we are in the UNSAT phase (see question below).
   Differentiability implies that one can minimize the cost function efficiently by gradient descent iterates. By efficiently we mean that each iterate takes $O(N)$ computations to be evaluated, and that convergence happens exponentially fast with properly chosen learning rate.

2. We know that in the high dimensional limit the space of solutions of this classification problem has a SAT/UNSAT transition at a critical value α_C. How many global minima of $E(J)$ do you expect to find in the SAT region? And how many in the UNSAT region?

   In the SAT region there are infinitely many solutions, as we have a fraction of the sphere which satisfies the constraints. When all constraints are satisfied, i.e. all points correctly labeled, all items in the sum (1) can be made zero by picking one of the perfect classifiers, hence there are infinitely many degenerate and connected global minima.
   In the UNSAT region, by convexity there is only one global minimum, with non-zero value of the energy function.

3. What do you expect the ground state energy density to be in the SAT and UNSAT phases?

   In the SAT phase, the constraints can be all satisfied. Thus, the ground state energy density will be zero. In the UNSAT phase, we expect that the ground state energy density will be positive, as we expect a finite fraction of constraints to be violated.

We will now study the empirical risk minimisation of the energy function $E(J)$ for a generic per-point cost $v(x) = 0$ for $x \geq 0$, and $v(x) > 0$ for $x < 0$.

4. Write the Gibbs distribution associated to the cost function $E(J)$, thought of as an energy function. Is the energy function properly scaled for large N?

   The Gibbs distribution is

$$p(J) = \frac{1}{Z} \exp(-\beta E(J)) \,. \tag{2}$$

   It is properly normalized, as each term in the sum of (1) is a scalar quantity as long as v does not depend on N or P, and the sum makes the energy of order $O(P) = O(N)$ in the proportional scaling $P = \alpha N$.

5. Write the averaged replicated partition function for the Gibbs measure of point 2. In which limit should we consider this quantity to access properties of the global minimum?

   The partition function is the same as in graded homework 1, ex 5.2.1, with an extra limit to low temperatures

$$\mathbb{E}_{\xi,\sigma,J^*} Z^n = \mathbb{E}_\xi \prod_{a=0}^{n} \int dJ^a \, \Pi_{\mu=1}^{p} \exp\left(-\beta v\left(\frac{1}{\sqrt{N}} J_a^T \xi^\mu - \kappa\right)\right) . \tag{3}$$

   Notice that we used the usual trick to set $\sigma^\mu = 1$.
   The limit we are interested in is the $\beta \to \infty$ limit (i.e. low temperature), as in that limit the Gibbs measure concentrates on the ground state of the energy function.

6. Argue without many computations that the averaged free energy is

$$f(\beta) = -\frac{1}{2\beta}\left[\log(1-q) + \frac{q}{1-q}\right] - \frac{\alpha}{\beta}\int dz \, \frac{e^{-\frac{z^2}{2q}}}{\sqrt{2\pi q}} \log\left[\int dr \, \frac{e^{-\frac{r^2}{2(1-q)}}}{\sqrt{2\pi(1-q)}} e^{-\beta v(r+z-\kappa)}\right] \tag{4}$$

   to be optimized over q.

   As argued in the graded homework 1, exercise 5.2.1., one can follow the computation for the Gardner's volume with the substitution $\theta(\cdot) \to \exp(-\beta v(\cdot))$.

Recall that the free energy is defined as $f = -\phi/\beta$, where ϕ is the free entropy $\phi = N^{-1} \log Z$. The two quantities are interchangeable at any finite β, but in the limit $\beta \to \infty$ the free entropy $\phi = O(-\beta)$, so it is useful to divide by $-\beta$ to access a finite quantity. Also, recall that at the extremiser q^* the free entropy has the decomposition $\phi = s - \beta e$, where e and s are the average values of the energy density and of the entropy density, hence $f = e - s/\beta$. The last equation shows us that $\lim_{\beta\to\infty} f(\beta) = e_{\text{ground state}}$, which in our problem is the training error, i.e. the value of the cost function $E(J)/N$ at its global minimum.

In the following, assume that $v(x)$ is convex for $x < 0$.

7. We already know that there is going to be a critical value of α_C such that the energy is zero for $\alpha < \alpha_C$. Argue that $q \to 1$ if we take the limit $\beta \to \infty$ in the UNSAT region.

   By uniqueness of the global minimum in the UNSAT region, as β grows, the level sets of the energy function shrink, so that the overlap of two samples of the Gibbs distribution become closer and closer between each other, hence $q \to 1$.

8. Show that

$$\lim_{\beta\to\infty} \frac{-1}{\beta} \frac{e^{-\frac{z^2}{2q}}}{\sqrt{2\pi q}} \log\left[\int dr \, \frac{e^{-\frac{r^2}{2(1-q)}}}{\sqrt{2\pi(1-q)}} e^{-\beta v(r+z-\kappa)}\right] = \frac{e^{-\frac{z^2}{2}}}{\sqrt{2\pi}} \min_r \left[\frac{r^2}{2x} + v(r+z-\kappa)\right] . \tag{5}$$

   Use the ansatz $q = 1 - \chi/\beta + \dots$ with $\chi > 0$.

   First, we have

$$\frac{e^{-\frac{z^2}{2q}}}{\sqrt{2\pi q}} \approx \frac{e^{-\frac{z^2}{2}}}{\sqrt{2\pi}} , \qquad \frac{e^{-\frac{r^2}{2(1-q)}}}{\sqrt{2\pi(1-q)}} \approx \frac{e^{-\frac{\beta r^2}{2\chi}}}{\sqrt{2\pi\chi/\beta}} . \tag{6}$$

   The first remark is important to argue that there is no hidden dependency on β in the integration measure for z. We are ready to compute the limit by using a saddle-point in β

$$-\lim_{\beta\to\infty} \frac{1}{\beta} \log\left[\int dr \, \frac{e^{-\frac{r^2}{2(1-q)}}}{\sqrt{2\pi(1-q)}} e^{-\beta v(r+z-\kappa)}\right] = \min_r \left[\frac{r^2}{2\chi} + v(r+z-\kappa)\right] \tag{7}$$

Thus, we found that the free energy simplifies quite drastically in the large β limit, as one fo the Gaussian integrals can be substituted by a minimisation problem, usually easier to perform analytically or numerically. This is a recurrent point: when considering empirical risk minimisation problems, the $\beta \to \infty$ limit will lead to some simplifications.

We now focus on the case $v(x) = 0$ for $x > 0$ and $v(x) = x^2/2$ for $x < 0$. This per-point cost function is differentiable, and convex for $x < 0$, so it is a good candidate of an actual cost function we could use in practice to algorithmically solve the problem.

9. Solve the minimisation problem over r in the case $v(x) = 0$ for $x > 0$ and $v(x) = x^2/2$ for $x < 0$. You should obtain

$$\min_r \left[\frac{r^2}{2\chi} + v(r+z-\kappa)\right] = \theta(\kappa - z) \frac{(z-\kappa)^2}{2(1+x)} . \tag{8}$$

   Fix $z \in \mathbb{R}$ and $\chi > 0$. The derivative of

$$h(r) = \begin{cases} \frac{r^2}{2\chi} + \frac{1}{2}(r+z-\kappa)^2 & r+z < \kappa \\ \frac{r^2}{2\chi} & r+z > \kappa \end{cases} . \tag{9}$$

   equals

$$h'(r) = \begin{cases} r(1+\chi^{-1}) + z - \kappa & r+z < \kappa \\ r\chi^{-1} & r+z > \kappa \end{cases} . \tag{10}$$

   Setting the derivative to zero gives us

$$\begin{cases} r = -\frac{z-\kappa}{1+1/\chi} & r+z < \kappa \\ r = 0 & r+z > \kappa \end{cases} . \tag{11}$$

Thus, all $r > \kappa - z$ are stationary points where the function $h(z)$ achieves the global minimum zero. When instead $r + z < 0$, we need to check that

$$z - \frac{z-\kappa}{1+1/\chi} < \kappa \iff \frac{z+\chi\kappa}{1+\chi} < 0 \iff z < \kappa \tag{12}$$

Thus, the minimum of $h(z)$ equals

$$\begin{cases} \frac{(z-\kappa)^2}{2(1+\chi)} & z < \kappa \\ 0 & z > \kappa \end{cases} . \tag{13}$$

This is equivalent to the expression in the solution

10. Show that the leading order for large β of the energy density in the UNSAT phase is given by

$$e(\chi) = -\frac{1}{2\chi} + \frac{\alpha}{2\alpha_c(1+\chi)} \tag{14}$$

where α_c is the SAT/UNSAT threshold we computed in the previous lectures, i.e.

$$\frac{1}{\alpha_c} = \int_{-\infty}^{\kappa} dz \, \frac{e^{-\frac{z^2}{2}}}{\sqrt{2\pi}} (z-\kappa)^2 . \tag{15}$$

We first want to look at the large β behavior of the first piece of the free energy

$$\frac{1}{2\beta}\left[\log(1-q) + \frac{q}{1-q}\right] \approx \frac{1}{2\beta}\left[\log\chi/\beta + \frac{\beta-\chi}{\chi}\right] \approx \frac{1}{2\chi} \tag{16}$$

The rest is given by the solution of the last point, giving

$$-\frac{1}{2\chi} + \frac{\alpha}{2(1+\chi)} \int_{-\infty}^{\kappa} dz \, \frac{e^{-\frac{z^2}{2}}}{\sqrt{2\pi}} (z-\kappa)^2 = -\frac{1}{2\chi} + \frac{\alpha}{2\alpha_c(1+\chi)} \tag{17}$$

11. Find the state equation for χ in the UNSAT phase, and use it to find the energy as a function of α both in the SAT and in the UNSAT phase

$$e(\alpha) = \frac{1}{2}\left(\sqrt{\frac{\alpha}{\alpha_c}} - 1\right)^2 \theta(\alpha - \alpha_c) . \tag{18}$$

In the SAT phase the energy density is zero, as there exists configurations that satisfy all constraints (i.e. have zero energy) and the energy function is non-negative. In the UNSAT phase, the state equation for χ is found by estremising $e(\chi)$, that is we want a value of χ^* such that $e'(\chi^*) = 0$,

$$e'(\chi^*) = 0 \iff \frac{1}{\chi^2} - \frac{\alpha/\alpha_c}{(1+\chi)^2} = 0 \iff \frac{1+\chi}{\chi} = \sqrt{\frac{\alpha}{\alpha_c}} \iff \chi_* = \left(\sqrt{\frac{\alpha}{\alpha_c}} - 1\right)^{-1} . \tag{19}$$

Notice that χ diverges when $\alpha \to \alpha_c$, from which we see that this solution is valid only for $\alpha > \alpha_c$ (we know that the UNSAT phase is for large α).

Remark: This gives us an independent way of computing the SAT/UNSAT transition. Suppose we did not know that α_c was the SAT/UNSAT threshold. Then, the fact that we know that our scaling ansatz $q \to 1$ breaks at the SAT/UNSAT threshold, along with the observation that χ diverges at α_c, allow us to deduce that α_c is indeed the SAT/UNSAT transition.
Finally, the energy is just $e(\chi^*)$, which can be easily reduced in the form given above.
The theta comes from noticing that the transition between the SAT and UNSAT phases is at $\alpha = \alpha_c$, allowing to write the energy density in the two phases compactly.

Thus, we showed that

- the large β limit leads in general to simplifications in the computations.
- the critical α_c delimiting the SAT and UNSAT phases arises naturally in the empirical risk minimisation computation as the point at which the overlap converges to 1 in the large β limit.

Analogous considerations would arise in the teacher-student version of the classification problem, with the added difficulty of having to deal with at least 3 order parameters.