# Statistical Physics of Computation - Exercises

Emanuele Troiani, Vittorio Erba, Yizhou Xu

September 2024

## Week 8

### 8.1 The average error of the BO estimator and of the Gibbs estimator

Consider a generic inference problem where you generate a hidden signal $X_* \sim P_0$, where $P_0$ is the prior distribution, and observe it through a noisy channel, obtaining the data/observation $Y \sim P_{\text{out}}$ ("out" stands for "output channel"). Think of $X$ as a vector with $N$ components, $X \in \mathbb{R}^N$, and $Y$ as a vector with $P$ components, $Y \in \mathbb{R}^P$. Consider the posterior distribution $P_{\text{posterior}}(X|Y) \propto P_0(X)P_{\text{out}}(Y|X)$, and suppose that you are in the Bayes Optimal (BO) setting, i.e. you know both $P_0$ and $P_{\text{out}}$, so you know the posterior.

We want to find expressions for the errors of the BO estimator w.r.t. to mean square error (MSE) loss (i.e. an expression for the MMSE) and of the Gibbs estimator, both as functions of the overlap order parameters associated to the posterior distribution. We define

$$\begin{aligned} Q_* &= \frac{1}{N}\mathbb{E}_{X\sim P_0}||X||^2 \\ Q &= \frac{1}{N}\mathbb{E}_Y\mathbb{E}_{X\sim P_{\text{posterior}}(\cdot|Y)}||X||^2 \\ m &= \frac{1}{N}\mathbb{E}_{Y,X_*}\mathbb{E}_{X\sim P_{\text{posterior}}(\cdot|Y)}X_*^T X \\ q &= \frac{1}{N}\mathbb{E}_Y\mathbb{E}_{X_1\sim P_{\text{posterior}}(\cdot|Y)}\mathbb{E}_{X_2\sim P_{\text{posterior}}(\cdot|Y)}X_1^T X_2 \end{aligned} \tag{1}$$

$Q_*$ is the self-overlap (norm) of the signal, $Q$ is the self-overlap (norm) of a sample from the posterior, $m$ is the overlap of a sample from the posterior with the hidden signal, and $q$ is the overlap between two independent samples from the posterior, and all of these quantities are averaged over the observation $Y$ defining the posterior.

To be very explicit, the averages are defined as

$$\begin{aligned} \mathbb{E}_{X\sim P_0} f(X) &= \int dX\, f(X)P_0(X)\,, \\ \mathbb{E}_Y f(Y) &= \int dY\, dX_*\, f(Y)P_{\text{out}}(Y|X_*)P_0(X_*)\,, \\ \mathbb{E}_{Y,X_*} f(Y,X_*) &= \int dY\, dX_*\, f(Y,X_*)P_{\text{out}}(Y|X_*)P_0(X_*)\,, \\ \mathbb{E}_{X\sim P_{\text{posterior}}(\cdot|Y)} f(X) &= \int dX\, f(X)P_{\text{posterior}}(X|Y)\,. \end{aligned} \tag{2}$$

**These definitions (both the order parameters and the averages) should become natural to you, even if they are not right now. So convince yourself that they make sense, and learn them.**

1. Show that for any estimator $\hat{X} : Y \to \hat{X}(Y)$, i.e. a function taking as input the observation, and outputting some estimate of the hidden signal that generated the observation, one has

$$\frac{1}{N}\mathbb{E}_{Y,X_*}||X_* - \hat{X}(Y)||^2 = \frac{1}{N}\mathbb{E}_{X^*\sim P_0}||X_*||^2 - \frac{2}{N}\mathbb{E}_{Y,X_*}X_*^T\hat{X}(Y) + \frac{1}{N}\mathbb{E}_Y||\hat{X}(Y)||^2\,. \quad (3)$$

   Use that $||z||^2 = z^T z$ and expand the product. In the first term there is no dependence on $Y$, and in the last one no dependence on $X_*$, so the two associated averages drop.

2. Show that $Q = Q_*$.

   We can apply Nishimori's identity, telling us that a sample from the posterior is equivalent to the hidden signal under average over the observation. Then we have

$$\begin{aligned} Q &= \frac{1}{N}\mathbb{E}_Y\mathbb{E}_{X\sim P_{\text{posterior}}(\cdot|Y)}||X||^2 \\ &= \frac{1}{N}\mathbb{E}_{Y,X_*}||X_*||^2 \\ &= \frac{1}{N}\mathbb{E}_{X_*}||X_*||^2 \\ &= Q_*\,. \end{aligned} \quad (4)$$

We start by considering the BO estimator. In class, we saw that the BO estimator w.r.t. the MSE is the average of the posterior, i.e.

$$\hat{X}_{\text{BO, MSE}}(Y) = \mathbb{E}_{X\sim P_{\text{posterior}}(\cdot|Y)}X\,. \quad (5)$$

3. Show that

$$\frac{1}{N}\mathbb{E}_{Y,X_*}||X_* - \hat{X}_{\text{BO, MSE}}(Y)||^2 = Q - 2m + q\,. \quad (6)$$

   Take the result of point 1, and recognize that the first term is directly the definition of $Q_* = Q$. For the second term, we have

$$\begin{aligned} \frac{1}{N}\mathbb{E}_{Y,X_*}X_*^T\hat{X}_{\text{BO, MSE}}(Y) &= \frac{1}{N}\mathbb{E}_{Y,X_*}X_*^T\mathbb{E}_{X\sim P_{\text{posterior}}(\cdot|Y)}X \\ &= \frac{1}{N}\mathbb{E}_{Y,X_*}\mathbb{E}_{X\sim P_{\text{posterior}}(\cdot|Y)}X_*^T X \\ &= m\,. \end{aligned} \quad (7)$$

   For the third term we have

$$\begin{aligned} \frac{1}{N}\mathbb{E}_Y||\hat{X}_{\text{BO, MSE}}(Y)||^2 &= \frac{1}{N}\mathbb{E}_Y\hat{X}_{\text{BO, MSE}}(Y)^T\hat{X}_{\text{BO, MSE}}(Y) \\ &= \frac{1}{N}\mathbb{E}_Y\mathbb{E}_{X_1\sim P_{\text{posterior}}(\cdot|Y)}\mathbb{E}_{X_2\sim P_{\text{posterior}}(\cdot|Y)}X_1^T X_2 \\ &= q\,. \end{aligned} \quad (8)$$

4. Argue finally that

$$\frac{1}{N}\mathbb{E}_{Y,X_*}||X_* - \hat{X}_{\text{BO, MSE}}(Y)||^2 = Q - q\,. \quad (9)$$

By Nishimori's identities, we know that a sample from the posterior and the hidden signal are equivalent under the average over the observation $Y$. Thus

$$\begin{aligned} q &= \frac{1}{N}\mathbb{E}_Y \mathbb{E}_{X_1 \sim P_{\text{posterior}}(\cdot|Y)} \mathbb{E}_{X_2 \sim P_{\text{posterior}}(\cdot|Y)} X_1^T X_2 \\ &= \frac{1}{N}\mathbb{E}_{Y,X_*} \mathbb{E}_{X \sim P_{\text{posterior}}(\cdot|Y)} X_*^T X \\ &= m\,, \end{aligned} \tag{10}$$

from which the result follows.

We now instead consider the Gibbs estimator. Recall that in the context of the classification task, we defined the Gibbs estimator as a uniform sample from the solution space of the task. In the context of inference, the Gibbs estimator is a sample from the posterior distribution (in the classification task, the posterior was exactly the uniform measure over the solution space, hence the use of the same name). In this case, we consider the error of the Gibbs estimator on average, as it is a random estimator.

4. Show that
$$\frac{1}{N}\mathbb{E}_{Y,X_*} \mathbb{E}_{X \sim P_{\text{posterior}}(\cdot|Y)} ||X_* - X||^2 = 2(Q - q)\,. \tag{11}$$
Notice the factor 2 difference with the BO estimator.

   Take the result of point 1, and recognize that the first term is directly the definition of $Q_* = Q$. For the second term, we have
$$\frac{1}{N}\mathbb{E}_{Y,X_*} \mathbb{E}_{X \sim P_{\text{posterior}}(\cdot|Y)} X_*^T X = m = q\,, \tag{12}$$
   where we used the answer of point 3. For the third term we have
$$\begin{aligned} \frac{1}{N}\mathbb{E}_{Y,X_*} \mathbb{E}_{X \sim P_{\text{posterior}}(\cdot|Y)} X^T X &= \frac{1}{N}\mathbb{E}_{Y,X_*} \mathbb{E}_{X \sim P_{\text{posterior}}(\cdot|Y)} ||X||^2 \\ &= \frac{1}{N}\mathbb{E}_{Y} \mathbb{E}_{X \sim P_{\text{posterior}}(\cdot|Y)} ||X||^2 \\ &= Q\,. \end{aligned} \tag{13}$$

5. Argue that the Gibbs sampler is on average worse than the BO estimator.

   By definition of Bayes Optimality, we already know that the BO estimator is equal or better than any other estimator, so in particular of any sample from the posterior, and hence of their average. More explicitly, we can notice that the average MSE of the Gibbs estimator, minus the MSE of the BO estimator equals
$$2(Q - q) - (Q - q) = Q - q \geq 0 \tag{14}$$
   as the scalar product of two vectors $q$ is always smaller than the product of their norms $Q$.

6. When is the Gibbs sampler (on average) as effective as the BO estimator?

   The two errors are equal only if $Q = q$, i.e. if the overlap is maximal. This is equivalent by Nishimori's identity to
$$\frac{m}{Q} = \frac{m}{\sqrt{Q}\sqrt{Q_*}} = 1\,. \tag{15}$$

That is, the the angle between a sample from the posterior and the hidden signal is zero, on average over the observation $Y$ and over the sample from the posterior $X$. In other words $X = X_*$. Thus, the Gibbs sampler achieves BO performance exactly when the BO estimator perfectly recovers the hidden signal. In all other cases, $Q \neq q$ and the Gibbs sampler is worse by a factor 2.

## 8.2 Bayesian learning of a scalar variable

We now consider the following scalar inference problem. We generate a hidden signal $x^* \in \mathbb{R}$ from a prior distribution $\mathrm{P}_0(x)$. We then observe only a noisy version of the signal

$$y = x^* + \sqrt{\Delta} z \tag{16}$$

where $z$ is an independent Gaussian variable with mean zero and variance 1, and $\Delta > 0$ plays the role of a noise-to-signal ratio. Given $y$, we want to Bayes-optimally estimate the signal $x^*$. We first need to set up the Bayesian machinery.

1. Write the output channel distribution $P_{\text{out}}(y|x)$, i.e. the probability of observing $y$ given a certain signal $x$.

   We observe that, conditioned on $x$, the observation $y$ is a Gaussian variable with mean $x$ and variance $\Delta$, so we have
   $$P_{\text{out}}(y|x) = N(y; x; \Delta)\,. \tag{17}$$

2. Use Bayes theorem to show that the posterior distribution, i.e. the probability that the signal is $x$ given our observation $y$, satisfies
   $$P_{\text{posterior}}(x|y) = P_0(x)\frac{e^{-\frac{(y-x)^2}{2\Delta}}}{Z} \tag{18}$$
   where $Z$ is the normalization factor.

   Recall that by Bayes theorem
   $$P_{\text{posterior}}(x|y) = \frac{1}{Z}P_{\text{out}}(y|x)P_0(x) = \frac{1}{Z}e^{-\frac{(y-x)^2}{2\Delta}}P_0(x)\,. \tag{19}$$

3. We need to find a good estimator $\hat{x}$ for our signal. We will use the Bayes Optimal estimator with respect to the MSE, i.e. the mean of the posterior distribution. Argue that we have:
   $$\hat{x}(y) = \frac{\int \mathrm{d}x\, x\, P_0(x) e^{-\frac{(y-x)^2}{2\Delta}}}{\int \mathrm{d}x P_0(x) e^{-\frac{(y-x)^2}{2\Delta}}} \tag{20}$$

   This is just the definition of the average of the posterior, after explicitly evaluating the normalization constant $Z$.

4. Suppose now that $P_0$ is a standard Gaussian. Show that

$$\hat{x}(y) = \frac{y}{1+\Delta}\,. \tag{21}$$

First we have

$$P_0(x)e^{-\frac{(y-x)^2}{2\Delta}} = \frac{1}{\sqrt{2\pi}}\exp\left\{-\frac{(y-x)^2}{2\Delta} - \frac{x^2}{2}\right\} = \frac{1}{\sqrt{2\pi}}\exp\left\{-\frac{x^2}{2}\left(\frac{1+\Delta}{\Delta}\right) + \frac{xy}{\Delta} - \frac{y^2}{2\Delta}\right\} \tag{22}$$

From which we derive by Gaussian integration that

$$\int \mathrm{d}x P_0(x)e^{-\frac{(y-x)^2}{2\Delta}} = \frac{1}{\sqrt{2\pi}}\sqrt{\frac{\Delta}{1+\Delta}}e^{-\frac{y^2}{2(1+\Delta)}} \tag{23}$$

and

$$\int \mathrm{d}x\, x\, P_0(x)e^{-\frac{(y-x)^2}{2\Delta}} = \frac{1}{\sqrt{2\pi}}\frac{y}{1+\Delta}\sqrt{\frac{\Delta}{1+\Delta}}e^{-\frac{y^2}{2(1+\Delta)}} \tag{24}$$

Putting the two results together gives the result.

5. Show that the MSE that this estimator achieves is

$$\mathbb{E}_{x^*,y}\left[(\hat{x}(y) - x^*)^2\right] = \frac{\Delta}{1+\Delta} \tag{25}$$

We just compute it:

$$\begin{aligned}
&\mathbb{E}_{x^*,y}\left[(\hat{x}(y) - x^*)^2\right] = \\
&\mathbb{E}_{x^*,y}\left[\left(\frac{y}{1+\Delta} - x^*\right)^2\right] = \\
&\mathbb{E}_{x^*,z}\left[\left(\frac{x^* + \sqrt{\Delta}z}{1+\Delta} - x^*\right)^2\right] = \\
&\mathbb{E}_{x^*,z}\left[\left(\frac{\sqrt{\Delta}}{1+\Delta}z - x^*\frac{\Delta}{1+\Delta}\right)^2\right] = \\
&\frac{\Delta}{(1+\Delta)^2}\mathbb{E}_{x^*,z}\left[\left(z - x^*\sqrt{\Delta}\right)^2\right] = \\
&\frac{\Delta}{(1+\Delta)^2}(1+\Delta) = \\
&\frac{\Delta}{1+\Delta}\,.
\end{aligned} \tag{26}$$

The only non-trivial passage is to express the observation $y$ as $y = x^* + \sqrt{\Delta}z$, moving from an average over $y$ to an average over the noise $z$.

6. If $\Delta \to \infty$ there is no information about the signal in the observation, as it is fully erased by the large amount of noise. In such a case, given that we know the prior, we may just sample a candidate estimate for the lost signal from the prior and hope that it achieves a good performance. What is, on average, the MSE of a sample from the prior? What is the

MSE of the BO estimator when $\Delta \to \infty$? Discuss what is the BO estimator doing in this limit.

We have

$$\mathbb{E}_{x^*,x\sim P_0}||x-x^*||^2 = 2\mathbb{E}_{x\sim P_0}||x||^2 = 2 \tag{27}$$

while the MSE of the BO estimator is just 1. Thus, the BO estimator is doing something smarter than sampling the prior $P_0$. Indeed, by inspecting the expression for the BO estimator, we see that for $\Delta \to \infty$, $\hat{x}(y) \to 0$. Apparently, it is less wrong to just return the zero estimator than to sample the prior! This reminds us that Bayes optimality is always defined w.r.t. a given error measure, and thus the associated BO estimator will pick up the quirks of that estimator.

7. We now change the prior $P_0$. Find the BO estimator w.r.t. to the MSE error for the case in which $x^* = 1$ with probability $p$ and zero otherwise. You should get:

$$\hat{x}(y) = \frac{1}{1+\exp\left\{\frac{1-2y}{2\Delta}\right\}\frac{1-p}{p}} \tag{28}$$

We can write the prior on $x$ as

$$P_0(x) = (1-p)\delta(x) + p\delta(x-1) \tag{29}$$

We then have

$$\int \mathrm{d}x P_0(x) e^{-\frac{(y-x)^2}{2\Delta}} = (1-p)e^{-\frac{y^2}{2\Delta}} + pe^{-\frac{(y-1)^2}{2\Delta}} \tag{30}$$

and

$$\int \mathrm{d}x x P_0(x) e^{-\frac{(y-x)^2}{2\Delta}} = pe^{-\frac{(y-1)^2}{2\Delta}} \tag{31}$$

Taking the ratio gives the estimator.

Notice that in the last point we obtained an estimator which gives us as an estimate a continuous value in $[0,1]$, instead of just telling us whether the signal was $x_* = 1$ or $x_* = 0$. This is expected, as we asked for the estimate that minimizes the MSE, without specifying that it should be related to the support of the prior $P_0$. If we wanted an estimator respecting the constraint that $x \in \{0,1\}$, we could have used the maximum-a-posteriori estimator (MAP), i.e.

$$\hat{x}_{\mathrm{MAP}}(y) = \mathrm{argmax}_{x\in\{0,1\}} P_{\mathrm{posterior}}(x|y) \tag{32}$$

giving as an estimate the most likely signal to have generated the observation $y$.

8. Compute the MAP estimator for the prior of point 7.

We have

$$\begin{aligned}\hat{x}_{\mathrm{MAP}}(y) &= \mathrm{argmax}_{x\in\{0,1\}} P_{\mathrm{posterior}}(x|y) \\ &= \begin{cases} 0 & \text{if} \quad P_{\mathrm{posterior}}(0|y) > P_{\mathrm{posterior}}(1|y) \\ \{0,1\} & \text{if} \quad P_{\mathrm{posterior}}(0|y) = P_{\mathrm{posterior}}(1|y) \\ 1 & \text{if} \quad P_{\mathrm{posterior}}(0|y) < P_{\mathrm{posterior}}(1|y) \end{cases}\end{aligned} \tag{33}$$

Now the condition $P_{\text{posterior}}(0|y) > P_{\text{posterior}}(1|y)$ can be rewritten as

$$\begin{aligned} P_0(0)\frac{e^{-\frac{y^2}{2\Delta}}}{Z} &> P_0(1)\frac{e^{-\frac{(y-1)^2}{2\Delta}}}{Z} \\ (1-p)e^{-\frac{y^2}{2\Delta}} &> pe^{-\frac{(y-1)^2}{2\Delta}} \\ \frac{1-p}{p} &> e^{\frac{y^2}{2\Delta}-\frac{(y-1)^2}{2\Delta}} \\ 2\Delta\log\frac{1-p}{p} &> y^2-(y-1)^2 \\ 2\Delta\log\frac{1-p}{p} &> 2y-1 \\ \frac{1}{2}+\Delta\log\frac{1-p}{p} &> y\,. \end{aligned} \tag{34}$$

Thus, the MAP estimator equals

$$\begin{aligned} \hat{x}_{\text{MAP}}(y) &= \text{argmax}_{x\in\{0,1\}} P_{\text{posterior}}(x|y) \\ &= \begin{cases} 0 & \text{if} \quad y < \frac{1}{2}+\Delta\log\frac{1-p}{p} \\ \{0,1\} & \text{if} \quad y = \frac{1}{2}+\Delta\log\frac{1-p}{p} \\ 1 & \text{if} \quad y > \frac{1}{2}+\Delta\log\frac{1-p}{p} \end{cases} \end{aligned} \tag{35}$$

We see that as a function of the noise-to-signal ratio $\Delta$ and the probability of the hidden signal being zero or one $p$, the MAP estimator is just a threshold function over $y$, rounding to 0 if $y$ is sufficiently small/negative, and rounding to 1 if $y$ is sufficiently large, as one may intuitively expect.