PD Dr. Torsten Enßlin
Vishal Johnson, Ming Huang

Exercise sheet 3

Exercise 3 - 1

Attention!
Consider a scenario in which Alice wants to communicate some information to Bob. The information concerns a continuous quantity s and resulting outcomes. Alice's knowledge state is $\mathrm{I_A}$ which gives rise to the probability distribution $p(s) = \mathcal{P}(s|\mathrm{I_A})$. Similarly, Bob's knowledge state is $\mathrm{I_B}$ which gives rise to the probability distribution $q(s) = \mathcal{P}(s|\mathrm{I_B})$. The aim is to find an appropriate loss function to quantify "distances" between the information contents $\mathrm{I_A}$ and $\mathrm{I_B}$ (and therefore probability distributions $p(s)$ and $q(s)$).

We wish the loss function to satisfy certain properties:

1. We assume that the loss function obeys *locality* - that the loss corresponding to outcome s_* must only depend on the probability $q(s_*)$ induced in Bob about event s_*. Therefore, the loss function takes the form $\mathcal{L}(s, q(s))$.
2. Furthermore, we assume that the loss function obeys *properness* - that the loss function takes on its minimal value when $q(s) = p(s)$.

Alice calculates the average loss L by averaging the above loss function over her probability distribution:

$$\mathrm{L} = \langle \mathcal{L}(s, q(s)) \rangle_{p(s)}. \tag{1}$$

a) Calculate the sensitivity of the loss L to the probability $q(s_*)$ that Bob's knowledge state assigns to signal value s_*. In order to do this, calculate the functional derivate $\delta \mathrm{L}/\delta q(s_*)$. (2 points)
Hint: Try to solve the problem for a discrete signal $\{s_i\}$ first.

b) For a proper loss function $\mathcal{L}$, the functional derivate evaluated at $q = p$ must be minimal. Therefore $\delta \mathrm{L}/\delta q(s_*)|_{q=p} = 0$. Calculate the functional relation satisfied by $\mathcal{L}$ under this condition. Remember that $q(s)$ must be normalised. (2 points)

c) Find out the loss function which satisfies the above functional relation. (1 point)

d) There are scenarios in which certain signal values s deserve more attention than other values of the signal. Using a weighting scheme $w(s)$ (with $w(s) > 0$) to quantify the attention deserved by different signal values one can define an attention function

$$\mathcal{A}(s) = \frac{w(s)\mathcal{P}(s)}{\int w(s)\mathcal{P}(s)ds} \tag{2}$$

corresponding to a probability distribution $\mathcal{P}(s)$.

Consider $\mathcal{A}_q(s)$ and $\mathcal{A}_p(s)$, the attention functions corresponding to the probability distributions $q(s)$ and $p(s)$. Alice wants to ensure that her communication influences Bob to place maximal attention $\mathcal{A}_q(s_*)$ on what actually happens (s_*) in the end.

First, show that properness in probability and properness in attention are equivalent. (2 points)

e) Repeat the above analysis for attention functions $\mathcal{A}_q(s)$ and $\mathcal{A}_p(s)$. Note that expectation averages are still to be performed over probabilities. (4 points)

For a more detailed discussion please check out the paper https://arxiv.org/abs/2307.11423.

Exercise 3 - 2

Your knowledge I about a quantity $\mathbf{x} = (x_1, \dots x_n)^{\mathrm{T}} \in \mathbb{R}^n$ is characterized by being separable,

i.e. $\mathcal{P}(\mathbf{x}|I) = \prod_{i=1}^{n} \mathcal{P}(x_i|I)$ and by being isotropic, i.e. $\mathcal{P}(\mathbf{x}|I) = \mathcal{P}(\mathbf{O}\,\mathbf{x}|I)$ for all orthonormal transformations with $\mathbf{O}^\dagger = \mathbf{O}^{-1}$.

Using the definition of the information Hamiltonian:

$$\mathcal{H}(\mathbf{x}|I) = \sum_{i=1}^{n} \mathcal{H}(x_i|I) \equiv \sum_{i=1}^{n} h_i(x_i) \ , \tag{3}$$

we may write

$$\prod_{i=1}^{n} \mathcal{P}(x_i|I) = \prod_{i=1}^{n} \mathcal{P}((\mathbf{O}\,\mathbf{x})_i|I) \tag{4}$$

$$\sum_{i=1}^{n} \mathcal{H}(x_i|I) = \sum_{i=1}^{n} \mathcal{H}((\mathbf{O}\,\mathbf{x})_i|I) \tag{5}$$

$$\sum_{i=1}^{n} h_i(x_i) = \sum_{i=1}^{n} h_i((\mathbf{O}\,\mathbf{x})_i) \ . \tag{6}$$

a) Let $\mathbf{x} = r\, e^{(i)}$ be parallel to the i-th unit vector of $\mathbb{R}^n$, $e^{(i)}_j = \delta_{ij}$, and $\mathbf{O}$ such that $\mathbf{O}e^{(i)} = e^{(j)}$ with $i \neq j$. Show that $h_i(r) = h_j(r) + \mathrm{const(i,j)}$. (3 points)

b) Use the result of a) to show that $\mathcal{H}(\mathbf{x}|I) = \sum_i h(x_i) + const.$ with $h(x_i) \equiv h_0(x_i)$. (1 point)

c) Given a general $\mathbf{x} \in \mathbb{R}^n$ with length $r = |\mathbf{x}| = \sqrt{\sum_i x_i^2}$, we can choose $\mathbf{O}$ such that $\mathbf{O}\mathbf{x} = (r, 0, \ldots, 0)$. Show that in this case

$$\frac{h'(x_j)}{x_j} = \frac{h'(r)}{r} \quad \forall j \ , \tag{7}$$

where h' denotes the derivative of h. (3 points)

d) Given the result of c), derive the general functional form of h. (2 points)

e) Finally, derive an expression for $\mathcal{P}(\mathbf{x}|I)$. (2 points)

Exercise 3 - 3

Let $P(s) = \mathcal{G}(s, S)$ with $s = (s_1, ..., s_n)^t$ be a real multivariate zero-centered Gaussian with Covariance $\langle ss^t \rangle = S$. We would like to fit another Gaussian distribution $P'(s) = \mathcal{G}(s, S')$ to it that has a diagonal covariance matrix $S'_{ij} = \delta_{ij}\sigma_i$. Here δ_{ij} denotes the Kronecker delta.

a) What is the optimal approximating Gaussian $P'(s)$ to $P(s)$, as parameterized by σ, obtained through minimizing the loss

$$\sigma = \arg\min_\sigma \mathrm{KL}(P(s), P'(s)) \ ?$$

(2 points)

b) What is the least updating fit of $P'(s)$ to $P(s)$, as obtained through

$$\sigma = \arg\min_\sigma \mathrm{KL}(P'(s), P(s)) \ ?$$

(2 points)

c) Let

$$S = \frac{1}{2} \begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & 0.2 \end{pmatrix} \begin{pmatrix} 1 & 1 \\ -1 & 1 \end{pmatrix} .$$

Visualize the Gaussian $\mathcal{G}(s, S)$ as well as the two different ways to fit a diagonal Gaussian to it introduced above. To visualize them, use a computer plotting samples of each of the three distributions. You can draw a sample from a multivariate Gaussian distribution by applying the square root of its covariance matrix to a white noise sample. (optional)

Exercise 3 - 4

Assume the measurement of a signal s, which yields the data d, leads to

$$\mathcal{P}(s) = \mathcal{G}(s, S)$$
$$\mathcal{P}(s|d) = \mathcal{G}(s - m, D).$$

a) Calculate the amount of information in terms of entropy one gains via the measurement (3 points).

b) Now assume that your signal prior in the above measurement was flat. How much information in terms of entropy does one gain via the measurement? Explain the result (1 point).

This exercise sheet will be discussed during the exercises.
Group 01, Wednesday 18:00 - 20:00, Theresienstr. 37, A 449,
Group 02, Thursday, 10:00 - 12:00, Theresienstr. 37, A 249,

https://wwwmpa.mpa-garching.mpg.de/ ensslin/lectures/lectures.html