6th homework set, Due May 30

1. (1p.) Determine $D(P\|Q)$

   (a) when P and Q are binomial distributions

   $$P(k) = \binom{n}{k} p^k (1-p)^{n-k},\ Q(k) = \binom{n}{k} q^k (1-q)^{n-k},\ k \in \{0, \dots, n\}. \tag{1}$$

   (b) when P and Q are Poisson distributions

   $$P(k) = \frac{\lambda^k}{k!} e^{-\lambda},\ Q(k) = \frac{\mu^k}{k!} e^{-\mu},\ k \in \{0, 1, \dots\}. \tag{2}$$

   Hint: You may use that the expected value $\sum P(k)k$ of the binomial or Poisson distribution is known from probability theory.

2. (2.5p.) Find the (ideal) codelength for encoding the sequence 2221111222 using (i) the coding process assigned to the class of i.i.d. processes (see page 481 of the lecture notes) and (ii) the coding process assigned to the class of (first order) Markov sources (see page 485 of the lecture notes). Interpret the result. (Actual codes are not required, only the codelengths)

3. (2.5p+2.5p) (Universal source coding with fixed length codes of rate R)

   (a) Let $A_n \subset \mathcal{X}^n$ be the union of all type classes T_P^n with $H(P) \leqslant R$. Show that

   $$\frac{1}{n} \log |A_n| \to R \tag{3}$$

   and that

   $$Q^n(A_n) \to 1 \tag{4}$$

   for every distribution Q on $\mathcal{X}$ with $H(Q) < R$. More exactly, prove that

   $$Q^n(A_n^c) = 2^{-n \cdot e_Q(R) + o(n)},\ e_Q(R) = \min_{P: H(P) \geqslant R} D(P||Q). \tag{5}$$

   Hint: Use the results of Section 2 of the lecture notes.

   Remark: For each n take a **fixed length** n-code C_n as follows. Let $l_n = \lceil \log |A_n| \rceil$ and choose an injective mapping $f_n : A_n \to \{0\} \times \{0,1\}^{l_n}$. Then let $C_n : \mathcal{X}^n \to \{0,1\}^{l_n+1}$ be the code which encodes the element of A_n using f_n and encode all the other sequences of $\mathcal{X}^n$ with the all 1's sequence of length $l+1$. According to this exercise the code sequence $\{C_n, n = 1, 2, \dots\}$ uses asymptotically R bits per source symbol, moroever, it is true simultaneously for each Q with $H(Q) < R$ that the Q^n probability that the source sequence can not be recovered from its code tends to 0 with exponent $e_Q(R)$ as $n \to \infty$.

   (b) Show that the error exponent $e_Q(R)$ is best possible for every Q, i.e., for arbitrary $\tilde{A}_n$ satisfying (3) above, always

   $$\liminf_{n \to \infty} \frac{1}{n} \log Q^n(\tilde{A}_n^c) \geqslant -e_Q(R). \tag{6}$$

   Hint: Check that for any fixed $\varepsilon > 0$, there is n_o such that $\frac{|T_p^n \cap \tilde{A}_n^c|}{|T_p^n|} \geqslant \frac{1}{2}$ for each n-type P with $H(P) > R + \varepsilon$ if $n \geqslant n_o$.

4. (3.5p.) (Glimpse into the general theory)

Read the general definition of the KL divergence below!

More generally, if P and Q are probability measures over a set $\mathcal{X}$, and P is absolutely continuous with respect to Q, then the Kullback–Leibler divergence from Q to P is defined as

$$D_{\mathrm{KL}}(P \parallel Q) = \int_{\mathcal{X}} \log\left(\frac{dP}{dQ}\right) dP,$$

If P is not absolutely continous with respect to Q, then the KL divergence is defined to be infinity

General definition of the KL divergence from wikipedia

where $\frac{dP}{dQ}$ is the Radon–Nikodym derivative of P with respect to Q, and provided the expression on the right-hand side exists. Equivalently (by the chain rule), this can be written as

$$D_{\mathrm{KL}}(P \parallel Q) = \int_{\mathcal{X}} \log\left(\frac{dP}{dQ}\right) \frac{dP}{dQ}\, dQ,$$

which is the entropy of Q relative to P. Continuing in this case, if μ is any measure on $\mathcal{X}$ for which $p = \frac{dP}{d\mu}$ and $q = \frac{dQ}{d\mu}$ exist (meaning that p and q are absolutely continuous with respect to μ), then the Kullback–Leibler divergence from Q to P is given as

$$D_{\mathrm{KL}}(P \parallel Q) = \int_{\mathcal{X}} p \log\left(\frac{p}{q}\right) d\mu.$$

The general KL divergence is also always nonnegative, and equals 0 iff the measures P and Q are equal

Let $\mathcal{E}$ be an exponential family of distributions P_θ, $\theta = (\theta_1, \ldots, \theta_k) \in \mathcal{H}$ on an arbitrary measurable space $(\mathcal{X}, \mathcal{F})$, defined by

$$\frac{\mathrm{d}P_\theta}{\mathrm{d}\mu}(x) = \mathrm{e}^{-\Lambda(\theta)+\sum_{j=1}^{k}\theta_j f_j(x)}, \quad \Lambda(\theta) = \ln \int \mathrm{e}^{\sum_{j=1}^{k}\theta_j f_j(x)}\mu(\mathrm{d}x),$$

where $f_1, \ldots, f_k$ are given (measurable) functions on $(\mathcal{X}, \mathcal{F})$ and $\mathcal{H} = \{\theta : \Lambda(\theta) < \infty\}$.

Given a sample $x = (x_1, \ldots, x_n)$ the MLE is the distribution P_{ML} that maximizes the normalized log-likelihood function $l(\theta) = \frac{1}{n} \cdot \log \prod_{i=1}^{n} \frac{\mathrm{d}P_\theta}{\mathrm{d}\mu}(x_i)$.

(a) Show that $l(\theta) = \sum_{j=1}^{k} \alpha_j \theta_j - \Lambda(\theta)$, where $\alpha_j = \frac{1}{n}\sum_{i=1}^{n} f_j(x_i)$!

(b) Prove that P_{ML} equals the reversed I-projection (onto $\mathcal{E}$) of any element P of the linear family $\mathcal{L} = \{P : \int f_j(x)P(\mathrm{d}x) = \alpha_j\}$ for which there exists $P_{\theta^0} \in \mathcal{E}$ with $\mathrm{D}(P\|P_{\theta^0}) < \infty$!

Hint: Prove that for P and P_{θ^0} above, and for all $P_\theta \in \mathcal{E}$

$$\mathrm{D}(P\|P_\theta) = \mathrm{D}(P\|P_{\theta^0}) + \sum_{j=1}^{k} \alpha_j \theta_j^0 - \Lambda(\theta^0) - \sum_{j=1}^{k} \alpha_j \theta_j + \Lambda(\theta).$$

(c) Prove that if $\mathcal{L} \cap \mathcal{E} \neq \varnothing$, then P_{ML} equals the single element of $\mathcal{L} \cap \mathcal{E}$! You can use without proof that if $P^* \in \mathcal{L} \cap \mathcal{E}$ then the Pythagorean identity holds, i.e.,

$$\mathrm{D}(P\|P_\theta) = \mathrm{D}(P\|P^*) + \mathrm{D}(P^*\|P_\theta), \; P \in \mathcal{L}, \; P_\theta \in \mathcal{E}.$$

Remark: In the non-discrete case the MLE can exist even if $\mathcal{L} \cap \mathcal{E} = \varnothing$.

As you can see, you can get two extra points for the second homework! Have a nice work!