4th homework set, Due May 12

1. (1p.) Let a distribution Q, a linear family $\mathcal{L}$ and a convex closed subfamily $\mathcal{L}' \subseteq \mathcal{L}$ on the finite set A be given with $S(Q) = S(\mathcal{L}) = A$. Denote by P^* the I-projection of Q onto $\mathcal{L}$. Prove that the I-projections of Q and of P^* onto $\mathcal{L}'$ are the same.

2. (3p.) Let $\mathcal{E}$ be the family of binomial distributions with $n = 5$ and $p \in (0,1)$, i.e.,

$$\mathcal{E} = \{\mathbb{P} : \mathbb{P}(a) = \binom{5}{a} p^a (1-p)^{5-a}, a \in \{0,1,2,3,4,5\}, \text{ for some } p \in (0,1).\} \tag{1}$$

   (a) Show that $\mathcal{E}$ is an exponential family!

   (b) We observe 200 independent drawing from an unknown distribution on $A = \{0,1,2,3,4,5\}$. The type of the observed sample $\hat{\mathbb{P}}_{200} = (\hat{\mathbb{P}}_{200}(0), \hat{\mathbb{P}}_{200}(1), \hat{\mathbb{P}}_{200}(2), \hat{\mathbb{P}}_{200}(3), \hat{\mathbb{P}}_{200}(4), \hat{\mathbb{P}}_{200}(5))$ equals

$$(0.05, 0.34, 0.31, 0.24, 0.04, 0.02). \tag{2}$$

   Test the null hypothesis with type 1 error probability $\varepsilon = 0.05$ that the sample come from a distirbution in $\mathcal{E}$ using the method outlined in Remark 4.2. of the lecture notes, i.e., calculate $\frac{400}{\log e} D(\hat{\mathbb{P}}_{200}||\mathbb{P}^*)$ and check whether it exceeds the 0.95 quantile of the chi-squared distribution with appropriate degree of freedom.

   Hint: Theorem 3.2 is useful for determining $\mathbb{P}^*$.

3. (3p.) Let $\mathcal{L}$ be the linear family of distributions on $\Omega = \{0,\ldots,r_1\} \times \{0,\ldots,r_2\}$ with prescribed marginals $(P(0\cdot),\ldots,P(r_1\cdot))$ and $(P(\cdot 0),\ldots,P(\cdot r_2))$. For any $Q \in \mathcal{P}(\Omega)$ with $S(Q) = \Omega$, the I-projection P^* of Q to $\mathcal{L}$ can be computed via iterative proportional fitting. Show that P^* can be computed also by the iterative algorithm

$$b_0(i,j) = Q(i,j) \tag{3}$$

$$b_{n+1}(i,j) = b_n(i,j)\sqrt{\frac{P(i\cdot)}{b_n(i\cdot)} \cdot \frac{P(\cdot j)}{b_n(\cdot j)}}, \text{ where } b_n(i\cdot) = \sum_j b_n(i,j),\ b_n(\cdot j) = \sum_i b_n(i,j), \tag{4}$$

   i.e., $b_n(i,j) \to P^*(i,j)$ for each $(i,j) \in \Omega$. Let Ξ be the exponential family corresponding to $\mathcal{L}$ (taking for Q the uniform distribution). Characterize the members of Ξ!

   Hint: Apply the theorem on generalized iterative scaling.

4. (3p.) Define distributions $\mathbb{Q}_n$ on the sets $\{0,1\}^n$ recursively, by $\mathbb{Q}_1 = (\frac{1}{2}, \frac{1}{2})$ and

$$\mathbb{Q}_n(x_1^n) = \mathbb{Q}_{n-1}(x_1^{n-1})\frac{\mathrm{N}(x_n|x_1^{n-1}) + 1}{n+1},\ n \geqslant 2,$$

   where $\mathrm{N}(x_n|x_1^{n-1})$ denotes the number of occurrences of the bit x_n in $x_1^{n-1} = x_1 \ldots x_{n-1}$.

   (a) Show that for all $k \in \{0,1,\ldots,n\}$

$$\mathbb{Q}_n(x_1^n) = \frac{1}{(n+1)\binom{n}{k}} \quad \text{if the type of } x_1^n \text{ equals } \left(\frac{k}{n}, \frac{n-k}{n}\right),$$

   and consequently, if the true probability of x_1^n is $\mathbb{P}^n(x_1^n) = \prod_{i=1}^n \mathbb{P}(x_i)$ (for any $\mathbb{P} = (\mathbb{P}(0), \mathbb{P}(1))$), this probability is bounded above by $(n+1)\mathbb{Q}_n(x_1^n)$.

   (b) Draw the conclusion that for any memoryless source with alphabet $\{0,1\}$, if $\mathbb{Q}_n$ is used as coding distribution instead of the true $\mathbb{P}^n$, the loss in average length $\mathrm{D}(\mathbb{P}^n||\mathbb{Q}_n)$ does not exceed $\log(n+1)$.