Problem Set 3

Information and Coding Theory

February 26, 2021 — Clayton Seitz

Problem 0.1. *A single dice is rolled and we gain a dollar if the outcome is 2,3,4,5 and lose a dollar if the outcome is 1 or 6. Find the expected gain and the maximum entropy distribution over the possible outcomes of a roll.*

Solution.

Let Q be the uniform distribution over the dice universe $\chi = \{1, 2, 3, 4, 5, 6\}$ where an outcome of a roll is $x \in \chi$. Furthermore, let $\phi(x)$ be the gain given the outcome of a roll x according the problem definition

$$\phi(x) = \begin{cases} 1 & x \in \{2, 3, 4, 5\} \\ -1 & x \in \{1, 6\} \end{cases}$$

and $\bar{x} \sim Q^n$ be a draw of a sequence of n rolls from the product distribution Q^n. We can then calculate the expected gain over n rolls where $\bar{x} = \{x_1 \ldots x_n\}$ denotes the outcome of a sequence of rolls

$$\begin{aligned} \mathop{\mathbf{E}}_{\bar{x}\sim Q^n} [\phi(\bar{x})] &= \sum_n \left(\mathop{\mathbf{E}}_{x_n \sim Q} [\phi(x_n)] \right) \\ &= n \cdot \left(\sum_{x\in\chi} \phi(x) \cdot q(x) \right) \\ &= n \cdot \left(\frac{1}{6} \sum_{x\in\chi} \phi(x) \right) \\ &= \frac{n}{3} \end{aligned}$$

Now, we would like to find the maximum entropy distribution P^* over support χ such that

$$\mathop{\mathbf{E}}_{\bar{x}\sim(P^*)^n}[\phi(\bar{x})] > \alpha$$

where $\alpha = \frac{n}{3}$. We begin by defining the linear family of distributions that satisfy this constraint on the expected gain

$$\mathcal{L} = \left\{ P : \mathop{\mathbf{E}}_{\bar{x}\sim P^n}[\phi(\bar{x})] = n \cdot \left(\sum_{x\in\chi} \phi(x)\cdot p(x) \right) > \alpha \right\}$$

We would like to find the distribution P^* such that $P^* = \mathbf{Proj}_{\mathcal{L}}(Q)$. To determine P^*, we include two constraints in the Lagrangian weighted by coefficients $\lambda_0, \lambda_1 \in \mathbb{R}$. For a particular distribution $P \in \mathcal{L}$ we have

$$\begin{aligned}\mathbf{\Lambda}(P,\lambda_0,\lambda_1) &= D(P||Q) + \lambda_0 \cdot \left(\sum_x p(x) - 1 \right) + \lambda_1 \cdot (\alpha - \Phi) \\ &= \sum_{x\in\chi} \left(p(x)\log\frac{p(x)}{q(x)} + \lambda_0 \cdot \left(p(x) - \frac{1}{d} \right) + \lambda_1 \cdot \left(\frac{\alpha}{d} - \phi(x) \right) \right)\end{aligned}$$

where we have let $\Phi = \sum_{x\in\chi}\phi(x)\cdot p(x)$ denote the expected gain and defined a linear penalty for deviation away from α that subtracts from the Lagrangian for $\Phi > \alpha$. We find the extremum by setting the gradient of this Lagrangian with respect to $p(x)$ to zero for every $x \in \chi$

$$\log\left(\frac{p^*(x)}{q(x)}\right) + \frac{1}{\ln 2} + \lambda_0 - \lambda_1 \cdot \phi(x) = 0$$

Therefore, the optimal distribution can be found from P by computing the following for each $x \in \chi$

$$p^*(x) = q(x) \cdot 2^{\,\lambda_1\,\cdot\,\phi(x) - \lambda_0}$$

■

Problem 0.2. *Exponential families and maximum entropy*

Solution.

$$\begin{aligned}
H(Q) &= -\sum_{x\sim\chi} Q(x) \log \exp\left\{\lambda_0 + \sum_{i\sim[k]} \lambda_i f_i(x)\right\} \\
&= -\frac{1}{\ln 2} \sum_{x\sim\chi} Q(x) \left\{\lambda_0 + \sum_{i\sim[k]} \lambda_i f_i(x)\right\} \\
&= -\frac{1}{\ln 2} \left(\lambda_0 + \sum_{x\sim\chi} Q(x) \left\{\sum_{i\sim[k]} \lambda_i f_i(x)\right\}\right) \\
&= -\frac{1}{\ln 2} \left(\lambda_0 + \sum_{i\sim[k]} \lambda_i \alpha_i\right)
\end{aligned}$$

Now we will show that the KL-Divergence is simply the difference of the entropies of Q and P

$$\begin{aligned}
D(P||Q) &= \sum_{x\sim\chi} p(x) \log \frac{p(x)}{q(x)} \\
&= -\frac{1}{\ln 2} \sum_{x\sim\chi} p(x) \left\{\lambda_0 + \sum_{i\sim[k]} \lambda_i f_i(x)\right\} - H(P) \\
&= -\frac{1}{\ln 2} \left(\lambda_0 + \sum_{i\sim[k]} \lambda_i \alpha_i\right) - H(P) \\
&= H(Q) - H(P)
\end{aligned}$$

using the result from above. Finally, using the fundamental lower bound on the KL-Divergence we can show that Q is the maximum entropy distribution in the family $\mathcal{L}$

$$D(P||Q) = H(Q) - H(P) \geq 0$$

which requires that $H(Q) \geq H(P)$.

■

Problem 0.3. *Minimax rates for denoising*

Solution.

We will start by writing out the form of the KL-Divergence between the joint distributions over (X, Y) denoted as $P_f, P_g \in \Pi$ for a set of so far undefined functions Π. Using the chain rule for KL-Divergence,

$$\begin{aligned} D(P(X,Y)||Q(X,Y)) &= D(P(X)||Q(X)) + D(P(Y|X)||Q(Y|X)) \\ &= D(P(Y|X)||Q(Y|X)) \\ &= D(\mathcal{N}(f(x), \sigma^2)||\mathcal{N}(g(x), \sigma^2)) \end{aligned}$$

which is the KL-Divergence between two Gaussians. This can be computed as follows for the special case that σ is the same for both distributions

$$\begin{aligned} D(\mathcal{N}(f(x), \sigma^2)||\mathcal{N}(g(x), \sigma^2)) &= \frac{1}{\ln 2} \int_0^1 \exp\left(-(x - f(x))^2/2\sigma\right) \\ &\cdot \ln\left(\frac{\exp\left(-(x - f(x))^2/2\sigma\right)}{\exp\left(-(x - g(x))^2/2\sigma\right)}\right) dx \\ &= \frac{1}{2\ln 2 \cdot \sigma^2} \int_0^1 \exp\left(-(x - f(x))^2\right) \\ &\cdot \left((x - g(x))^2 - (x - f(x))^2\right) dx \\ &= \frac{1}{2\ln 2 \cdot \sigma^2} \int_0^1 |f(x) - g(x)|^2 dx \\ &= \frac{1}{2\ln 2 \cdot \sigma^2} \cdot ||f(x) - g(x)||_2^2 \end{aligned}$$

Next we will prove a lower bound on the minimax loss for n samples when we have a set of functions S. s.t. $f, g \in S$. First, we can manipulate the result from above and show that

$$D(P_f||P_g) = \frac{1}{2\ln 2 \cdot \sigma^2} \cdot ||f(x) - g(x)||_2^2 \leq \frac{32\delta^2}{\ln 2 \cdot \sigma^2}$$

since we have said that $||f - g||_2^2 \leq 8\delta$. At the same time, we can use what we know about lower bounds for minimax rates via multiple hypotheses. Let's say that a sequence of (X, Y) pairs $\bar{z}$ is drawn from the joint distribution P_s

where $s \in S$ and a tester T determines which P_s the data was drawn from (which amounts to denoising the data by determining the function s that generated it). The probability of error by T that outputs the distribution $s \in S$ the data was drawn from is lower bounded by

$$\mathbf{Pr}\left[T(\bar{z}) \neq s\right]\right] \geq 1 - \frac{n \cdot \underset{s_1, s_2 \in S}{\mathbf{E}}\left[D(P_{s_1}||P_{s_2})\right] + 1}{\log |S|}$$

Also, we define the loss function to be $\ell = ||f - g||_2^2$

$$\begin{aligned} M_n(\Pi, \ell) &\geq \ell(\delta) \cdot \inf_T \left\{\mathbf{Pr}\left[T(\bar{x}) \neq s\right]\right]\right\} \\ &= \delta^2 \cdot \left(1 - \frac{n \cdot \underset{s_1, s_2 \in S}{\mathbf{E}}\left[D(P_{s_1}||P_{s_2})\right] + 1}{\log |S|}\right) \\ &= \delta^2 \cdot \left(1 - \frac{n \cdot \left(32\delta^2/\sigma^2 \ln 2\right) + 1}{\log |S|}\right) \end{aligned}$$

where we substituted the result from above since the expectation over all possible pairs of functions can be at most the largest distance between two functions. Now, we define the set of functions S to be a series of non-intersecting bump functions which are L-Lipschitz since

$$\begin{aligned} |B_\epsilon(x_1) - B_\epsilon(x_2)| &= L \cdot ||x_2| - |x_1|| \\ &\leq L \cdot |x_2 - x_1| \end{aligned}$$

and if we make the assumption that $\epsilon < 1$

$$\begin{aligned} \int_{-1}^{1} (B_\epsilon(x))^2 dx &= \int_{-\epsilon}^{\epsilon} (B_\epsilon(x))^2 dx \\ &= 2\int_0^\epsilon (L \cdot (\epsilon - |x|)^2 dx \\ &= 2L^2 \int_0^\epsilon (x - \epsilon)^2 dx \\ &= \frac{2L^2\epsilon^3}{3} \end{aligned}$$

Next, we will show that the squared L_2 distance can be written in terms of the Hamming distance between the two position vectors

$$\begin{aligned}||f_a - f_b||_2^2 &= \int_{-1}^{1} |f_a(x) - f_b(x)|^2 dx \\ &= \int_{-1}^{1} |\sum_{i=1}^{m} (a_i - b_i) B_\epsilon(x - z_i)|^2 dx \\ &= \int_{-1}^{1} |\sum_{a_i \neq b_i} B_\epsilon(x - z_i)|^2 dx \\ &= \sum_{a_i \neq b_i} \int_{-1}^{1} \left(B_\epsilon(x - z_i)\right)^2 dx \\ &= \frac{2L^2\epsilon^3}{3} d_H(a, b)\end{aligned}$$

Finally, there exists a constant c_0 such that

$$\mathcal{M}_n(\Pi, \ell) \geq c_0 \cdot \left(\frac{\sigma^2 \cdot L}{n}\right)^{2/3}$$

■