EE/Ma 126b Information Theory - Homework Set #1

Ling Li, ling@cs.caltech.edu

January 10, 2001

Since problems 2, 3, and 5 in this homework all deal with the minimum of the relative entropy, here I introduce a theorem to relieve the total work:

Theorem: For a closed convex set $E \subset \mathcal{P}$ and distribution Q, let $P^* \in E$ be the distribution that achieves the minimum distance to Q, i.e.,

$$D(P^*||Q) = \min_{P \in E} D(P||Q).$$

Then (1) P^* exists and is unique; (2) if $Q \in E$ then $P^* = Q$, otherwise $P^* \in \partial E$, where ∂E is the boundary of set E.

Proof: $D(P||Q)$ is a continuous function with respect to P, so it can achieve its minimum on a closed set. Thus P^* exists. Suppose there are two distributions $P_1, P_2 \in E$ achieving the minimum of $D(P||Q)$, i.e.,

$$D(P_1||Q) = D(P_2||Q) = \min_{P \in E} D(P||Q).$$

Let $P_\lambda = \lambda P_1 + (1-\lambda)P_2$ for $0 < \lambda < 1$. The convexity of $D(P||Q)$ gives

$$D(P_\lambda||Q) \leq \lambda D(P_1||Q) + (1-\lambda)D(P_2||Q) = \min_{P \in E} D(P||Q). \tag{1}$$

Since E is a convex set, $P_\lambda \in E$, thus $\min_{P \in E} D(P||Q) \leq D(P_\lambda||Q)$. So the equality of (1) must hold, requiring that $P_1 = P_2$ since $D(P||Q)$ is strictly convex in P. Then P^* is unique.

If $Q \in E$, obviously $P^* = Q$ achieves the minimum of $D(P||Q)$. Otherwise, $Q \notin E$, $D(P^*||Q) = \min_{P \in E} D(P||Q) > 0$. Let $\tilde{P}_\lambda = \lambda P^* + (1-\lambda)Q$ for $0 < \lambda < 1$. We have

$$D(\tilde{P}_\lambda||Q) \leq \lambda D(P^*||Q) + (1-\lambda)D(Q||Q) = \lambda D(P^*||Q) < \min_{P \in E} D(P||Q).$$

Thus $\tilde{P}_\lambda \notin E$ for $0 < \lambda < 1$. Thus P^* is in the boundary of E, i.e., $P^* \in \partial E$, since no neighborhood of P^* is a subset of E. $\square$

1.1 *Lempel-Ziv.* The string is parsed as $0, 00, 000, 1, 10, 101, 0000, 01, 1010, 1$. The number of phrases is 10 and we need $\lceil \log 10 \rceil = 4$ bits to describe the location of the prefix to the phrase. Thus the Lempel-Ziv code for the string is

$$\boxed{(0000,0)(0001,0)(0010,0)(0000,1)(0100,0)(0101,1)(0011,0)(0001,1)(0110,0)(0000,1)}.$$

Note that:

- We can use only 3 bits to describe the location, since the locations with number larger than 7 are not used in this code.
- The last phrase is the same as the 4^{th} phrase in the string. Because the string ends there, we can not find a phrase that has not been marked off before.

1.2 *Error exponent for universal codes.* $H(P)$ is a concave function of P. Thus $E = \{P : H(P) \geq R\}$ is a closed convex set, and its boundary is $\partial E = \{P : H(P) = R\}$.[1] For practical reason, we assume that $R \leq \log|\mathcal{X}|$.

(a) From the theorem, if $H(Q) \geq R$, i.e., $Q \in E$, then $P^* = Q$. If $H(Q) < R$, then $P^* \in \partial E$, i.e., $H(P^*) = R$. That is, P^* is the solution to the optimization problem

$$\min D(P||Q)$$

subject to

$$\sum_x P(x) = 1, H(P) = R.$$

We can use the method of Lagrange multipliers to find P^*. Let

$$L = D(P||Q) - \lambda_1 (H(P) - R) - \lambda_2 \left(\sum_x P(x) - 1 \right).$$

From

$$\frac{\partial L}{\partial P(x)} = \log e \left(\ln \frac{P(x)}{Q(x)} + 1 \right) + \lambda_1 \log e (\ln P(x) + 1) - \lambda_2 = 0,$$

we get

$$P(x) = 2^{\frac{\lambda_2}{1+\lambda_1}} e^{-1} Q(x)^{\frac{1}{1+\lambda_1}}.$$

Let $\lambda = \frac{1}{1+\lambda_1}$. From the constraint $\sum_x P(x) = 1$, we get

$$2^{\frac{\lambda_2}{1+\lambda_1}} e^{-1} = \frac{1}{\sum_x Q(x)^\lambda}.$$

Thus

$$P(x) = \frac{Q(x)^\lambda}{\sum_x Q(x)^\lambda}. \tag{2}$$

λ can be decided by the constraint $H(P) = R$. However, we will get two λ's, one for the minimum, one for the maximum. Since $H(Q) < R$, we get the range for the minimum is $0 \leq \lambda < 1$.[2] So

$$\boxed{P^*(x) = \begin{cases} Q(x), & \text{when } H(Q) \geq R; \\ \frac{Q(x)^\lambda}{\sum_x Q(x)^\lambda}, \ 0 \leq \lambda < 1 \text{ decided by } H(P^*) = R, & \text{when } H(Q) < R. \end{cases}}$$

[1] The proof (together with some proofs in problems 3 and 5) is straightforward and is omitted here.

[2] We can differentiate $H(P)$ and $D(P||Q)$ with respect to λ for (2):

$$\begin{aligned} \frac{dH(P)}{d\lambda} &= -\lambda \left[\sum_x P(x)(\ln Q(x))^2 - (\sum_x P(x) \ln Q(x))^2 \right] \log e, \\ \frac{dD(P||Q)}{d\lambda} &= (\lambda - 1) \left[\sum_x P(x)(\ln Q(x))^2 - (\sum_x P(x) \ln Q(x))^2 \right] \log e. \end{aligned}$$

(b) $P_e^{(n)} \to 0 \Leftrightarrow D(P^*||Q) > 0 \Leftrightarrow H(Q) < R$. For binary X, the range of $Q(0)$ or $Q(1)$ is $\boxed{[0,q) \cup (1-q,1]}$, where $H(q) = R$ and $q \leq \frac{1}{2}$.

1.3 *Sequential projection.* $\mathcal{P}_1$ and $\mathcal{P}_2$ are both closed convex sets. Thus $\mathcal{P}_1 \cap \mathcal{P}_2$ is also a closed convex set. We will analyze this problem in 3 cases:

- If $Q \in \mathcal{P}_1$. Thus $P^* = Q$, and R^* minimizes $D(R||P^*)$ since R^* minimizes $D(R||Q)$ over all $R \in \mathcal{P}_1 \cap \mathcal{P}_2$.
- If $Q \notin \mathcal{P}_1$ and $P^* \in \mathcal{P}_1 \cap \mathcal{P}_2$. Thus $Q \notin \mathcal{P}_1 \cap \mathcal{P}_2$ and the theorem says that the one that minimizes $D(R||Q)$ over all $R \in \mathcal{P}_1 \cap \mathcal{P}_2$ exists and is unique. Thus $R^* = P^*$. R^* also minimizes $D(R||P^*)$.
- Otherwise, $Q \notin \mathcal{P}_1$ and $P^* \notin \mathcal{P}_1 \cap \mathcal{P}_2$. By the method of Lagrange multipliers, let

$$L = D(P||Q) - \tilde{\lambda}_0 \left(\sum_x p(x) - 1 \right) - \sum_{i=1}^{r} \lambda_i \left(\sum_x p(x) h_i(x) - \alpha_x \right).$$

From

$$\frac{\partial L}{\partial p(x)} = \log e \left(\ln \frac{p(x)}{q(x)} + 1 \right) - \tilde{\lambda}_0 - \sum_{i=1}^{r} \lambda_i h_i(x) = 0,$$

we get P^* is of the form

$$P^*(x) = Q(x) \cdot 2^{\lambda_0 + \sum_{i=1}^{r} \lambda_i h_i(x)},$$

where $\lambda_0 = \tilde{\lambda}_0 - \log e$, λ_i $(i = 1, 2, \ldots, r)$ can be decided by the constraints.[3] Similarly, R^* is of the form

$$R^*(x) = Q(x) \cdot 2^{\delta_0 + \sum_{i=1}^{r} \delta_i h_i(x) + \sum_{j=1}^{s} \zeta_j g_j(x)}.$$

For R' that minimizes $D(R||P^*)$ over all $R \in \mathcal{P}_1 \cap \mathcal{P}_2$, we also have

$$R'(x) = P^*(x) \cdot 2^{\delta'_0 + \sum_{i=1}^{r} \delta'_i h_i(x) + \sum_{j=1}^{s} \zeta'_j g_j(x)} = Q(x) \cdot 2^{(\lambda_0 + \delta'_0) + \sum_{i=1}^{r} (\lambda_i + \delta'_i) h_i(x) + \sum_{j=1}^{s} \zeta'_j g_j(x)},$$

in the same form as R^*. Since the parameters in R^* and R' are decided by the same constraints and by the theorem R^* is unique, we have

$$R^* = R'.$$

That is, R^* minimizes $D(R||P^*)$ over all $R \in \mathcal{P}_1 \cap \mathcal{P}_2$. $\square$

Since x^2 is a strictly convex function of x,

$$\sum_x P(x)(\ln Q(x))^2 > (\sum_x P(x) \ln Q(x))^2$$

when Q is not a uniform distribution. Thus $H(P)$ gets smaller along λ when $\lambda > 0$, and $D(P||Q)$ gets smaller along λ when $\lambda < 1$. Since when $\lambda = 1$, $P = Q$ and $H(P) = H(Q)$, so $H(P) \geq R > H(Q)$ gives $\lambda < 1$. And the minimum of $D(P||Q)$ requires λ be as near to 1 as possible. When $R \leq \log|\mathcal{X}|$, $\lambda = 0$ gives a P satisfying $H(P) = \log|\mathcal{X}| \geq R$. Thus the λ for the minimum also satisfies $\lambda \geq 0$ (and that of the maximum satisfies $\lambda \leq 0$).

The cases that $R > \log|\mathcal{X}|$ or Q is a uniform distribution (which also means $R > H(Q) = \log|\mathcal{X}|$) are trivial.

[3] Like in problem 2, we need to distinguish the minimum and maximum of $D(P||Q)$.

1.4 *Counting.* Let Q be the uniform distribution and $E = \{P : \sum_{i=1}^{m} P(i)g(i) \geq \alpha\}$. We have

$$D(P||Q) = \log|\mathcal{X}| - H(P),$$

and

$$P^* = \arg\min_{P\in E} D(P||Q) = \arg\max_{P\in E} H(P).$$

Since E is a closed convex set, Sanov's theorem says

$$\frac{1}{n}\log Q^n(E) \to -D(P^*||Q) = H(P^*) - \log|\mathcal{X}|.$$

Thus from

$$|T(E)| = \sum_{P\in E\cap\mathcal{P}_n} |T(P)| = \frac{Q^n(E)}{|\mathcal{X}|^{-n}},$$

we have

$$\boxed{\frac{1}{n}\log|T(E)| = \frac{1}{n}\log Q^n(E) + \log|\mathcal{X}| \to H(P^*)}. \tag{3}$$

Since

$$\frac{1}{n}\sum_{i=1}^{n} g(x_i) \geq \alpha \Leftrightarrow \sum_{i=1}^{m} P_{x^n}(i)g(i) \geq \alpha \Leftrightarrow P_{x^n} \in E \cap \mathcal{P}_n$$

and

$$H^* = \max_{P:\sum_{i=1}^{m} P(i)g(i)\geq\alpha} H(P) = H(P^*),$$

we get the number of sequences $x^n \in \mathcal{X}^n$ satisfying $\frac{1}{n}\sum_{i=1}^{n} g(x_i) \geq \alpha$ is $|T(E)|$, which is approximately equal to 2^{nH^*} according to (3), to first order in the exponent for n sufficiently large. □

1.5 *Joint distributions and product distributions.* Obviously, E is a closed convex set.

(a) We want to

$$\min D(P||Q_0)$$

while subject to

$$\begin{aligned}
\sum_{x,y} P(x,y) &= 1, & (4)\\
-\sum_{x,y} P(x,y)\log Q(x) - H(X) &= 0, & (5)\\
-\sum_{x,y} P(x,y)\log Q(y) - H(Y) &= 0, & (6)\\
-\sum_{x,y} P(x,y)\log Q(x,y) - H(X,Y) &= 0. & (7)
\end{aligned}$$

Let

$$\begin{aligned}
L = {} & D(P||Q_0) - \tilde{\lambda}_0\left(\sum_{x,y} P(x,y) - 1\right) - \tilde{\lambda}_1\left(-\sum_{x,y} P(x,y)\log Q(x) - H(X)\right)\\
& - \tilde{\lambda}_2\left(-\sum_{x,y} P(x,y)\log Q(y) - H(Y)\right) - \tilde{\lambda}_3\left(-\sum_{x,y} P(x,y)\log Q(x,y) - H(X,Y)\right).
\end{aligned}$$

The minimum distribution P^* satisfies $\frac{\partial L}{\partial P(x,y)} = 0$, i.e.,

$$\log e \left(\ln \frac{P^*(x,y)}{Q_0(x,y)} + 1 \right) - \tilde{\lambda}_0 + \tilde{\lambda}_1 \log Q(x) + \tilde{\lambda}_2 \log Q(y) + \tilde{\lambda}_3 \log Q(x,y) = 0.$$

Thus

$$\boxed{P^*(x,y) = Q_0(x,y) e^{\left(\tilde{\lambda}_0 - \tilde{\lambda}_1 \log Q(x) - \tilde{\lambda}_2 \log Q(y) - \tilde{\lambda}_3 \log Q(x,y)\right) \ln 2 - 1}}.$$

By assigning

$$\lambda_0 = \tilde{\lambda}_0 \ln 2 - 1, \lambda_1 = -\tilde{\lambda}_1 \ln 2, \lambda_2 = -\tilde{\lambda}_2 \ln 2, \lambda_3 = -\tilde{\lambda}_3 \ln 2,$$

we have P^* is of the form

$$\boxed{P^*(x,y) = Q_0(x,y) e^{\lambda_0 + \lambda_1 \log Q(x) + \lambda_2 \log Q(y) + \lambda_3 \log Q(x,y)}}, \tag{8}$$

where λ_0, λ_1, λ_2, and λ_3 ($\tilde{\lambda}_0$, $\tilde{\lambda}_1$, $\tilde{\lambda}_2$, and $\tilde{\lambda}_3$) are chosen to satisfy the constraints (4)–(7). Since E is a closed convex set, the theorem gives that P^* is unique. □

(b) Let $\lambda_0 = 0$, $\lambda_1 = \lambda_2 = -\ln 2$, and $\lambda_3 = \ln 2$ in (8), we get

$$P^*(x,y) = Q_0(x,y) \frac{Q(x,y)}{Q(x)Q(y)} = Q(x,y).$$

It is also easy to see that $Q(x,y)$ satisfies the constraints (4)–(7), i.e.,

$$\begin{aligned}
\sum_{x,y} Q(x,y) &= 1, \\
-\sum_{x,y} Q(x,y) \log Q(x) - H(X) &= -\sum_{x} Q(x) \log Q(x) - H(X) = 0, \\
-\sum_{x,y} Q(x,y) \log Q(y) - H(Y) &= -\sum_{y} Q(y) \log Q(y) - H(Y) = 0, \\
-\sum_{x,y} Q(x,y) \log Q(x,y) - H(X,Y) &= 0.
\end{aligned}$$

So the distribution in E closest to the product distribution is the joint distribution. □