EE/Ma 126b Information Theory - Homework Set #3

Ling Li, ling@cs.caltech.edu

January 24, 2001

3.1 *Maximum entropy.* The support set is $S = \mathcal{R}^+$. The maximum entropy density with constraints

$$\int_S f(x)dx = 1, \quad \int_S xf(x)dx = \alpha_1, \quad \int_S (\ln x)f(x)dx = \alpha_2,$$

is of the form

$$f(x) = e^{\lambda_0 + \lambda_1 x + \lambda_2 \ln x} = e^{\lambda_0} x^{\lambda_2} e^{\lambda_1 x}, \quad x \in S,$$

where the parameters λ_0, λ_1, and λ_2 are chosen so that f satisfies the constraints. In order to satisfy $\int_S f(x)dx = 1$, $\boxed{\lambda_1 < 0 \text{ and } \lambda_2 > -1}$.* By changing a variable,

$$1 = \int_S f(x)dx = e^{\lambda_0} \int_0^\infty x^{\lambda_2} e^{\lambda_1 x} dx = e^{\lambda_0} (-\lambda_1)^{-\lambda_2 - 1} \int_0^\infty x^{\lambda_2} e^{-x} dx,$$

so

$$e^{\lambda_0} = \frac{(-\lambda_1)^{\lambda_2+1}}{\Gamma(\lambda_2 + 1)}, \quad \text{or} \quad \lambda_0 = (\lambda_2 + 1)\ln(-\lambda_1) - \ln\Gamma(\lambda_2 + 1),$$

where $\Gamma(t) = \int_0^\infty x^{t-1} e^{-x} dx$ is the Euler gamma function. Thus we know

$$\boxed{f(x) = \frac{x^{k-1} e^{-x/b}}{\Gamma(k) b^k}, \quad x > 0}$$

is a $\boxed{\textit{gamma distribution}}$ with shape parameter $k = \lambda_2 + 1$ and scale parameter $\boxed{b = -\lambda_1^{-1}}$. From the second constraint, $EX = kb = -k\lambda_1^{-1} = \alpha_1$, we get $\boxed{k = -\alpha_1 \lambda_1}$. The last parameter λ_1 can be decided by the last constraint.

3.2 *Maximum entropy with marginals.* For any joint distribution $p(x, y)$ that has the fixed marginals $p(x)$ and $p(y)$, we claim that $p^*(x, y) = p(x)p(y)$ maximizing the entropy $H(X, Y) = H(p(x, y))$. And $p^*(x, y)$ is the only maximizing distribution.

Proof: For any distribution $p(x, y)$ that satisfies $\sum_y p(x, y) = p(x)$ and $\sum_x p(x, y) = p(y)$, we have $H(X, Y) = H(X) + H(Y|X) \le H(X) + H(Y)$ with equality iff X and Y are independent. Thus $H(X, Y)$ gets its maximum $H(X) + H(Y)$ iff $p(x, y) = p(x)p(y)$.

*$\lambda_1 > 0$ makes $f(x) \to \infty$ when $x \to \infty$; $\lambda_1 = 0$ makes $f(x) = e^{\lambda_0} e^{\lambda_1 x}$ and its integral doesn't converge on S; when $\lambda_2 \le -1$, the integral of $f(x)$ on $(0, 1)$ does not converge.

Thus, the maximum entropy distribution $p(x, y)$ for the problem is

$x \backslash y$	1	2	3	$p(x)$
1	1/3	1/12	1/12	1/2
2	1/6	1/24	1/24	1/4
3	1/6	1/24	1/24	1/4
$p(y)$	2/3	1/6	1/6	

3.3 *Rate distortion function with infinite distortion.* Considering distribution $p(\hat{x}|x)$ such that

$$Ed(X, \hat{X}) = \sum_{x,\hat{x}} p(x)p(\hat{x}|x)d(x, \hat{x}) = \frac{1}{2}\left[p(\hat{x} = 0|x = 1) + p(\hat{x} = 1|x = 0) \cdot \infty\right] \le D,$$

we have $p(\hat{x} = 1|x = 0) = 0$ and thus $p(\hat{x} = 0|x = 0) = 1$. Let p denote $p(\hat{x} = 0|x = 1)$ for convenience. Then we have

$$p(\hat{x} = 0) = 1 \cdot p(x = 0) + p \cdot p(x = 1) = \frac{1 + p}{2}.$$

So

$$p(x = 0|\hat{x} = 0) = \frac{p(\hat{x} = 0|x = 0)p(x = 0)}{p(\hat{x} = 0)} = \frac{1}{1 + p}, \quad p(x = 0|\hat{x} = 1) = 0,$$

and

$$\begin{aligned} I(X; \hat{X}) &= H(X) - H(X|\hat{X}) \\ &= 1 - p(\hat{x} = 0)H\left(\frac{1}{1 + p}\right) - p(\hat{x} = 1)H(0) \\ &= 1 - \frac{1 + p}{2}H\left(\frac{1}{1 + p}\right). \end{aligned} \tag{1}$$

Differentiate (1), we have

$$\frac{\partial I(X; \hat{X})}{\partial p} = \frac{1}{2(1 + p)}\log p - \frac{1}{2}H\left(\frac{1}{1 + p}\right) = \frac{1}{2}\log\frac{p}{1 + p} < 0$$

since $\frac{p}{1+p} < 1$. So the minimum of $I(X; \hat{X})$ is achieved at the maximum of p. From

$$Ed(X, \hat{X}) = \frac{1}{2}p \le D$$

and $0 \le p \le 1$, the maximum of p is $\min\{2D, 1\}$. So

$$\begin{aligned} R(D) &= \min_{Ed(X,\hat{X}) \le D} I(X; \hat{X}) = \left[1 - \frac{1 + p}{2}H\left(\frac{1}{1 + p}\right)\right]\Bigg|_{p=\min\{2D,1\}} \\ &= \boxed{\begin{cases} 1 - \frac{1+2D}{2}H\left(\frac{1}{1+2D}\right), & 0 \le D < 1/2; \\ 0, & D \ge 1/2. \end{cases}} \end{aligned}$$

3.4 *Rate distortion for binary source with asymmetric distortion.* Let p_{01} denote the cross probability $p(\hat{x} = 1|x = 0)$ and p_{10} denote $p(\hat{x} = 0|x = 1)$. The distortion is

$$Ed(X, \hat{X}) = \sum_{x,\hat{x}} p(x)p(\hat{x}|x)d(x, \hat{x}) = \frac{1}{2}(ap_{01} + bp_{10}).$$

Since

$$p(\hat{x}=0) = p(\hat{x}=0|x=0)p(x=0) + p(\hat{x}=0|x=1)p(x=1) = \frac{1}{2}(1-p_{01}+p_{10}),$$

the mutual information is

$$\begin{aligned} I(X;\hat{X}) &= H(\hat{X}) - H(\hat{X}|X) \\ &= H\left(\frac{1}{2}(1-p_{01}+p_{10})\right) - \frac{1}{2}H(p_{01}) - \frac{1}{2}H(p_{10}). \end{aligned} \tag{2}$$

Thus

$$\frac{\partial I(X;\hat{X})}{\partial p_{01}} = \frac{1}{2}\log\frac{1-p_{01}+p_{10}}{1+p_{01}-p_{10}} - \frac{1}{2}\log\frac{1-p_{01}}{p_{01}}, \tag{3}$$

$$\frac{\partial I(X;\hat{X})}{\partial p_{10}} = \frac{1}{2}\log\frac{1+p_{01}-p_{10}}{1-p_{01}+p_{10}} - \frac{1}{2}\log\frac{1-p_{10}}{p_{10}}. \tag{4}$$

Since $\log(\cdot)$ is a monotonic increasing function, and

$$\frac{1-p_{01}+p_{10}}{1+p_{01}-p_{10}} - \frac{1-p_{01}}{p_{01}} = \frac{p_{01}+p_{10}-1}{p_{01}(1+p_{01}-p_{10})},$$

$$\frac{1+p_{01}-p_{10}}{1-p_{01}+p_{10}} - \frac{1-p_{10}}{p_{10}} = \frac{p_{01}+p_{10}-1}{p_{10}(1-p_{01}+p_{10})},$$

and $|p_{01}-p_{10}| \le 1$, we have $\operatorname{sgn}\left(\frac{\partial I(X;\hat{X})}{\partial p_{01}}\right) = \operatorname{sgn}\left(\frac{\partial I(X;\hat{X})}{\partial p_{10}}\right) = \operatorname{sgn}(p_{01}+p_{10}-1)$. Then we can decrease $I(X;\hat{X})$ by decreasing p_{01} and/or p_{10} when $p_{01}+p_{10} > 1$, or by increasing p_{01} and/or p_{10} when $p_{01}+p_{10} < 1$. The minimum of $I(X;\hat{X})$, which is 0, is achieved when $p_{01}+p_{10} = 1$.[†] However, $Ed(X,\hat{X}) \le D$ restricts that $ap_{01}+bp_{10} \le 2D$. So

(a) When $a \le 2D$ or $b \le 2D$, $p_{01}+p_{10}=1$ can be achieved, either when $p_{01}=1, p_{10}=0$ or when $p_{01}=0, p_{10}=1$. So now $\boxed{R(D)=0}$.

(b) When $a > 2D$ and $b > 2D$, $p_{01}+p_{10} < 1$. However, from the above discussion, the minimum of $I(X;\hat{X})$ is achieved at the boundary of $ap_{01}+bp_{10} = 2D$, since we can always increase p_{01} and/or p_{10} when $ap_{01}+bp_{10} < 2D$, to decrease the mutual information. Thus we can use the method of Lagrange multipliers. Let

$$L = I(X;\hat{X}) - \lambda(ap_{01}+bp_{10}-2D)$$

and solve $\frac{\partial L}{\partial p_{01}} = \frac{\partial L}{\partial p_{10}} = \frac{\partial L}{\partial \lambda} = 0$, i.e.,

$$\frac{\partial I(X;\hat{X})}{\partial p_{01}} - \lambda a = \frac{\partial I(X;\hat{X})}{\partial p_{10}} - \lambda b = ap_{01}+bp_{10}-2D = 0.$$

After eliminating the parameter λ and using (3) and (4), we get

$$\begin{cases} (a+b)\log\frac{1-p_{01}+p_{10}}{1+p_{01}-p_{10}} + a\log\frac{1-p_{10}}{p_{10}} - b\log\frac{1-p_{01}}{p_{01}} = 0, \\ ap_{01}+bp_{10}-2D = 0. \end{cases}$$

Solving these equations and then using (2), we can get the minimum of $I(X;\hat{X})$, i.e., $R(D)$. (If no solutions are found, the minimum of $I(X;\hat{X})$ is at one of the two ends: $(p_{01}=\frac{2D}{a}, p_{10}=0)$ and $(p_{01}=0, p_{10}=\frac{2D}{b})$.)

[†] When $p_{01}+p_{10}=1$, $H\left(\frac{1}{2}(1-p_{01}+p_{10})\right) = H(p_{10})$ and $H(p_{01}) = H(1-p_{10}) = H(p_{10})$, so $I(X;\hat{X}) = 0$.

3.5 *Shannon lower bound for the rate distortion function.* Let $\mathcal{P}(D) = \{\mathbf{p} : \sum_{i=1}^m p_i d_i \le D\}$. Thus

$$\phi(D) = \max_{\mathbf{p}\in\mathcal{P}(D)} H(\mathbf{p}). \tag{5}$$

(a) For any $D', D'' \ge 0$, and $\lambda \in [0,1]$, let

$$\mathbf{p}' = \arg\max_{\mathbf{p}\in\mathcal{P}(D')} H(\mathbf{p}), \quad \mathbf{p}'' = \arg\max_{\mathbf{p}\in\mathcal{P}(D'')} H(\mathbf{p}),$$

and $\mathbf{p}^{(\lambda)} = \lambda\mathbf{p}' + (1-\lambda)\mathbf{p}''$. Thus the concavity of $H(\mathbf{p})$ gives

$$H(\mathbf{p}^{(\lambda)}) \ge \lambda H(\mathbf{p}') + (1-\lambda)H(\mathbf{p}'') = \lambda\phi(D') + (1-\lambda)\phi(D''). \tag{6}$$

Since $\mathbf{p}' \in \mathcal{P}(D')$ and $\mathbf{p}'' \in \mathcal{P}(D'')$, we have $\sum_{i=1}^m p'_i d_i \le D'$ and $\sum_{i=1}^m p''_i d_i \le D''$, and

$$\sum_{i=1}^m p_i^{(\lambda)} d_i = \lambda\sum_{i=1}^m p'_i d_i + (1-\lambda)\sum_{i=1}^m p''_i d_i \le \lambda D' + (1-\lambda)D'',$$

i.e., $\mathbf{p}^{(\lambda)} \in \mathcal{P}(\lambda D' + (1-\lambda)D'')$. Together with (5) and (6), this gives

$$\phi(\lambda D' + (1-\lambda)D'') = \max_{\mathbf{p}\in\mathcal{P}(\lambda D'+(1-\lambda)D'')} H(\mathbf{p}) \ge H(\mathbf{p}^{(\lambda)}) \ge \lambda\phi(D') + (1-\lambda)\phi(D'').$$

So $\phi(D)$ is a concave function of D. Besides, since $\mathcal{P}(D) \subseteq \mathcal{P}(D')$ when $D \le D'$, we know $\phi(D)$ is also a non-decreasing function of D. □

(b) If $Ed(X,\hat{X}) \le D$, i.e.,

$$\sum_{\hat{x}} p(\hat{x})D_{\hat{x}} = \sum_{\hat{x}} p(\hat{x})\sum_x p(x|\hat{x})d(x,\hat{x}) = \sum_{x,\hat{x}} p(x,\hat{x})d(x,\hat{x}) \le D, \tag{7}$$

we have

$$\begin{aligned} I(X;\hat{X}) &= H(X) - H(X|\hat{X}) && (8)\\ &= H(X) - \sum_{\hat{x}} p(\hat{x})H(X|\hat{X}=\hat{x}) && (9)\\ &\ge H(X) - \sum_{\hat{x}} p(\hat{x})\phi(D_{\hat{x}}) && (10)\\ &\ge H(X) - \phi\left(\sum_{\hat{x}} p(\hat{x})D_{\hat{x}}\right) && (11)\\ &\ge H(X) - \phi(D). && (12) \end{aligned}$$

Here

- (8) is (2.39) in Cover's book;
- (9) is the definition of the conditional entropy;
- Since for fixed $\hat{x}$, $\{d(x,\hat{x})|x\in\mathcal{X}\}$ is a permutation of $\{d_1, d_2, \ldots, d_m\}$, thus from $D_{\hat{x}} = \sum_x p(x|\hat{x})d(x,\hat{x})$ we know that one permutation of $\{p(x|\hat{x})|x\in\mathcal{X}\}$, say $\mathbf{p}$, satisfies $\sum_i p_i d_i = D_{\hat{x}}$, i.e., $\mathbf{p} \in \mathcal{P}(D_{\hat{x}})$. Since permutation doesn't change the entropy, we have

$$H(X|\hat{X}=\hat{x}) = H(\mathbf{p}) \le \phi(D_{\hat{x}}). \tag{13}$$

  This explains (10);

- (11) is because of the concavity of $\phi(D)$ and $p(\hat{x}) \geq 0$ and $\sum_{\hat{x}} p(\hat{x}) = 1$;
- From (7) and $\phi(D)$ is non-decreasing, we finally get (12).

(c) From (b) we know if $Ed(x, \hat{x}) \leq D$ then $I(X; \hat{X}) \geq H(X) - \phi(D)$. So

$$\boxed{R(D) = \min_{Ed(x,\hat{x}) \leq D} I(X; \hat{X}) \geq H(X) - \phi(D)}. \tag{14}$$

Let

$$\mathbf{p}^* = \arg \max_{\mathbf{p} \in \mathcal{P}(D)} H(\mathbf{p}).$$

Since for any fixed $\hat{x}$, $\{d(x, \hat{x}) | x \in \mathcal{X}\}$ is a permutation of $\{d_1, d_2, \ldots, d_m\}$, we can make $p(x|\hat{x})$ a permutation of $\mathbf{p}^*$ with the same order as $d(x, \hat{x})$. Thus we have

- $D_{\hat{x}} = \sum_x p(x|\hat{x}) d(x, \hat{x}) = \sum_i p_i^* d_i$ is the same for all $\hat{x} \in \hat{\mathcal{X}}$;
- $D_{\hat{x}} = \sum_i p_i^* d_i \leq D$, since $\mathbf{p}^* \in \mathcal{P}(D)$;
- $H(X|\hat{X} = \hat{x}) = H(\mathbf{p}^*) = \phi(D) = \phi(D_{\hat{x}})$, since we also have $\mathbf{p}^* \in \mathcal{P}(D_{\hat{x}})$ and $D_{\hat{x}} \leq D$.

For such $p(x|\hat{x})$, we have the equalities of (13), (10), (11), and (12). Thus the lower bound of $R(D)$ can be achieved, i.e., $R(D) = H(X) - \phi(D)$.

However, till now we have not prove that such $p(x|\hat{x})$ *meets* the source distribution. If any distribution of $\hat{X}$, together with such $p(x|\hat{x})$, can not satisfy the given source distribution, then we can not claim $R(D) = H(X) - \phi(D)$. Luckily, if in addition, we assume that the source has a uniform distribution and the rows of the distortion matrix are permutations of each other, such $p(x|\hat{x})$ can meet the source distribution.

Let $\hat{X}$ also be uniformly distributed. Since the rows of the distortion matrix are permutations of each other, our way to produce $p(x|\hat{x})$ assures that $\left\{p(x|\hat{x}) | \hat{x} \in \hat{\mathcal{X}}\right\}$ for fixed x is a permutation of that of a different x. Thus

$$p(x) = \sum_{\hat{x}} p(x|\hat{x}) p(\hat{x}) = \frac{1}{|\hat{\mathcal{X}}|} \sum_{\hat{x}} p(x|\hat{x})$$

is invariant for all $x \in \mathcal{X}$, i.e., X is uniformly distributed. So now the source distribution is met and we have $R(D) = H(X) - \phi(D)$. □