Massachusetts Institute of Technology
Department of Electrical Engineering and Computer Science
6.437 Inference and Information
Spring 2015

Problem Set 7

Issued: Tuesday, April 7, 2015 **Due:** Tuesday, April 14, 2015

Problem 7.1 (practice)

Consider a Gaussian random variable $\mathsf{x} \sim \mathcal{N}(0, \sigma_{\mathsf{x}}^2)$, where σ_{x}^2 is fixed.

(a) Calculate the differential entropy $h(\mathsf{x})$.

Now consider a model $p_{\mathsf{y}|\mathsf{x}}(y \,|\, x) \sim \mathcal{N}(x, \sigma_{\mathsf{y}|\mathsf{x}}^2)$, where $\sigma_{\mathsf{y}|\mathsf{x}}^2$ is fixed.

(b) Calculate the mutual information between x and y.

Problem 7.2

Suppose we have observations of the form

$$\mathsf{y} \triangleq \sqrt{\rho}\mathsf{x} + \mathsf{z},$$

where x and z are independent Gaussian random variables each with zero-mean and unit-variance, and where $\rho > 0$ is some constant.

(a) Determine the Bayes least-squares estimate $\hat{x}_{\mathrm{BLS}}(y)$ of x based on the observed value y, and compute the corresponding mean-square estimation error $\lambda_{\mathrm{BLS}}(\rho)$, expressing your answer as a function of ρ.

Hint: You may find it convenient to recall that the sum of two independent Gaussian random variables is also a Gaussian random variable.

(b) Show that

$$h(\mathsf{y}|\mathsf{x}) = h(\mathsf{y} - a\mathsf{x}|\mathsf{x}),$$

for any constant a.

(c) Show that

$$I(\mathsf{x};\mathsf{y}) = h(\mathsf{y}) - h(\mathsf{z}),$$

and then use this result to express

$$I(\rho) \triangleq I(\mathsf{x};\mathsf{y}).$$

as a function of ρ.

(d) Establish that

$$\frac{d}{d\rho} I(\rho) = \frac{1}{2} \lambda_{\mathrm{BLS}}(\rho).$$

Problem 7.3

Consider N flips of a biased coin: i.e., let y be a vector of N independent identically distributed Bernoulli(x) observations, where $x \in [0,1]$. That is, $\mathsf{y}_i = 1$ with probability x, and $\mathsf{y}_i = 0$ otherwise. Consider a symmetric Beta(θ, θ) prior on x:

$$p_{\mathsf{x}}(x) = c(\theta)\, x^{\theta-1}(1-x)^{\theta-1}, \quad 0 \le x \le 1,$$

where $c(\theta) = (\Gamma(2\theta))/(\Gamma(\theta))^2$ and Γ is the Gamma function. Note that the uniform prior corresponds to $\theta = 1$.

A few facts about Gamma function and Beta function are summarized as follows:

- $\Gamma(z) \triangleq \int_0^\infty x^{z-1} e^{-x} \mathrm{d}x$
- $\Gamma(z) = (z-1)\Gamma(z-1)$
- $\mathrm{Beta}(x,y) \triangleq \int_0^1 t^{x-1}(1-t)^{y-1}\mathrm{d}t = \frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)}$

(a) For $N = 5$, plot, as a function of θ over the range $0 < \theta \le 1$, the normalized mutual information $I(\mathsf{y}; \mathsf{x})/N$ obtained with the Beta prior. Plot additional curves for $N = 10, 15,$ and 20.

Hint: In MATLAB, the `quad()` function can perform the necessary integration

(b) For the Beta prior, determine the form of the associated predictor (mixture) distribution $q(\mathsf{y}_N \mid \mathsf{y}_{N-1}, \ldots, \mathsf{y}_1)$ as a function of the Beta prior parameter θ. How does varying θ affect the predictor for small N? For large N?

Now consider the uniform prior on x.

(c) Fix N and ϵ. Let q^* be the uniform mixing distribution (in Bayesian terminology, the marginal probability of y under the uniform prior). Show that

$$D(p_{\mathsf{y}|\mathsf{x}}(\mathbf{y} \mid x) \,\|\, q^*) \le \log(N+1)$$

for all x. Then determine for what fraction ρ of values of $x \in [0,1]$ we have

$$D(p_{\mathsf{y}|\mathsf{x}}(\mathbf{y} \mid x) \,\|\, q^*) \ge (1-\epsilon)\log(N+1).$$

Plot the associated ρ versus ϵ curve for $N = 5$.

To determine the fraction ρ, you may want to just look at a finite set of values of x (say M uniformly spaced samples between 0 and 1, where M is reasonably large) and evaluate $D(\cdot\|\cdot)$ for each of those, keeping track of the fraction that are above the threshold. This gives an approximation to what we want.

(d) Now vary N, and examine the progression of ρ versus ϵ curves as N gets larger. What appears to happen as $N \to \infty$? (You can confirm this analytically by finding the limits as $N \to \infty$.)

Here's a handy approximation for the entropy of a binomial random variable n with parameters N and p, valid for large N:

$$H(\mathsf{n}) \approx \frac{1}{2}\left(1 + \log\left(2\pi N p\,(1-p)\right)\right).$$

Problem 7.4

(a) Determine a conjugate prior family for exponential models of the form

$$p_{\mathsf{y}|\mathsf{x}}(y \mid x) = x e^{-xy}.$$

where $y \geq 0$.

(b) Determine a conjugate prior family for Poisson models of the form

$$p_{\mathsf{y}|\mathsf{x}}(y \mid x) = \frac{x^y e^{-x}}{y!} \qquad y = 0, 1, 2, \ldots.$$

Problem 7.5

Let $\mathbf{z} = [\mathsf{z}_1, \ldots, \mathsf{z}_N]^{\mathrm{T}}$ be a vector of N i.i.d. variables distributed according to

$$p_{\mathsf{z}}(z; x) = \begin{cases} x/3, & z = 0, \\ 2x/3, & z = 1, \\ 1 - x, & z = 2, \end{cases}$$

where $x \in [0, 1]$ is an unknown parameter. We do not observe $\mathbf{z}$ directly, but instead get access to the corresponding sequence of measurements $\mathbf{y} = [y_1, \ldots, y_N]^{\mathrm{T}}$ such that

$$y_n = g(z_n) = \begin{cases} 0, & z_n = 0, 2, \\ 1, & z_n = 1, \end{cases}$$

for $n = 1, \ldots, N$.

In a particular run of the experiment, we observe that 1/4 of all elements in the observed sequence $\mathbf{y}$ are ones, and the rest are zeros (i.e., $\sum_{n=1}^{N} y_n = N/4$).

(a) Find the ML estimate $\hat{x}_{\mathrm{ML}}$ of the parameter x for that particular run.

Let $\mathcal{P}_{\mathsf{z}} \triangleq \{p_{\mathsf{z}}(\cdot; x) : x \in [0,1]\}$ be the set of distributions $p_{\mathsf{z}}(\cdot; x)$ parameterized by x as defined above, and

$$\hat{\mathcal{P}}^{\mathcal{Z}}(\mathbf{y}) \triangleq \left\{ \hat{p}_{\mathsf{z}} : \sum_{z:g(z)=y} \hat{p}_{\mathsf{y}}(z) = \hat{p}_{\mathsf{y}}(y;\mathbf{y}) \quad \text{for all } y \in \mathcal{Y} \right\}$$

be the set of distributions $\hat{p}_{\mathsf{z}}$ that are consistent with the empirical distribution $\hat{p}_{\mathsf{y}}(\cdot;\mathbf{y})$ for the observed sequence $\mathbf{y}$.

(b) Draw $\mathcal{P}_{\mathsf{z}}$ and $\hat{\mathcal{P}}^{\mathcal{Z}}(\mathbf{y})$ on the probability simplex. Mark clearly the points of intersection of the sets with each other and with the simplex boundaries. You can assume that N is large enough so that the set of empirical distributions is dense.

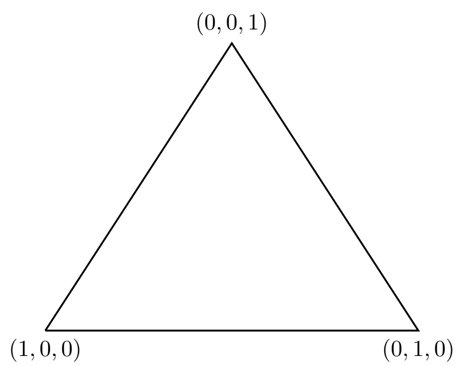

Figure 1: Probability simplex for z, where a point (p_0, p_1, p_2) corresponds to a distribution $(p_{\mathsf{z}}(0), p_{\mathsf{z}}(1), p_{\mathsf{z}}(2))$.

(c) In this part, we estimate the parameter x via alternating projections. Recall that in iteration i of the algorithm ($i = 1, 2, \ldots$), the E-step constructs a distribution

$$\hat{p}_{\mathsf{z}}^{(i)} = \arg\min_{\hat{p}_{\mathsf{z}} \in \hat{\mathcal{P}}^{\mathcal{Z}}(\mathbf{y})} D\left(\hat{p}_{\mathsf{z}} \| p_{\mathsf{z}}\left(\cdot; x^{(i-1)}\right)\right).$$

The M-step of the algorithm then determines the next estimate of the parameter

$$x^{(i)} = \arg\min_{x} D\left(\hat{p}_{\mathsf{z}}^{(i)} \| p_{\mathsf{z}}\left(\cdot; x\right)\right),$$

which in turn defines a distribution $p_{\mathsf{z}}^{(i)} = p_{\mathsf{z}}\left(\cdot; x^{(i)}\right) \in \mathcal{P}_{\mathsf{z}}$.

Suppose we initialize the algorithm with $x^{(0)} = 3/4$.

(i) Determine $\hat{p}_{\mathsf{z}}^{(1)}$.

(ii) Determine $p_{\mathsf{z}}^{(1)}$.

(iii) Will this process converge? If yes, determine the limit distribution $p_{\mathsf{z}}^{(\infty)}$.

(d) Now suppose $x^{(0)} = 0$. Will the algorithm converge? If yes, determine the limit distribution $p_{\mathsf{z}}^{(\infty)}$.

Problem 7.6

Let us develop and analyze the Arimoto-Blahut algorithm for computing model capacity. To start, let $\mathcal{X}$ and $\mathcal{Y}$ be arbitrary finite alphabets. Let $\{q_{\mathsf{y}|\mathsf{x}}(\cdot|x),\ x \in \mathcal{X}\}$ denote the class of models whose capacity we wish to compute. Recall that the model capacity is given by

$$C \triangleq \max_{p_{\mathsf{x}} \in \mathcal{P}^{\mathcal{X}}} I(\mathsf{x};\mathsf{y}) \tag{1}$$

where

$$I(\mathsf{x};\mathsf{y}) = \sum_{x\in\mathcal{X},y\in\mathcal{Y}} p_{\mathsf{x}}(x)\, q_{\mathsf{y}|\mathsf{x}}(y|x) \log \frac{q_{\mathsf{y}|\mathsf{x}}(y|x)}{q_{\mathsf{y}}(y)}, \qquad \text{with} \quad q_{\mathsf{y}}(y) = \sum_{x'\in\mathcal{X}} p_{\mathsf{x}}(x')\, q_{\mathsf{y}|\mathsf{x}}(y|x'),$$

and with $\mathcal{P}^{\mathcal{X}}$ denoting the set of all distributions on $\mathcal{X}$.

Next, let

$$\varphi(p_{\mathsf{x}}, p_{\mathsf{x}|\mathsf{y}}) \triangleq \sum_{x\in\mathcal{X},y\in\mathcal{Y}} p_{\mathsf{x}}(x)\, q_{\mathsf{y}|\mathsf{x}}(y|x) \log \frac{p_{\mathsf{x}|\mathsf{y}}(x|y)}{p_{\mathsf{x}}(x)}$$

for any $p_{\mathsf{x}} \in \mathcal{P}^{\mathcal{X}}$ and $p_{\mathsf{x}|\mathsf{y}} \in \mathcal{P}^{\mathcal{X}|\mathcal{Y}}$, where $\mathcal{P}^{\mathcal{X}|\mathcal{Y}}$ is the set of all conditional distributions for x given y.

(a) Show that for any $p_{\mathsf{x}} \in \mathcal{P}^{\mathcal{X}}$,

$$I(\mathsf{x};\mathsf{y}) = \max_{p_{\mathsf{x}|\mathsf{y}} \in \mathcal{P}^{\mathcal{X}|\mathcal{Y}}} \varphi(p_{\mathsf{x}}, p_{\mathsf{x}|\mathsf{y}}) \tag{2}$$

and that the associated maximizing distribution is

$$p_{\mathsf{x}|\mathsf{y}}(x|y) = \frac{p_{\mathsf{x}}(x)\, q_{\mathsf{y}|\mathsf{x}}(y|x)}{q_{\mathsf{y}}(y)}.$$

Using (2) in (1), note that model capacity can be equivalently expressed as

$$C = \max_{p_{\mathsf{x}} \in \mathcal{P}^{\mathcal{X}}} \max_{p_{\mathsf{x}|\mathsf{y}} \in \mathcal{P}^{\mathcal{X}|\mathcal{Y}}} \varphi(p_{\mathsf{x}}, p_{\mathsf{x}|\mathsf{y}}). \tag{3}$$

The Arimoto-Blahut algorithm evaluates (3) via an alternating maximization procedure. Specifically, starting from some $p_{\mathsf{x}}^{(0)}$, we compute, for $k = 1, 2, \dots,$

$$p_{\mathsf{x}|\mathsf{y}}^{(k)} = \operatorname*{arg\,max}_{p_{\mathsf{x}|\mathsf{y}} \in \mathcal{P}^{\mathcal{X}|\mathcal{Y}}} \varphi(p_{\mathsf{x}}^{(k-1)}, p_{\mathsf{x}|\mathsf{y}}), \tag{4a}$$

$$p_{\mathsf{x}}^{(k)} = \operatorname*{arg\,max}_{p_{\mathsf{x}} \in \mathcal{P}^{\mathcal{X}}} \varphi(p_{\mathsf{x}}, p_{\mathsf{x}|\mathsf{y}}^{(k)}). \tag{4b}$$

Hence, the estimate of capacity after the kth iteration is

$$C_k = \max_{p_{\mathsf{x}} \in \mathcal{P}^{\mathcal{X}}} \varphi(p_{\mathsf{x}}, p_{\mathsf{x}|\mathsf{y}}^{(k)}) = \varphi(p_{\mathsf{x}}^{(k)}, p_{\mathsf{x}|\mathsf{y}}^{(k)}).$$

(b) Determine $\alpha_{k-1}(x)$ such that (4) corresponds to the following update step:

$$p_{\mathsf{x}}^{(k)}(x) = \frac{p_{\mathsf{x}}^{(k-1)}(x)\, e^{\alpha_{k-1}(x)}}{\sum_{x' \in \mathcal{X}} p_{\mathsf{x}}^{(k-1)}(x')\, e^{\alpha_{k-1}(x')}}.$$

Hint: It may be convenient to optimize (4a) and (4b) separately.

In the remainder of the problem, we interpret the Arimoto-Blahut algorithm as an alternating divergence minimization procedure. To this end, we define the sets

$$\mathcal{P} = \left\{ p_{\mathsf{x}}\, q_{\mathsf{y}|\mathsf{x}} \colon p_{\mathsf{x}} \in \mathcal{P}^{\mathcal{X}} \right\}$$
$$\mathcal{Q} = \left\{ p_{\mathsf{x}|\mathsf{y}}\, q_{\mathsf{y}|\mathsf{x}} \colon p_{\mathsf{x}|\mathsf{y}} \in \mathcal{P}^{\mathcal{X}|\mathcal{Y}} \right\}$$

Note that $\mathcal{Q}$ is a set of measures (nonnegative functions), but not distributions, i.e., $p_{\mathsf{x}|\mathsf{y}}(x|y) q_{\mathsf{y}|\mathsf{x}}(y|x)$ is nonnegative but does not sum to one. We can still use divergence to compare measures, and the associated information geometry holds. In particular, the divergence of measure $\nu(x,y)$ from measure $\mu(x,y)$ is

$$D(\mu \| \nu) \triangleq \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} \mu(x,y) \log \frac{\mu(x,y)}{\nu(x,y)}.$$

(c) Show that the Arimoto-Blahut algorithm is an alternating divergence minimization procedure. Specifically, show that

(i) Eq. (4a) is a reverse I-projection from $p_{\mathsf{x}}^{(k-1)} q_{\mathsf{y}|\mathsf{x}}$ into $\mathcal{Q}$.

(ii) Eq. (4b) is an I-projection from $p_{\mathsf{x}|\mathsf{y}}^{(k)}\, q_{\mathsf{y}|\mathsf{x}}$ into $\mathcal{P}$.

(d) Determine which, if either, of the sets $\mathcal{P}$ and $\mathcal{Q}$ are convex.

Problem 7.7 (practice)

Consider N light bulbs whose lifetime is uniformly distributed between 0 and x, where x is an unknown scalar parameter. All N light bulbs are installed in one room and turned on at the same time. You come into the room after time a, (a is a known constant), observe the state of each bulb and then immediately leave the room. The goal is to estimate the lifetime parameter x.

Let z_n be the lifetime of bulb n and y_n be the binary random variable that is equal to 1 if bulb n was still on when you entered the room ($n \in \{1, \ldots, N\}$). We assume that all bulbs are independent of each other.

Suppose when you enter the room, the first $K > 0$ bulbs are on, and the rest are burnt out. In other words, you observe $y_n = 1$ for $n \in \{1, \ldots, K\}$, and $y_n = 0$ for $n \in \{K+1, \ldots, N\}$.

(a) Determine the probability of observing $\mathbf{y} = [y_1, \ldots, y_N]$ defined above.

(b) Determine the maximum likelihood estimate of the parameter x given the observed data $\mathbf{y}$.

In the remainder of this problem, we investigate the performance of the alternating projections algorithm that uses $\mathbf{z} = [\mathsf{z}_1, \ldots, \mathsf{z}_N]$ as full data.

(c) Determine function $f(\cdot)$ such that any empirical probability distribution $\hat{p}_{\mathsf{z}}(\cdot)$ consistent with the observed data must satisfy

$$\mathbb{E}_{\hat{p}_{\mathsf{z}}(\cdot)}\left[f(\mathsf{z})\right] = \frac{K}{N}.$$

Hint: It might be useful to think of the expectation as an integral, i.e., for any distribution $q(\cdot)$,

$$\mathbb{E}_q\left[f(\mathsf{z})\right] = \int_z f(z)\, q(z)\, dz.$$

(d) Suppose we initialize the algorithm with $x^{(0)} > a$. Determine the empirical probability distribution $\hat{p}_{\mathsf{z}}^{(0)}(\cdot)$ that we obtain after performing the E-step of the algorithm, i.e.,

$$\hat{p}_{\mathsf{z}}^{(0)}(\cdot) = \underset{\hat{p}_{\mathsf{z}}(\cdot) \in \hat{\mathcal{P}}^{\mathcal{Z}}}{\arg\min}\, D(\hat{p}_{\mathsf{z}}(\cdot) \| p_{\mathsf{z}}(\cdot; x^{(0)})),$$

where $\hat{\mathcal{P}}^{\mathcal{Z}}$ is the set of all empirical probability distributions $\hat{p}_{\mathsf{z}}(\cdot)$ consistent with the observed data.

(e) Determine the next value of the parameter $x^{(1)}$ that results from performing the M-step of the algorithm, i.e.,

$$x^{(1)} = \arg\min_{x} D(\hat{p}_{\mathsf{z}}^{(0)}(\cdot) \| p_{\mathsf{z}}(\cdot; x)).$$

(f) Will the algorithm converge? If yes, will it converge to the correct estimate?