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

Problem Set 8

Issued: Tuesday, April 28, 2015 **Due:** Tuesday, May 5, 2015

Problem 8.1 (practice)
Let x be a continuous random variable distributed according to a distribution $p(\cdot)$. Bob would like to estimate the probability

$$P(a) = \mathbb{P}(\mathsf{x} \geq a) = \int_{y \geq a} p(y)\, dy.$$

when a is large. It turns out that $p(\cdot)$ is the normal Gaussian distribution with zero mean and unit variance, but Bob does not know this.

In this problem we use the notation $f(t) \doteq e^{\lambda t^2}$ as $t \to \infty$ to mean that

$$\lim_{t \to \infty} \frac{\ln f(t)}{t^2} = \lambda,$$

and use the notation $\dot{\geq}$ and $\dot{\leq}$ analogously.

Hint: In the following parts, you may find the following fact useful: when x is a standard normal random variable, $P(a) \doteq e^{-a^2/2}$ as $a \to \infty$.

(a) First, Bob considers a naïve Monte Carlo approach by generating N i.i.d. samples $\mathbf{x} = [x_1, \ldots, x_N]$ from the distribution p and constructs an estimate of $P(a)$ via

$$\hat{P}_1(\mathbf{x}) = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}_{x_i \geq a},$$

where $\mathbb{1}_{x \geq a}$ is equal to 1 if $x \geq a$ and is 0 otherwise.

(i) Show that for $N \leq 1/(2P(a))$,

$$\mathbb{P}\left(\hat{P}_1(\mathsf{x}) = 0\right) \geq \frac{1}{2},$$

and conclude that for such values of N

$$\mathbb{P}\left(|\hat{P}_1(\mathsf{x}) - P(a)| \geq P(a)\right) \geq \frac{1}{2},$$

i.e., with high probability the relative error in the estimate will be large.

Hint: Use the inequality

$$(1-b)^c \geq 1 - cb, \quad \text{for all } b \in [0,1] \text{and } c \geq 1.$$

(ii) From part (a)-i., we know that for $\hat{P}_1(\mathbf{x})$ to be a useful approximation of $P(a)$, we need at least $N = N(a) \geq 1/(2P(a))$. Show that this implies $N(a) \stackrel{.}{\geq} e^{a^2/2}$ as $a \to \infty$, i.e., the number of samples must grow at least exponentially with a^2 if Bob uses a naïve Monte Carlo approach.

In order to estimate $P(a)$ with fewer samples, Bob uses an importance-sampling approach instead of the naïve Monte Carlo method. In this alternative approach, N i.i.d. samples $\mathbf{y} = [y_1, \ldots, y_N]$ generated from a mean a, unit variance Gaussian distribution q to construct an importance-sampling estimate of $P(a)$ via

$$\hat{P}(\mathbf{y}) = \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}_{y_i \geq a} \frac{p(y_i)}{q(y_i)}.$$

(b) Show that

$$\mathbb{E}\big[\hat{P}(\mathbf{y})\big] = P(a),$$

and

$$\mathrm{var}\big[\hat{P}(\mathbf{y})\big] = \frac{1}{N}\big(\alpha F(a) + \beta P(a)^2\big)$$

where

$$F(a) \triangleq \int_{y \geq a} \frac{p(y)^2}{q(y)} dy,$$

and α and β are constants that do not depend on a or N. Determine α and β.

(c) Show that $F(a) \doteq e^{\gamma a^2}$ as $a \to \infty$ for a constant γ that does not depend on a, and determine γ.

(d) By Chebyshev's inequality and using that $\hat{P}(\mathbf{y})$ is unbiased from part (a), we have that for any small constant $\delta > 0$,

$$\mathbb{P}\left(|\hat{P}(\mathbf{y}) - P(a)| \geq \delta P(a)\right) \leq \frac{\mathrm{var}\big[\hat{P}(\mathbf{y})\big]}{\delta^2 P(a)^2}.$$

Show using parts (b) and (c) that when we choose $N = N(a)$ such that

$$\frac{\mathrm{var}\big[\hat{P}(\mathbf{y})\big]}{\delta^2 P(a)^2} = \delta,$$

it follows that $N(a) \stackrel{.}{\leq} e^{\kappa a^2}$ where $\kappa = 0$.

The fact that $\kappa = 0$ shows that the number of samples needed for a reasonable estimate of $P(a)$ using importance sampling grows at most subexponentially in a^2.

Problem 8.2
Recall that the set of ϵ-typical sequences $\mathbf{x} = (x_1, \ldots, x_N)$ with respect to the distribution p_{x} over alphabet $\mathcal{X}$ is

$$\mathcal{T}_{\mathsf{x}}(\epsilon) = \left\{ \mathbf{x} \in \mathcal{X}^N \colon \left| \frac{1}{N} \log p_{\mathbf{x}}^N(\mathbf{x}) + H(\mathsf{x}) \right| \leq \epsilon \right\}, \qquad \text{with} \quad p_{\mathbf{x}}^N(\mathbf{x}) = \prod_{n=1}^{N} p_{\mathsf{x}}(x_n),$$

where $H(\mathsf{x})$ is the entropy associated with the distribution p_{x}.

More generally, we say a pair of sequences $\mathbf{x} = (x_1, \ldots, x_N)$ and $\mathbf{y} = (y_1, \ldots, y_N)$ are ϵ-*jointly*-typical with respect to the joint distribution $p_{\mathsf{x},\mathsf{y}}$ if the super-symbol sequence $\mathbf{z} = (\mathbf{x}, \mathbf{y})$ is ϵ-typical with respect to $p_{\mathsf{z}} = p_{\mathsf{x},\mathsf{y}}$, *and* if $\mathbf{x}$ and $\mathbf{y}$ are each ϵ-typical with respect to the corresponding marginals p_{x} and p_{y}. With a slight abuse of notation, we use $\mathcal{T}_{\mathsf{x},\mathsf{y}}(\epsilon)$ to denote the associated ϵ-jointly-typical set.

In this problem, $H(\mathsf{x}, \mathsf{y})$ denotes the joint entropy of x and y, and $I(\mathsf{x}; \mathsf{y})$ denotes the mutual information between x and y, where x and y follow a fixed joint distribution $p_{\mathsf{x},\mathsf{y}}(x, y)$.

The ϵ-jointly-typical set and its sequences satisfy the following properties, which you may find useful:

(i) If $(\mathbf{x}, \mathbf{y})$ is an i.i.d. sequence of length N with elements distributed according to $p_{\mathsf{x},\mathsf{y}}$, then $\mathbb{P}\left((\mathbf{x}, \mathbf{y}) \in \mathcal{T}_{\mathsf{x},\mathsf{y}}(\epsilon)\right) \geq 1 - \epsilon$, for N sufficiently large;

(ii) For any $(\mathbf{x}, \mathbf{y}) \in \mathcal{T}_{\mathsf{x},\mathsf{y}}(\epsilon)$ we have

$$2^{-N(H(\mathsf{x},\mathsf{y})+\epsilon)} \leq p_{\mathbf{x},\mathbf{y}}^N(\mathbf{x}, \mathbf{y}) \leq 2^{-N(H(\mathsf{x},\mathsf{y})-\epsilon)}, \qquad \text{with} \quad p_{\mathbf{x},\mathbf{y}}^N(\mathbf{x}, \mathbf{y}) = \prod_{n=1}^{N} p_{\mathsf{x},\mathsf{y}}(x_n, y_n);$$

(iii) $(1 - \epsilon)2^{N(H(\mathsf{x},\mathsf{y})-\epsilon)} \leq |\mathcal{T}_{\mathsf{x},\mathsf{y}}(\epsilon)| \leq 2^{N(H(\mathsf{x},\mathsf{y})+\epsilon)}$, where the lower bound works for sufficiently large N and the upper bound is valid for all N.

Finally, the ϵ-*conditionally*-typical set $\mathcal{T}_{\mathsf{y}|\mathsf{x}}(\epsilon, \mathbf{x})$ corresponding to a sequence $\mathbf{x} \in \mathcal{T}_{\mathsf{x}}(\epsilon)$, is defined as the subset of sequences in $\mathcal{T}_{\mathsf{y}}(\epsilon)$ that are ϵ-jointly-typical with the sequence $\mathbf{x}$, i.e.,

$$\mathcal{T}_{\mathsf{y}|\mathsf{x}}(\epsilon, \mathbf{x}) = \left\{ \mathbf{y} \in \mathcal{T}_{\mathsf{y}}(\epsilon) \colon (\mathbf{x}, \mathbf{y}) \in \mathcal{T}_{\mathsf{x},\mathsf{y}}(\epsilon) \right\}.$$

(a) Determine a finite α (that is not a function of ϵ or N) such that for all sequences $\mathbf{y} \in \mathcal{T}_{\mathsf{y}|\mathsf{x}}(\epsilon, \mathbf{x})$ and any $\mathbf{x} \in \mathcal{T}_{\mathsf{x}}(\epsilon)$,

$$\left| \frac{1}{N} \log p_{\mathbf{y}|\mathbf{x}}^N(\mathbf{y}|\mathbf{x}) + H(\mathsf{y}|\mathsf{x}) \right| \leq \alpha\epsilon, \qquad \text{with} \quad p_{\mathbf{y}|\mathbf{x}}^N(\mathbf{y}|\mathbf{x}) = \frac{p_{\mathbf{x},\mathbf{y}}^N(\mathbf{x}, \mathbf{y})}{p_{\mathbf{x}}^N(\mathbf{x})},$$

where $H(\mathsf{y}|\mathsf{x})$ denotes the conditional entropy of y given x.

(b) Determine a finite β (that is not a function of ϵ or N) such that for any $\mathbf{x} \in \mathcal{T}_{\mathsf{x}}(\epsilon)$,

$$\left|\mathcal{T}_{\mathsf{y}|\mathsf{x}}(\epsilon, \mathbf{x})\right| \leq 2^{N(H(\mathsf{y}|\mathsf{x})+\beta\epsilon)}.$$

(c) Let $\tilde{\mathbf{x}} = (\tilde{x}_1, \ldots, \tilde{x}_N)$ and $\tilde{\mathbf{y}} = (\tilde{y}_1, \ldots, \tilde{y}_N)$ be realizations of independent sequences each generated i.i.d. according to marginals p_{x} and p_{y}, respectively, i.e., realizations $(\tilde{\mathbf{x}}, \tilde{\mathbf{y}})$ from the joint distribution that is i.i.d. with

$$p_{\tilde{\mathsf{x}},\tilde{\mathsf{y}}}(\tilde{x}, \tilde{y}) = p_{\mathsf{x}}(\tilde{x})\, p_{\mathsf{y}}(\tilde{y}), \qquad \tilde{x} \in \mathcal{X},\ \tilde{y} \in \mathcal{Y}.$$

Determine a finite γ (that is not a function of ϵ or N) such that

$$\mathbb{P}\left((\tilde{\mathbf{x}}, \tilde{\mathbf{y}}) \in \mathcal{T}_{\mathsf{x},\mathsf{y}}(\epsilon)\right) \leq 2^{-N(I(\mathsf{x};\mathsf{y})-\gamma\epsilon)},$$

where $I(\mathsf{x};\mathsf{y})$ denotes the mutual information associated with $p_{\mathsf{x},\mathsf{y}}(x, y)$.
Hint: Please recall carefully the exact definition of ϵ-jointly-typical sequences.

In the remainder of the problem, consider a classification problem with m classes. Each class H_i is defined by a feature vector $\mathbf{x}^{(i)} \triangleq (x_1^{(i)}, \ldots, x_N^{(i)}) \in \mathcal{X}^N$. The m feature vectors can be modeled as being independently drawn at random (with replacement) from $\mathcal{X}^N$ according to the same distribution $p_{\mathsf{x}}^N\left(\mathbf{x}^{(i)}\right) = \prod_{n=1}^N p_{\mathsf{x}}\left(x_n^{(i)}\right)$. The prior over the m classes is governed by the distribution $p_{\mathsf{H}}(\cdot)$, where $p_{\mathsf{H}}(H_i)$ denotes the probability that the true class $\mathsf{H} = H_i$. The true class H is independent with all feature vectors $\mathbf{x}^{(i)}$.

Under the condition that the unknown true class $\mathsf{H} = H_i$, the observation $\mathbf{y} = (y_1, \ldots, y_N)$ will be generated by the feature vector $\mathbf{x}^{(i)}$ according to

$$p_{\mathbf{y}|\mathbf{x},\mathsf{H}}\left(\mathbf{y}|\mathbf{x}^{(i)}, H_i\right) = \prod_{n=1}^N p_{\mathsf{y}|\mathsf{x}}(y_n|x_n^{(i)}).$$

With the observation $\mathbf{y}$ (and the feature vectors $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(m)}$) we decide which class is being observed using the rule

$$\hat{\mathsf{H}}(\mathbf{y}) = H_i \quad \text{if} \quad (\mathbf{x}^{(i)}, \mathbf{y}) \in \mathcal{T}_{\mathsf{x},\mathsf{y}}(\epsilon),$$

where $i \in \{1, \ldots, m\}$.

The decision device declares an error if the observed sequence $\mathbf{y}$ is ϵ-jointly-typical either with more than one feature vector, or with no feature vectors.

(d) Determine a finite δ (that is not a function of ϵ or N) such that for sufficiently large N,

$$\mathbb{P}\left(\text{error} \;\middle|\; \mathsf{H} = H_1\right) \leq (m-1)2^{-N(I(\mathsf{x};\mathsf{y})-\delta\epsilon)} + \epsilon.$$

Hint: You may find useful the union bound $\mathbb{P}\left(\cup_{i=1}^n \mathcal{A}_i\right) \leq \sum_{i=1}^n \mathbb{P}\left(\mathcal{A}_i\right)$ for arbitrary collections of events $\mathcal{A}_1, \ldots, \mathcal{A}_n$. In addition, you may find useful the independence between $\mathbf{y}$ and $\mathbf{x}^{(i)}$ conditioned on $\mathsf{H} = H_1$, where $i \geq 2$.

(e) Assume $I(\mathsf{x};\mathsf{y}) > 0$. Show that the largest number of classes that can be distinguished with vanishing probability of an error grows exponentially with N. In other words, it is possible to distinguish $m = 2^{NR}$ classes where $R > 0$ is a constant, and $\mathbb{P}\,(\text{error}) \to 0$ as $\epsilon \to 0$ and $N \to \infty$.

Problem 8.3

Let $\mathsf{y}_1, \mathsf{y}_2, \ldots$ be i.i.d. random variables drawn according to the geometric distribution

$$\mathbb{P}\,(\mathsf{y} = k) = p^{k-1}(1-p), \quad k = 1, 2, \ldots.$$

Find good estimates (to first order in the exponent) of:

(a) $\mathbb{P}\left((1/N) \cdot \sum_{n=1}^{N} \mathsf{y}_n \geq \alpha\right)$, where $\alpha > 1/(1-p)$.

(b) $\mathbb{P}\left(\mathsf{y}_1 = k \,\middle|\, (1/N) \cdot \sum_{n=1}^{N} \mathsf{y}_n \geq \alpha\right)$, where $\alpha > 1/(1-p)$.
Hint: Please feel free to apply the conditional limit theorem to this problem, even if the size of the alphabet is infinite here.

(c) Evaluate parts (a) and (b) for $p = 1/2$, $\alpha = 4$.

Problem 8.4

Let $\mathbf{y} = [\mathsf{y}_1, \ldots, \mathsf{y}_N]^{\mathrm{T}}$ be a vector of N i.i.d. binary variables distributed according to a Bernoulli distribution with a known parameter $x \in [0, 1]$, i.e., $\mathbb{P}\,(\mathsf{y}_n = 1) = x$. Let $\gamma \in (0, 1)$ be a known constant.

(a) Find all values of x for which

$$\mathbb{P}\left(\frac{1}{N}\sum_{n=1}^{N} \mathsf{y}_n \geq \gamma\right) \doteq 1.$$

We now model the parameter of the Bernoulli distribution as a random variable x distributed uniformly over the unit interval $[0, 1]$. Note that while $\mathsf{y}_1, \ldots, \mathsf{y}_N$ are *conditionally* independent given x, this change in the model makes $\mathsf{y}_1, \ldots, \mathsf{y}_N$ dependent.

(b) Determine the exponent $\beta(\gamma)$ such that

$$\mathbb{P}\left(\frac{1}{N}\sum_{n=1}^{N} \mathsf{y}_n \geq \gamma\right) \doteq e^{-N\beta(\gamma)}.$$

(c) Does your answer to part (b) depend on the prior distribution on x? Explain.

Problem 8.5 (practice)
Suppose that $y_1, y_2, \ldots, y_N$ are i.i.d. Poisson random variables with parameter $x > 0$, so that each y_i has distribution

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

Recall that a Poisson random variable with parameter x has mean x and variance x.

(a) One estimator for x is the sample mean:

$$\hat{x}(\mathbf{y}) = \frac{1}{N} \sum_{i=1}^{N} y_i.$$

The asymptotic efficiency in probability of $\hat{x}$ can be expressed via

$$\delta(\epsilon, x) = \mathbb{P}\left(|\hat{x} - x| > \epsilon x\right) \doteq e^{-Nx\xi(\epsilon)},$$

where $0 < \epsilon < 1$. Determine $\xi(\epsilon)$, and show that $\xi(\epsilon) \approx \epsilon^2/2$ for small ϵ.

(b) Suppose that x is known to be one of two equally likely values, a or ae, where a is a known real positive number. The best achievable error probability for deciding the correct value is, asymptotically to first order in the exponent, $P_{\mathrm{e}} \doteq e^{-Na\kappa}$. Determine the constant κ. Show that $\kappa < 1$.

Problem 8.6
Let $y_1, \ldots, y_N$ be a sequence of i.i.d. discrete random variables. Consider the following binary hypothesis testing framework, where $0 < \epsilon < 1$ and $\epsilon \neq 1/2$.

$$H_0 : \quad p_{y_i}(y_i) = p_0(y_i) = \begin{cases} \epsilon & y_i = 1 \\ 1 - \epsilon & y_i = 0 \end{cases} \quad \text{for each } i$$

$$H_1 : \quad p_{y_i}(y_i) = p_1(y_i) = \begin{cases} 1 - \epsilon & y_i = 1 \\ \epsilon & y_i = 0 \end{cases} \quad \text{for each } i$$

(a) If we constrain the detection probability $P_{\mathrm{D}} \geq 0.99$ and want to minimize the false-alarm probability P_{F}, determine the exponent that governs the asymptotic decay of P_{F} with N.

(b) Suppose instead that we fix $P_{\mathrm{F}} \leq 0.01$ and want to maximize P_{D}. What is the exponent governing the rate at which $P_{\mathrm{D}} \to 1$?

(c) Consider the Bayesian case where there are non-zero priors on the two hypotheses. Asymptotically the probability of error P_e can be made to decay exponentially. Determine the largest achievable exponent, as a function of ϵ.

(d) How do the normalized exponents (divide by $(1-2\epsilon)^2$) compare in parts (a), (b), and (c) in the limit as $\epsilon \to 1/2$?

Problem 8.7 (practice)
Consider the binary hypothesis testing problem, with observations $y_1, \ldots, y_N$:

$$H_0 : (y_1, \ldots, y_N) \overset{\text{i.i.d.}}{\sim} p_0, \qquad H_1 : (y_1, \ldots, y_N) \overset{\text{i.i.d.}}{\sim} p_1,$$

for $p_0, p_1 \in \mathcal{P}(\mathcal{Y})$ with $2 \leq |\mathcal{Y}| < \infty$. Define the decision region

$$\mathcal{R}_N \triangleq \{\mathbf{y} \in \mathcal{Y}^N : \hat{H}(\mathbf{y}) = H_1\},$$

i.e., $\mathbf{y} = (y_1, \ldots, y_N) \in \mathcal{R}_N$ results in deciding H_1. Similarly, define the complement

$$\mathcal{Y}^N \setminus \mathcal{R}_N \triangleq \{\mathbf{y} \in \mathcal{Y}^N : \hat{H}(\mathbf{y}) = H_0\}.$$

Hence, the probabilities of false alarm and missed detection are, respectively,

$$P_F = P_0\{\mathcal{R}_N\} \triangleq \sum_{\mathbf{y} \in \mathcal{R}_N} p_0^N(\mathbf{y}), \qquad P_M = P_1\{\mathcal{Y}^N \setminus \mathcal{R}_N\} \triangleq \sum_{\mathbf{y} \in \mathcal{Y}^N \setminus \mathcal{R}_N} p_1^N(\mathbf{y}).$$

In this problem, we analyze and apply the *Hoeffding Test* defined by

$$\mathcal{R}_N^* \triangleq \left\{\mathbf{y} \in \mathcal{Y}^N : D(\hat{q}(\cdot\,;\mathbf{y}) \,\|\, p_0(\cdot)) \geq \lambda\right\},$$

where $\hat{q}(\cdot\,;\mathbf{y})$ is the type (empirical distribution) of $\mathbf{y}$ and $\lambda > 0$ is a constant.

(a) Show that under the Hoeffding test, P_F decays exponentially with the rate satisfying

$$P_F \,\dot{\leq}\, \exp(-N\lambda). \tag{1}$$

(b) Let the decision region $\tilde{\mathcal{R}}_N$ correspond to any test such that (1) is satisfied. In addition, $\tilde{\mathcal{R}}_N$ satisfies that if $\mathbf{y} \in \tilde{\mathcal{R}}_N$, then all observations in the type class $T(\hat{q}(\cdot\,;\mathbf{y}))$ also belongs to $\tilde{\mathcal{R}}_N$. Show $P_1\{\mathcal{Y}^N \setminus \mathcal{R}_N^*\} \,\dot{\leq}\, P_1\{\mathcal{Y}^N \setminus \tilde{\mathcal{R}}_N\}$.

Hint: Show that for any $\epsilon > 0$, we have $D(\hat{q}(\cdot\,;\mathbf{y}) \,\|\, p_0(\cdot)) \geq (\lambda - \epsilon)$ for $\mathbf{y} \in \tilde{\mathcal{R}}_N$ with a sufficiently large N.

(c) The Hoeffding test satisfies

$$P_M = P_1\{\mathcal{Y}^N \setminus \mathcal{R}_N^*\} \,\dot{\leq}\, \exp(-NJ(\lambda)). \tag{2}$$

for some $J(\lambda) > 0$. Express $J(\lambda)$ as an I-projection.

(d) In this part, we apply the Hoeffding test to an anomaly detection problem. Suppose we have $M = \lfloor \exp(NR) \rfloor$ sequences each of length N, denoted as $\mathbf{y}_1, \ldots, \mathbf{y}_M$. For some unknown $a \in \{1, \ldots, M\}$ each element in $\mathbf{y}_a$ is drawn from p_0, while all elements in $\mathbf{y}_1, \ldots, \mathbf{y}_{a-1}, \mathbf{y}_{a+1}, \ldots, \mathbf{y}_M$ are drawn from p_1 with each draw made independently. We estimate a from $\mathbf{y}_1, \ldots, \mathbf{y}_M$ using the following rule:

> If there exists a *unique* $a \in \{1, \ldots, M\}$ such that $D(\hat{q}(\cdot\,; \mathbf{y}_a) \,\|\, p_0(\cdot)) < \lambda$, set $\hat{a} = a$. If no such unique a exists, set $\hat{a} = 0$.

For this decision rule, determine an $R > 0$ in terms of λ such that $\mathbb{P}(\hat{a} \neq a) \to 0$ as $N \to \infty$. For simplicity assume that $a = 1$ in your analysis.
Hint: Decompose the error event, use the union bound $\mathbb{P}(\cup_{i=1}^{L} \mathcal{B}_i) \leq \sum_{i=1}^{L} \mathbb{P}(\mathcal{B}_i)$, and use (1) and (2).