Homework 1

Due: Wednesday, September 28, 2016

Notes: For positive integers k, $[k] := \{1, \ldots, k\}$ denotes the set of the first k positive integers. When $X \sim p$ and $Y \sim q$ are random variables over the same sample space, $D(X||Y)$, $D(X||q)$, and $D(p||Y)$ should all be read as $D(p||q)$. The homework is out of 60 points.

1. **Warm-up Problems**

   (a) (15 points) Two teams A and B play a best-of-five series that terminates as soon as one of the teams wins three games. Let X be the random variable representing the outcome of the series, written as a string of who won the individual games (e.g., possible values of X are AAA, $BAAA$, $ABABB$, etc.) Let Y be the number of games played before the series ends. Assuming that A and B are equally matched and the outcomes of different games in the series are independent, calculate $H(X)$, $H(Y)$, $H(Y|X)$, $H(X|Y)$, and $I(X;Y)$. Let p_A and q_A be the distributions of X and Y, respectively, given that A wins the series. Calculate $D(p_A||X)$ and $D(q_A||Y)$.

   (b) (5 points) Suppose X, Y, and Z are each Bernoulli(1/2) and are pairwise independent (i.e., $I(X;Y) = I(Y;Z) = I(X;Z) = 0$). What is the minimum possible value of $H(X,Y,Z)$?

2. **General Data Processing**

   (a) **(10 points)** Suppose we have two distributions p_1 and p_2 on $[k]$, and, for each $i \in [k]$, a conditional distribution q_i over $[\ell]$. Let $q_1(j) = \sum_{i=1}^{k} q_i(j)p_1(i)$ and $q_2(j) = \sum_{i=1}^{k} q_i(j)p_2(i)$ denote the marginal distributions over $[\ell]$ induced by p_1 and p_2, respectively. Prove the General Data Processing Inequality

   $$D(q_1||q_2) \leq D(p_1||p_2). \tag{1}$$

   Hint: Use the log-sum inequality, which states that, for all non-negative sequences $a_1, \ldots, a_n$ and $b_1, \ldots, b_n$, letting $a = \sum_{i=1}^{n} a_i$ and $b = \sum_{i=1}^{n} b_i$,

   $$\sum_{i=1}^{n} a_i \log \frac{a_i}{b_i} \geq a \log \frac{a}{b}.$$

   (b) As special cases of (1), show:

   i. **(5 points)** For random variables X and Y taking values in $[k]$ and function f with domain $[k]$,

   $$D(f(X)||f(Y)) \leq D(X||Y) \quad \text{and} \quad H(f(X)) \leq H(X).$$

   ii. **(5 points)** The Data Processing Inequality from class: for a Markov chain $X \mapsto Y \mapsto Z$, $I(X;Z) \leq I(X;Y)$.

3. **Plug-in estimator for differential entropy**

This problem derives convergence rates for an estimator of the differential entropy $H(p) = -\int_{\mathcal{X}} p(x) \log p(x)\, dx$ of a probability density p, given n IID samples $X_1, \ldots, X_n \sim p$. To simplify matters, we will make the following assumptions:

i) The sample space $\mathcal{X} = [0,1]^D$ is the D-dimensional unit cube.

ii) We know positive lower and upper bounds

$$0 < \kappa_1 \leq \inf_{x \in \mathcal{X}} p(x) \leq \sup_{x \in \mathcal{X}} p(x) \leq \kappa_2 < \infty$$

on the true density p.

The estimator in question is a plug-in estimator based on a truncated kernel density estimate (KDE). Specifically, the estimate $\widehat{H}_h$ is given by given by

$$\widehat{H}_h = H(\widehat{p}_h) = -\int_{\mathcal{X}} \widehat{p}_h(x) \log \widehat{p}_h(x)\, dx, \tag{2}$$

where, for some bandwidth $h > 0$ and kernel $K : \mathbb{R}^D \to \mathbb{R}$ with $\int_{\mathbb{R}^D} K(u)\, du = 1$,

$$\widehat{p}_h(x) = \min\left\{\kappa_2, \max\left\{\kappa_1, \frac{1}{nh^d}\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right)\right\}\right\}, \tag{3}$$

is a truncated KDE of p.

You may take for granted the following facts about the integrated squared bias and variance of the truncated KDE: [1] there exist constants $C_0, C_1 > 0$ such that, for all $h > 0$,

$$\int_{\mathcal{X}} \left(\mathbb{E}\left[\widehat{p}_h(x)\right] - p(x)\right)^2\, dx \leq C_0 h^{2\beta} \tag{4}$$

and

$$\int_{\mathcal{X}} \mathbb{V}\left[\widehat{p}_h(x)\right]\, dx \leq \frac{C_1}{nh^D}. \tag{5}$$

Here, the "Hölder" parameter $\beta > 0$ is a measure of smoothness of the probability density p. Larger β indicates smoother p, and hence less smoothing bias. The standard decomposition of mean-squared error into bias and variance gives

$$\int_{\mathcal{X}} \mathbb{E}\left[\left(\widehat{p}_h(x) - p(x)\right)^2\right] = \int_{\mathcal{X}} \left(\mathbb{E}\left[\widehat{p}_h(x)\right] - p(x)\right)^2 + \mathbb{V}\left[\widehat{p}_h(x)\right]\, dx \leq C_0 h^{2\beta} + \frac{C_1}{nh^D}.$$

Optimizing over h gives the rate $h \asymp n^{-\frac{1}{2\beta+D}}$ and plugging this back in gives the integrated MSE rate

$$\int_{\mathcal{X}} \mathbb{E}\left[\left(\widehat{p}_h(x) - p(x)\right)^2\right] \asymp n^{-\frac{2\beta}{2\beta+D}}.$$

In this problem, we will derive similar bounds for the plug-in entropy estimator, and study its optimal bandwidth and MSE.

[1] These results can be found in any text on nonparametric estimation, such as Tsybakov [2008], Section 1.2.

(a) **(5 points)** Prove the bias bound

$$\left|\mathbb{E}\left[\widehat{H}_h\right] - H\right| \leq C_B\left(h^\beta + h^{2\beta} + \frac{1}{nh^D}\right),$$

for some C_B depending only on C_0, C_1, κ_1, κ_2, and D. *(Hint: Along with inequalities* (4) *and* (5)*, a second-order Taylor expansion and Jensen's inequality may be useful.)*

(b) **(5 points)** This part will use McDiarmid's inequality:

Theorem 1. (McDiarmid's Inequality): *Suppose we have n independent random variables $X_1, \ldots, X_n$ taking values in a set Ω and a function $f : \Omega \to \mathbb{R}$ such that, for some constants $c_1, \ldots, c_n$,*

$$\sup_{x_1,\ldots,x_n,y\in\Omega} |f(x_1,\ldots,x_n) - f(x_1,\ldots,x_{i-1},y,x_{i+1},\ldots,x_n)| \leq c_i, \quad \text{for each } i \in [n].$$

Then, McDiarmid's inequality states that, for any $\varepsilon > 0$,

$$\mathbb{P}\left[|f(X) - \mathbb{E}\left[f(X)\right]| > \varepsilon\right] \leq 2\exp\left(-\frac{2\varepsilon^2}{\sum_{i=1}^n c_i^2}\right).$$

Essentially, if a function depends on many independent random variables, but not too much on any one of them, McDiarmid's inequality tells us that the function's distribution is is tightly concentrated around its expectation.

Use McDiarmid's inequality to derive the exponential concentration bound

$$\mathbb{P}\left[\left|\widehat{H}_h - \mathbb{E}\left[\widehat{H}_h\right]\right| > \varepsilon\right] \leq 2\exp\left(-C_E\varepsilon^2 n\right), \tag{6}$$

for the plug-in estimator $\widehat{H}_h$, for some C_E depending only on D, K, κ_1, and κ_2. *(Hint: The mean value theorem will be useful here.)*

(c) **(5 points)** Use (6) to prove the variance bound $\mathbb{V}\left[\widehat{H}\right] \leq \frac{C_V}{n}$, with C_V depending only on D, K, κ_1, and κ_2. *(Hint: Recall that, for a non-negative random variable X, $\mathbb{E}[X] = \int_0^\infty \mathbb{P}[X > x]\, dx$.)*

(d) **(5 points)** Combine the bias and variance bounds to derive a bound on the mean squared error (MSE) $\mathbb{E}\left[\left(\widehat{H}_h - H\right)^2\right]$ of $\widehat{H}_h$. Optimize this over h. What are the optimal bandwidth and MSE rates (asymptotically, as $n \to \infty$)? How do these compare to the optimal bandwidth and MSE rates for kernel density estimation (smaller, same, or larger)?

References

A.B. Tsybakov. *Introduction to Nonparametric Estimation.* Springer Publishing Company, Incorporated, 1st edition, 2008. ISBN 0387790519, 9780387790510.