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 $I(X;Y)$ (in bits). 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 $H(X,Y,Z)$?

   Solution:

   (a) There are

   - 2 ways the series can have length 3, each with probability 2^{-3}.
   - $2\binom{3}{1} = 6$ ways the series can have length 4, each with probability 2^{-4}.
   - There are $2\binom{4}{2} = 12$ ways the series can have length 5, each with probability 2^{-5}.

   Hence,

   $$H(X) = 2 \cdot 2^{-3} \log_2 2^3 + 6 \cdot 2^{-4} \log_2 2^4 + 12 \cdot 2^{-5} \log_2 2^5 = \boxed{\frac{33}{8}},$$

   and

   $$H(Y) = \frac{1}{4} \log_2 4 + 2 \cdot \frac{3}{8} \log_2 \frac{8}{3} = \boxed{\tfrac{1}{2} + \tfrac{3}{4} \log_2 \tfrac{8}{3} \approx 1.56}.$$

   Since Y is a (deterministic) function of X, $\boxed{H(Y|X) = 0}$,

   $$I(X;Y) = H(Y) = \boxed{\frac{1}{2} + \frac{3}{4} \log_2 \frac{8}{3} \approx 1.56},$$

   and

   $$H(X|Y) = H(X) - H(Y) = \boxed{\frac{29}{8} - \frac{3}{4} \log_2 \frac{8}{3} \approx 2.56.}$$

   Since $p_A(x)$ is precisely twice $\mathbb{P}(X = x)$ wherever p_A is supported, we have $D(p_A||X) = \mathbb{E}_{X' \sim p_A}[\log_2(2)] = \boxed{1}$. Finally, since Y is independent of whether A wins the series, q_A is identical to the distribution of Y, and so $\boxed{D(q_A||Y) = 0}$.

(b) Applying the Chain Rule (twice), the fact that Shannon entropy is nonnegative, and the fact that $H(Y|Z) = H(Y)$ (since Y and Z are independent)

$$\begin{aligned} H(X,Y,Z) &= H(X|Y,Z) + H(Y|Z) + H(Z) \\ &\geq H(Y|Z) + H(Z) = H(Y) + H(Z) = 2. \end{aligned}$$

This is achieved, for example, if $Z = X \oplus Y$ (where $\oplus$ denotes the exclusive or operation), since, in this case, any of $\{X, Y, Z\}$ is a function of the remaining two.

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 $X \mapsto Y \mapsto Z$, $I(X;Z) \leq I(X;Y)$.

Solution:

(a) For each $i \in [k]$ and $j \in [\ell]$, define $a_{j,i} := q_i(j)p_1(i)$ and $b_{j,i} := q_i(j)p_2(i)$. Note that $q_1(j) = \sum_{i=1}^{k} a_{j,i}$, and $q_2(j) = \sum_{i=1}^{k} b_{j,i}$. Hence, applying the log-sum inequality (to each term of the summation),

$$\begin{aligned} D\left(q_1||q_2\right) = \sum_{j=1}^{\ell} q_1(j) \log \frac{q_1(j)}{q_2(j)} &\leq \sum_{j=1}^{\ell} \sum_{i=1}^{k} a_{j,i} \log \frac{a_{j,i}}{b_{j,i}} \\ &= \sum_{i=1}^{k} \sum_{j=1}^{\ell} q_i(j)p_1(i) \log \frac{p_1(i)}{p_2(i)} \\ &= \sum_{i=1}^{k} p_1(i) \log \frac{p_1(i)}{p_2(i)} = D\left(p_1||p_2\right), \end{aligned}$$

since, for each $i \in [k]$, $\sum_{j=1}^{\ell} q_i(j) = 1$.

(b) i. The inequality

$$D(f(X)||f(Y)) \leq D(X||Y) \tag{2}$$

is precisely (1) in the case that q_i is a delta function $q_i(j) = 1_{\{j=f(i)\}}$ at $f(i)$. Since $H(f(X)|X) = 0$,

$$H(f(X)) = I(X; f(X)) = D(P_{(X,f(X))}||P_X P_{f(X)}).$$

Applying (2) with the bivariate function $(x, y) \mapsto (x, f(y))$ gives

$$H(f(X)) = D(P_{(X,f(X))}||P_X P_{f(X)}) \leq D(P_{(X,X)}||P_X P_X) = I(X; X) = H(X).$$

ii. Recall that, in general, $I(X; Y)$ is the expected divergence in the distribution of Y when given X. Applying the law of total probability, the fact that $P_{Z|X,Y} = P_{Z|Y}$, and inequality (1),

$$\begin{aligned} I(X; Z) = \mathbb{E}_X \left[D\left(P_{Z|X} \middle\| P_Z\right)\right] &= \mathbb{E}_X \left[D\left(\int_{\mathcal{Y}} P_{Z|X,Y} P_{Y|X} \middle\| \int_{\mathcal{Y}} P_{Z|Y} P_Y\right)\right] \\ &= \mathbb{E}_X \left[D\left(\int_{\mathcal{Y}} P_{Z|Y} P_{Y|X} \middle\| \int_{\mathcal{Y}} P_{Z|Y} P_Y\right)\right] \\ &\leq \mathbb{E}_X \left[D\left(P_{Y|X} \middle\| P_Y\right)\right] = I(X; Y). \end{aligned}$$

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

$$\widehat{H}_h = H(\widehat{p}_h) = -\int_{\mathcal{X}} \widehat{p}_h(x) \log \widehat{p}_h(x)\, dx, \tag{3}$$

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

$$\widehat{p}_h(x) = \min\left\{\kappa_2, \max\left\{\kappa_1, \frac{1}{nh^d} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right)\right\}\right\}, \tag{4}$$

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 $h > 0$,

$$\int_{\mathcal{X}} \left(\mathbb{E}\left[\widehat{p}_h(x)\right] - p(x)\right)^2 \, dx \leq C_0 h^{2\beta} \tag{5}$$

and

$$\int_{\mathcal{X}} \mathbb{V}\left[\widehat{p}_h(x)\right] \, dx \leq \frac{C_1}{nh^D}. \tag{6}$$

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.

(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* (5) *and* (6)*, 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 $\varepsilon > 0$,

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

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

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

$$\mathbb{P}\left[\left|\widehat{H}_h - \mathbb{E}\left[\widehat{H}_h\right]\right| > \varepsilon\right] \leq 2\exp\left(-C_E\varepsilon^2 n\right), \tag{7}$$

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 (7) 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)?

Note: When initially writing this problem, I had omitted the explicit assumption that there exists a known upper bound $\sup_{x\in\mathcal{X}} p(x) \leq \kappa_2 < \infty$. The existence of such a κ_2 *uniformly over the class of distributions under consideration* actually follows from assuming Hölder continuity. You were not required to show this, but, for completeness, we show this here in the case that p is β-Hölder continuous with constant L, for some $\beta \in (0, 1]$.

Proof: Since p is continuous and $\mathcal{X}$ is compact, $x^* := \operatorname{argmax}_{x\in\mathcal{X}} p(x)$ exists. Let $u := \left(\frac{p(x^*)}{2L}\right)^{1/\beta}$, so that, by the Hölder condition, for all $x \in B(x, u) \cap \mathcal{X}$, [2] $p(x) \geq p(x^*)/2$. Since p is a probability density,

$$1 \geq \int_{B(x^*,u)\cap\mathcal{X}} p(x)\, dx \geq \int_{B(x^*,u)\cap\mathcal{X}} \frac{p(x^*)}{2} \geq \frac{\mu(B(0,u))}{2^D} \cdot \frac{p(x^*)}{2},$$

(the 2^D term is due the fact that at least one orthant of $B(x, u)$ must lie entirely within $\mathcal{X}$). Since u increases with $p(x^*)$, the right side clearly increases unboundedly with $p(x^*)$, and so the latter must be bounded.

Since we need Hölder continuity to bound the bias of $\widehat{p}$ anyway, the existence of κ_2 is a very mild assumption. The existence of κ_1 is a much stronger assumption, but can only be weakened slightly.

Solution: (Based on Liu et al. [2012])

(a) Define $g : (\kappa_1, \kappa_2) \to \mathbb{R}$ by $g(z) = z \log z$. Define constants

$$G_1 := \sup_{z\in[\kappa_1,\kappa_2]} \left|g'(z)\right| = \max\left\{|1 + \log(\kappa_1)|, |1 + \log(\kappa_2)\right\}$$

[2] Here, $B(x, u)$ denotes the ball of radius u centered at x, in the same metric as the Hölder condition, and μ denotes its Lebesgue measure.

and

$$G_2 := \sup_{z \in [\kappa_1, \kappa_2]} |g''(z)| = \frac{1}{\kappa_1}.$$

Note that, for any $x, y \in [\kappa_1, \kappa_2]$, there exists $z \in [x, y] \cup [y, x] \subseteq [\kappa_1, \kappa_2]$ such that

$$g(x) - g(y) = g'(y)(y - x) + \frac{g''(z)}{2}(y - x)^2.$$

$$\begin{aligned}
\left|\mathbb{E}\left[\widehat{H}\right] - H\right| &\leq \left|\mathbb{E}\left[\int_{\mathcal{X}} g\left(\widehat{p}(x)\right) - g(p(x))\, dx\right]\right| \\
&\leq \left|\mathbb{E}\left[\int_{\mathcal{X}} g'(p(x))(p(x) - \widehat{p}(x)) + \frac{g''(\zeta(x))}{2}(p(x) - \widehat{p}(x))^2\, dx\right]\right| \\
&\leq \int_{\mathcal{X}} G_1 \left|p(x) - \mathbb{E}\left[\widehat{p}(x)\right]\right| + \frac{G_2}{2}\mathbb{E}\left[\left(p(x) - \widehat{p}(x)\right)^2\right] dx
\end{aligned}$$

Applying Jensen's inequality (since $\mathcal{X}$ has Lebesgue measure 1),

$$\begin{aligned}
\left|\mathbb{E}\left[\widehat{H}\right] - H\right| &\leq G_1 \sqrt{\int_{\mathcal{X}} \left(p(x) - \mathbb{E}\left[\widehat{p}(x)\right]\right)^2 dx} + \frac{G_2}{2}\int_{\mathcal{X}} \mathbb{E}\left[\left(p(x) - \widehat{p}(x)\right)^2\right] dx \\
&\leq G_1\sqrt{C_0 h^{2\beta}} + \frac{G_2}{2}\left(C_0 h^{2\beta} + \frac{C_1}{nh^D}\right) \\
&\leq C_B\left(h^\beta + h^{2\beta} + \frac{1}{nh^D}\right),
\end{aligned}$$

where the second inequality is by the given bounds on the integrated squared bias and integrated MSE of the kernel density estimator, and

$$C_B := \max\left\{G_1\sqrt{C_0}, \frac{G_2}{2}\max\left\{C_0, C_1\right\}\right\}. \quad \blacksquare$$

(b) For sake of applying McDiarmid's inequality, let $\widehat{p}'$ denote the KDE $\widehat{p}$ when the i^{th} sample X_i is replaced by an independent sample X_i', and let $\widehat{H}'$ denote the corresponding plug-in estimate. By the mean value theorem, for any $x, y > 0$,

$$|x \log x - y \log y| \leq (1 + \max\{|\log x|, |\log y|\})\,|x - y|.$$

Hence, letting $\kappa := 1 + \max\{|\log \kappa_1|, |\log \kappa_2|\}$, since both $\widehat{p}$ and $\widehat{p}^{(i)}$ lie in $[\kappa_1, \kappa_2]$,

$$\begin{aligned}
\left|\widehat{H} - \widehat{H}'\right| &= \left|\int_{\mathcal{X}} \widehat{p}(x)\log\widehat{p}(x) - \widehat{p}'(x)\log\widehat{p}'(x)\, dx\right| \\
&\leq \int_{\mathcal{X}} \left|\widehat{p}(x)\log\widehat{p}(x) - \widehat{p}'(x)\log\widehat{p}'(x)\right| dx \leq \kappa\int_{\mathcal{X}} |\widehat{p}(x) - \widehat{p}'(x)|\, dx.
\end{aligned}$$

Note that, since $\widehat{p}$ and $\widehat{p}'$ differ in only one sample, almost all terms in $p - \widehat{p}$ cancel out:

$$|\widehat{p}(x) - \widehat{p}'(x)| = \frac{1}{nh^D}\left|K\left(\frac{x - X_i}{h}\right) - K\left(\frac{x - X_i'}{h}\right)\right|$$

Now, applying the change of variables $u = \frac{x - X_i}{h}$, since the Jacobian of this transformation has determinant $|J_u x| = |hI_D| = h^D$,

$$\begin{aligned}\left|\widehat{H} - \widehat{H}'\right| &\leq \frac{\kappa}{nh^D} \int_{\mathcal{X}} \left| K\left(\frac{x - X_i}{h}\right) - K\left(\frac{x - X_i'}{h}\right)\right| dx \\ &\leq 2\frac{\kappa}{nh^D} \int_{\mathcal{X}} \left| K\left(\frac{x - X_i}{h}\right)\right| dx \\ &\leq 2\frac{\kappa}{n} \int_{\mathcal{X}} |K(u)| \, du = 2\frac{\kappa}{n}\|K\|_1.\end{aligned}$$

Hence, by McDiarmid's inequality, for $C_E = \frac{1}{2\kappa^2\|K\|_1^2}$,

$$\mathbb{P}\left[\left|\widehat{H} - \mathbb{E}\left[\widehat{H}\right]\right|\right] \leq 2\exp\left(-\frac{2\varepsilon^2}{\sum_{i=1}^n \left(2\frac{\kappa}{n}\|K\|_1\right)^2}\right) = 2\exp\left(-C_E n\varepsilon^2\right). \quad \blacksquare$$

(c) By the previous part, for $C_V := \frac{2}{C_E}$,

$$\begin{aligned}\mathbb{V}\left[\widehat{H}\right] &= \mathbb{E}\left[\left(\widehat{H} - \mathbb{E}\left[\widehat{H}\right]\right)^2\right] \\ &= \int_0^\infty \mathbb{P}\left[\left(\widehat{H} - \mathbb{E}\left[\widehat{H}\right]\right)^2 > \varepsilon\right] d\varepsilon \\ &= \int_0^\infty \mathbb{P}\left[\left|\widehat{H} - \mathbb{E}\left[\widehat{H}\right]\right| > \sqrt{\varepsilon}\right] d\varepsilon \\ &\leq 2\int_0^\infty \exp\left(-C_E\varepsilon n\right) d\varepsilon = \frac{2}{C_E n} = \frac{C_V}{n}. \quad \blacksquare\end{aligned}$$

(d) Combining parts (a) and (c) via the usual bias-variance decomposition of MSE gives

$$\begin{aligned}\mathbb{E}\left[\left(\widehat{H} - H\right)^2\right] &= \mathbb{E}^2\left[\widehat{H} - H\right] + \mathbb{V}\left[\widehat{H}\right] \\ &\leq C_B^2\left(h^\beta + h^{2\beta} + \frac{1}{nh^D}\right) + \frac{C_V}{n}. \qquad (8)\end{aligned}$$

Note that the variance does not depend on h, and so we can just optimize the bias bound over h. Also, since $h \to 0$ as $n \to \infty$, the $h^{2\beta}$ term is negligible; we replace it with a constant factor of 2. Hence, since the bias bound is convex in h, at the optimal bandwidth h_*, we have

$$0 = \frac{d}{dh}h^\beta + n^{-1}h^{-D}\Big|_{h=h_*} = \beta h_*^{\beta-1} - Dn^{-1}h_*^{-(D+1)} = h_*^{\beta+D} - \frac{D}{\beta}n^{-1},$$

and so $h_* \asymp n^{-\frac{1}{\beta+D}}$. Plugging this into (8) gives

$$\mathbb{E}\left[\left(\widehat{H} - H\right)^2\right] \asymp \left(n^{-\frac{\beta}{\beta+D}}\right)^2 + n^{-1} = n^{-\min\left\{1, \frac{2\beta}{\beta+D}\right\}}.$$

For any values of β and D, this rate is faster than the $n^{-\frac{2\beta}{2\beta+D}}$ optimal rate for density estimation, and uses a smaller bandwidth than $h \asymp n^{-\frac{1}{2\beta+D}}$. $\blacksquare$

References

Han Liu, Larry Wasserman, and John D Lafferty. Exponential concentration for mutual information estimation with application to forests. In *Advances in Neural Information Processing Systems*, pages 2537–2545, 2012.

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