APMA 1740/2610 2025: Homework 3

1. **Convergence in probability.** We say that a sequence of random variables $X_1, X_2, \ldots$ converges in probability to another random variable X if

$$\mathbb{P}\big(|X_n - X| > \epsilon\big) \to 0 \quad \text{as } n \to \infty \text{ for every } \epsilon > 0.$$

It is often the case that X is a deterministic constant, which is a trivial example of a random variable.

Let $\theta \in \mathbb{R}$ and suppose that $\widehat{\theta}_n$ is an estimator of θ. Use Markov's inequality[1] to prove the following: If the MSE of $\widehat{\theta}_n$ converges to zero (i.e., $\widehat{\theta}_n$ is consistent in MSE), then $\widehat{\theta}_n$ converges to θ in probability (i.e., $\widehat{\theta}_n$ is consistent in probability).

This is a very useful fact because working with expectations (MSE) is often much easier than working with probabilities.

We want to show that

$$\text{MSE}[\hat{\theta}_n] \to 0 \implies \mathbb{P}(\left|\hat{\theta}_n - \theta\right| > \varepsilon) \to 0$$

Consider

$$\begin{aligned}
\mathbb{P}(\left|\hat{\theta}_n - \theta\right| > \varepsilon) &= \mathbb{P}(\left|\hat{\theta}_n - \theta\right| \geq \varepsilon) - \mathbb{P}(\left|\hat{\theta}_n - \theta\right| = \varepsilon) \\
&= \mathbb{P}(\left|\hat{\theta}_n - \theta\right| \geq \varepsilon) \qquad (\varepsilon \to 0) \\
&= \mathbb{P}(\left|\hat{\theta}_n - \theta\right|^2 \geq \varepsilon^2) \\
&\leq \frac{1}{\varepsilon^2}\mathbb{E}[\left|\hat{\theta}_n - \theta\right|^2] \qquad \text{(Markov's Inequality)} \\
&= \frac{1}{\varepsilon^2}\text{MSE}[\hat{\theta}_n]
\end{aligned}$$

By assumption, $\text{MSE}[\hat{\theta}_n] \to 0$ so $\mathbb{P}(\left|\hat{\theta}_n - \theta\right| > \varepsilon) \to 0$ as well.

[1] Markov's inequality states that $\mathbb{P}\big(|Z| \geq \alpha\big) \leq \mathbb{E}(|Z|)/\alpha$ for every $\alpha > 0$. Note that you can apply Markov's inequality to $|Y|^k$ by taking $Z = |Y|^k$.

2. **Bias-variance tradeoff.** Consider the problem of flipping a coin n times and then guessing the probability of heads. In other words, let $X_{1:n}$ be iid Bernoulli(p) for $p \in [0,1]$ and let $\widehat{p} \triangleq \widehat{p}(X_{1:n})$ be an estimator of p. Consider the following possible estimators $\widehat{p}$ of p:

$\widehat{p}_1 \triangleq 0.5$	This estimator ignores the data and always guesses 0.5.
$\widehat{p}_2 \triangleq \bar{X}_n \triangleq n^{-1}\sum_{i=1}^n X_i$	This estimator is simply the fraction of ones in the data.
$\widehat{p}_3 \triangleq \frac{2}{n+2}\widehat{p}_1 + \frac{n}{n+2}\widehat{p}_2$	This estimator moves $\bar{X}_n$ closer to 0.5, more so for less data.

(a) Compute the bias, variance, and mean squared error (MSE) as functions of n and p for each of the three estimators.

Bias:

$$\mathbb{E}[\hat{p}_1] - p = \boxed{0.5 - p}$$

$$\mathbb{E}[\hat{p}_2] - p = \mathbb{E}[\bar{X}_n] - p = \frac{1}{n}\sum_{i=1}^n \mathbb{E}[X_i] - p = p - p = \boxed{0}$$

$$\mathbb{E}[\hat{p}_3] - p = \frac{2}{n+2}\mathbb{E}[\hat{p}_1] + \frac{n}{n+2}\mathbb{E}[\hat{p}_2] - p = \frac{2}{n+2}(0.5) + \frac{n}{n+2}(p) - p = \boxed{\frac{1-2p}{n+2}}$$

Variance:

$$\text{Var}[\hat{p}_1] = \text{Var}[0.5] = \boxed{0}$$

$$\text{Var}[\hat{p}_2] = \frac{1}{n^2}\sum_{i=1}^n \text{Var}[X_i] = \frac{1}{n^2}\cdot n \cdot p(1-p) = \boxed{\frac{p(1-p)}{n}}$$

$$\text{Var}[\hat{p}_3] = \frac{4}{(n+2)^2}\text{Var}[\hat{p}_1] + \frac{n^2}{(n+2)^2}\text{Var}[\hat{p}_2] = \frac{n^2}{(n+2)^2}\frac{p(1-p)}{n} = \boxed{\frac{np(1-p)}{(n+2)^2}}$$

Finally, since $\text{MSE}[\hat{\theta}] = \text{Var}[\hat{\theta}] + b^2$,

$$\text{MSE}[\hat{p}_1] = \boxed{(0.5-p)^2}$$

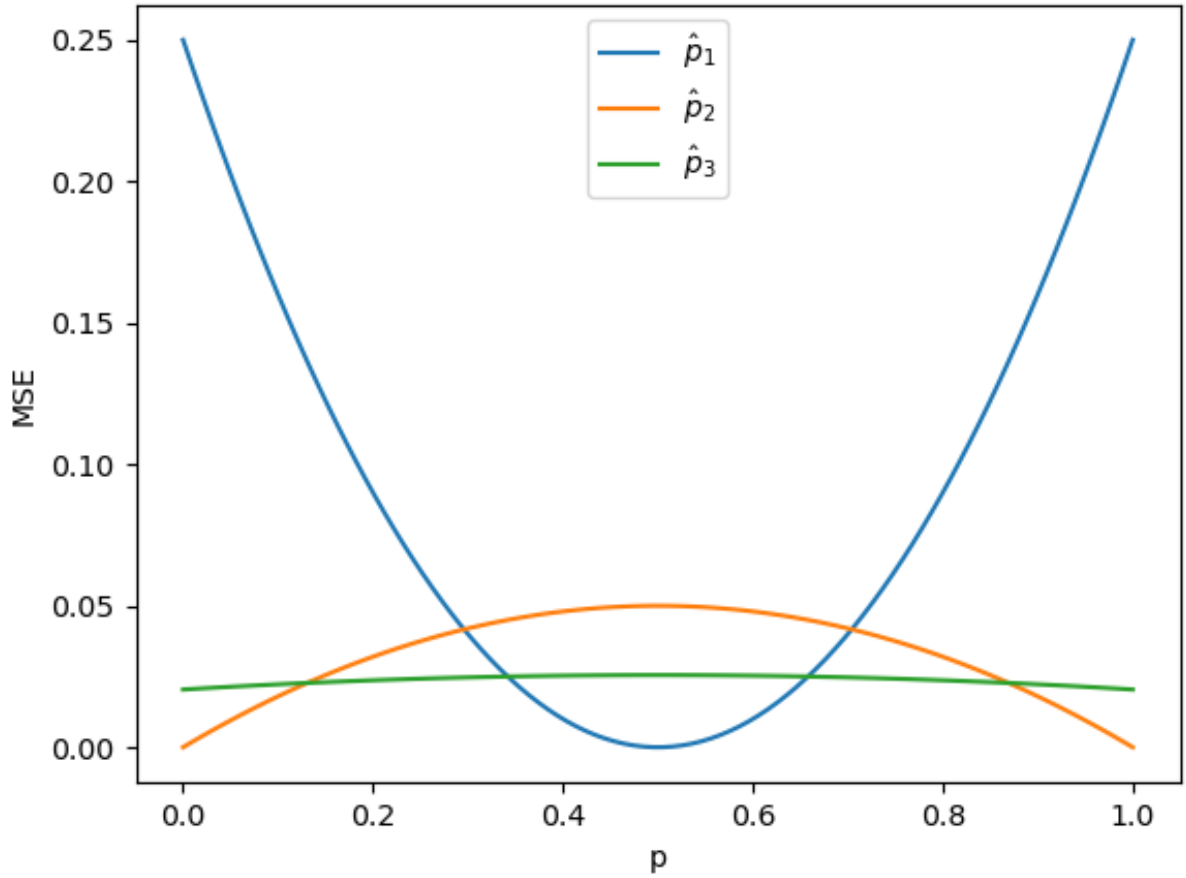
$$\text{MSE}[\hat{p}_2] = \boxed{\frac{p(1-p)}{n}}$$

$$\text{MSE}[\hat{p}_3] = \frac{np(1-p)}{(n+2)^2} + \left(\frac{1-2p}{n+2}\right)^2 = \frac{(4-n)p^2 + (n-4)p + 1}{(n+2)^2}$$

(b) Which, if any, of these estimators is unbiased? Which, if any, is consistent in probability? Why?

$\hat{p}_2$ is unbiased since $\mathbb{E}[\hat{p}_2] = p$ for all n. Both $\hat{p}_2$ and $\hat{p}_3$ are consistent in probability since $\text{MSE}[\hat{p}_2] \to 0$ and $\text{MSE}[\hat{p}_3] \to 0$ and by Problem 1, consistence in MSE implies consistence in probability.

(c) For the case $n = 5$, plot the three MSEs as a function of p on the same graph.



(d) For the case $n = 5$, are any of the three estimators uniformly better than the others? Why or why not?

No. $\hat{p}_1$ optimizes for the case $p = 0.5$ while $\hat{p}_2$ optimizes for the case $p = 0$ or $p = 1$. $\hat{p}_3$ is a compromise between the two.

(e) For each of the three estimators, describe scenarios where that estimator is the one you would choose. For example, suppose you have an a priori reason to believe that the coin is heavily weighted to almost always land on the same side. Which estimator might you choose?

If we strongly believed that the coin was fair, we would choose $\hat{p}_1$ as it has lowest MSE. If, however, we had reason to suspect the coin was heavily weighted towards one side, we would choose $\hat{p}_2$ which is unbiased and minimizes MSE at $p = 0$ and $p = 1$. In other cases, it likely makes sense to use $\hat{p}_3$ as it is a compromise between the two.

3. (Do not submit.) **Kernel density estimation.** Given an unknown density function f on $\mathbb{R}$, and a sample $X_{1:n} \sim iid\ f$, our goal is to construct an estimator $\hat{f}_n(x; X_{1:n})$ of f. (And preferably one that has some kind of guarantee to approach f as the sample size, n, goes to ∞.)

Let k (a "kernel") be some density function $k(x)$ on $\mathbb{R}$, with the properties

$$\int_{x=-\infty}^{\infty} xk(x)dx = 0 \quad \text{(mean zero)} \qquad \& \qquad \int_{x=-\infty}^{\infty} x^2k(x)dx = 1 \quad \text{(standard deviation one)}$$

Convince yourself that for any $w \in (0, \infty)$, $k_w(x) \triangleq \frac{1}{w}k(\frac{x}{w})$ is again a zero-mean density function, but now with standard deviation w (sometimes called the "bandwidth" of the kernel k_w).

The "kernel density estimator" of f, based on the kernel k and the observations $X_{1:n}$, is

$$\hat{f}_{n,w}(x; X_{1:n}) \triangleq \frac{1}{n}\sum_{k=1}^{n} k_w(x - X_k)$$

In words, we center a density (k_w) at each sample point. The estimator at any location x is just the average of the heights at x of these n centered densities. In this problem you will experiment with the results of using this method on a constructed (and hence known) density. The kernel will be the standard normal density, $\mathcal{N}(0, 1)$, and hence

$$k_w(x) = \frac{1}{w}\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2w^2}}$$

The constructed density is

$$f(x) \triangleq \frac{1}{10}\sum_{\ell=1}^{10} \phi(x, \ell, \cos^2(\ell))$$

where $\phi(x, \mu, \sigma^2)$ is the pdf of a normal random variable with mean μ and variance σ^2. This is called a mixture of Gaussians.[2]

(a) Plot f over the interval $[-2, 14]$ using bins of size 0.001. The function `normpdf` can simplify your code in Matlab (but note that it takes σ, not σ^2, as an argument). A worthwhile sanity check at this point is to generate a million samples from f and make a histogram to verify that you get the same shape.

(b) Generate a sample of size 200 from f. Using the standard normal pdf as the kernel, compute kernel density estimates of f for a variety of bandwidths.[3] Graph three of these — one with a (somewhat, but not excessively) too small bandwidth, one with a (somewhat, but not excessively) too large bandwidth, and one with a bandwidth that seems about right to you. Put the true f on the graphs, too, for comparison. (Try to pretend that you don't know the real f, and think about which of these bandwidths you would choose.)

(c) Repeat part (b) 50 times (generate new data each time) using the same three bandwidths that you selected. For each bandwidth, make a graph showing all 50 kernel density estimates and also the true density.[4] Interpret your plots in terms of bias and variance. Which bandwidth do you think has the best mean integrated squared error (MISE)?

(d) Since you know the true f, you can (numerically) compute ISE. (You can't do this in a real problem, because f is unknown.) For each of the three bandwidths used in part (c), compute the ISE for each of your 50 kernel density estimates from part (c) and average them to get an estimate of the MISE. Report your estimate of MISE for each bandwidth. Was your intuition from part (c) correct?

(e) Repeat part (d) at a selection of bandwidths ranging from 0.01 to 3. (The more the better, but this can be computationally expensive.) Plot the curve of estimated MISE as a function of bandwidth. You should see a more-or-less smooth curve with an apparently unique minimum. (You will need higher resolution at smaller values of w. I used $w \in [0.01 : 0.01 : 0.3,\ .4 : 0.1 : 1,\ 2,\ 3]$, i.e. w from 0.01 to 0.3 in steps of size 0.01; from 0.4 to 1 in steps of size 0.1; and $w = 2$ and $w = 3$, for a total of 39 values. But this might take more time on your computer than you're willing to wait.)

---

[2] Here is one way to create a sample of size n from this distribution, say $X_1, \ldots, X_n$. First sample n labels $L_1, \ldots, L_n$ uniformly from the set $\{1, 2, \ldots, 10\}$, for example, with `L = randi(10,n,1)` in Matlab. Now sample each X_i independently according to Normal$(L_i, \cos^2(L_i))$, for example, with `X = L + abs(cos(L)).*randn(n,1)` in Matlab.

[3] Something like this might work in Matlab: `x = -2:.001:14; fhat = zeros(size(x)); for k = 1:200, fhat = fhat + normpdf(x,X(k),w)/200; end`

[4] The matlab command `hold on` is useful here. Also, if you plot the true density last and use something like `plot(x,f,'k','linewidth',2)` then it can be a lot easier to distinguish the true density from your 50 estimates.

4. (Do not submit.) **Cross validation for kernel density estimation.**

In the previous problem, we explored the accuracy of various bandwidths with the luxury of knowing the target density. Of course we would not need to estimate the density in the first place if we it were already known. In this problem we will use the cross-validation approach developed in class to estimate a suitable bandwidth from the data. Stone's theorem assures us that the resulting estimator of the density is asymptotically optimal, and hence we can expect good behavior for a large sample. How good is the approach at a modest sample size of $n = 200$. (Modest, at least, in the context of nonparametric estimation.)

Recall that cross validation selects the bandwidth using

$$\widehat{h}_n \triangleq \arg\min_h \widehat{J}(h)$$

where

$$\widehat{J}(h) \triangleq \int_{-\infty}^{\infty} \left(\widehat{f}_{n,h}(x)\right)^2 dx - \frac{2}{n}\sum_{i=1}^{n} \widehat{f}^{(i)}_{n-1,h}(X_i; {}_iX)$$

and $\widehat{f}^{(i)}_{n-1,h}(x_i; {}_iX)$ is the kernel density estimator using all of the data except X_i. (*See below for a better way to write $\widehat{J}(h)$ when the kernel is the standard normal, and for some computational hints.*)

(a) Select 200 independent samples from f. Compute $\widehat{J}(h)$ at each $h \in \{.01, .02, \ldots, .99, 1\}$. Plot $\widehat{J}$.

(b) Choose the bandwidth, $\hat{h}^*$, that minimizes $\widehat{J}(h)$. Report the value of $\hat{h}^*$, and display both the density f and the kernel density estimator that uses the cross-validated bandwidth. Put both functions on the same plot.

(c) Repeat (b) 50 times, and display all fifty estimators in one plot. Superimpose f, as was done in Problem 3. (Be sure to compute a new bandwidth each time.) Use the 50 estimators to compute the approximate MISE for cross-validated estimation of this f.

(d) How does cross validation compare to a fixed, hand-selected bandwidth? Was the MISE using cross-validation better than each of the fixed bandwidths from Problem 3(d)? What were the minimum and maximum bandwidths chosen by cross-validation for your 50 datasets? Comment on your results.

Computational guidelines. You can simplify $\widehat{J}(h)$ for the standard normal base kernel by noting that

$$\begin{aligned}
\int \left(\widehat{f}_{n,h}(x)\right)^2 dx &= \int \left(\frac{1}{n}\sum_{i=1}^{n} \kappa_h(x - X_i)\right)^2 dx \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} \int \kappa_h(x - X_i)\kappa_h(x - X_j)dx \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} \frac{1}{2\pi h^2} \int \exp\left(\frac{-(x - X_i)^2 - (x - X_j)^2}{2h^2}\right) dx \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} \frac{1}{\sqrt{2\pi}(h\sqrt{2})} \exp\left(\frac{-(X_i - X_j)^2}{2(h\sqrt{2})^2}\right) = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} \kappa_{h\sqrt{2}}(X_i - X_j)
\end{aligned}$$

and

$$\begin{aligned}
-\frac{2}{n}\sum_{i=1}^{n} \widehat{f}^{(i)}_{n-1,h}(X_i) &= -\frac{2}{n}\sum_{i=1}^{n} \frac{1}{n-1}\sum_{j\neq i} \kappa_h(X_i - X_j) \\
&= -\frac{2}{n(n-1)}\sum_{i=1}^{n}\left(\left(\sum_{j=1}^{n} \kappa_h(X_i - X_j)\right) - \kappa_h(X_i - X_i)\right) \\
&= -\left(\frac{2}{n(n-1)}\sum_{i=1}^{n}\sum_{j=1}^{n} \kappa_h(X_i - X_j)\right) + \frac{2}{n-1}\kappa_h(0) \\
&= \frac{2\kappa_h(0)}{n-1} - \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} \frac{2n}{n-1}\kappa_h(X_i - X_j)
\end{aligned}$$

so that

$$\widehat{J}(h) = \frac{2\kappa_h(0)}{n-1} + \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\left(\kappa_{h\sqrt{2}}(X_i - X_j) - \frac{2n}{n-1}\kappa_h(X_i - X_j)\right),$$

where κ_h is the $\mathrm{N}(0, h^2)$ pdf. Precompute the distances $X_i - X_j$ and store them in a 200×200 matrix D. Then in Matlab you need only use:

```
sum(sum(normpdf(D,0,h*sqrt(2)) - 2*(n/(n-1))*normpdf(D,0,h)))
```

to get the double sum in the simplified expression for $\widehat{J}(h)$, which should evaluate very quickly.

5. **Estimating a pmf: consistency in mean square and in probability.** Let $\mathcal{I}$ be a discrete set (i.e., finite or countably infinite). For convenience, we will label the elements with integers, i.e., $\mathcal{I} \triangleq \{1, 2, \dots\}$. Let $p \triangleq (p_1, p_2, \dots)$ be an unknown probability mass function (pmf) on $\mathcal{I}$, i.e., $0 \le p_i \le 1$ for all $i \in \mathcal{I}$ and $\sum_{i \in \mathcal{I}} p_i = 1$. Suppose that $X_{1:n}$ is iid with common pmf p over $\mathcal{I}$.

(a) Find the maximum likelihood estimator (MLE) of p. Call it $\hat{p}^{\text{MLE}}$. Note that $\hat{p}^{\text{MLE}} = (\hat{p}_1^{\text{MLE}}, \hat{p}_2^{\text{MLE}}, \dots)$ is a pmf and that it depends on both n and the data $X_{1:n}$, which are suppressed in the notation.

$$\begin{aligned}
\hat{p}^{\text{MLE}} &= \arg\max_{\tilde{p}} \tilde{p}(X_1 = x_1, X_2 = x_2, \dots, X_n = x_n) \\
&= \arg\max_{\tilde{p}} \prod_{j=1}^{n} \tilde{p}(X_j = x_j) \\
&= \arg\max_{\tilde{p}} \prod_{i \in \mathcal{I}} \tilde{p}_i^{\sum_{j=1}^{n} \mathbb{1}\{X_j = i\}} \\
&= \arg\max_{\tilde{p}} \sum_{i \in \mathcal{I}} \sum_{j=1}^{n} \mathbb{1}\{X_j = i\} \log \tilde{p}_i \\
&= \arg\max_{\tilde{p}} \sum_{i \in \mathcal{I}} \frac{1}{n} \sum_{j=1}^{n} \mathbb{1}\{X_j = i\} \log \tilde{p}_i \\
&= \arg\max_{\tilde{p}} \sum_{i \in \mathcal{I}} \hat{p}_i \log \tilde{p}_i
\end{aligned}$$

Now we can use Lagrange multipliers with $\sum_{i \in \mathcal{I}} \tilde{p}_i = 1$:

$$\frac{\partial}{\partial \tilde{p}_i} \left[\sum_{i \in \mathcal{I}} \hat{p} \log \tilde{p}_i + \lambda (1 - \sum_{i \in \mathcal{I}} \tilde{p}_i) \right] = \frac{\hat{p}_i}{\tilde{p}_i} - \lambda = 0 \implies \tilde{p}_i = \frac{\hat{p}_i}{\lambda}$$

But

$$\sum_{i \in \mathcal{I}} \tilde{p}_i = \sum_{i \in \mathcal{I}} \frac{\hat{p}_i}{\lambda} = 1 \implies \sum_{i \in \mathcal{I}} \hat{p}_i = \lambda$$

and since the empirical distribution is a pmf, $\lambda = 1$ which finally tells us

$$\hat{p}_i^{\text{MLE}} = \hat{p}_i = \frac{1}{n} \sum_{j=1}^{n} \mathbb{1}\{X_j = i\}$$

(b) What is the MLE in this case? We have a name for it.

The empirical distribution.

(c) For fixed i, compute the mean squared error (MSE) of $\hat{p}_i^{\text{MLE}}$. Call it MSE_i. Note that MSE_i depends on n and on the true pmf p, which are suppressed in the notation.

$$\begin{aligned}
\text{MSE}_i[\hat{p}] &= \text{Var}[\hat{p}_i^{\text{MLE}}] + (\mathbb{E}[\hat{p}_i^{\text{MLE}}] - p_i)^2 \\
&= \text{Var}[\hat{p}_i] + \text{Bias}[hatp_i]^2
\end{aligned}$$

First,

$$\text{Bias}[\hat{p}_i] = \mathbb{E}[\hat{p}_i] - p_i = \frac{1}{n} \sum_{j=1}^{n} \mathbb{E}[\mathbb{1}\{X_j = i\}] - p_i = \frac{1}{n} \sum_{j=1}^{n} p_i - p_i = p_i - p_i = 0$$

$$\text{Bias}^2[\hat{p}_i] = 0$$

And

$$\begin{aligned}
\mathrm{Var}[\hat{p}_i] &= \mathrm{Var}\left[\frac{1}{n}\sum_{j=1}^{n}\mathbb{1}\{X_j = i\}\right] \\
&= \frac{1}{n^2}\sum_{j=1}^{n}\mathrm{Var}[\mathbb{1}\{X_j = i\}] \\
&= \frac{1}{n^2}\sum_{j=1}^{n}\mathbb{E}[\mathbb{1}\{X_j = i\}^2] - \mathbb{E}[\mathbb{1}\{X_j = i\}]^2 \\
&= \frac{1}{n^2}\sum_{j=1}^{n} p_i - p_i^2 \\
&= \frac{1}{n}(p_i - p_i^2) = \frac{p_i(1-p_i)}{n}
\end{aligned}$$

so

$$\mathrm{MSE}_i[\hat{p}] = \frac{p_i(1-p_i)}{n}$$

(d) For fixed i, show that $\hat{p}_i^{\mathrm{MLE}}$ is consistent in mean square and in probability. Hint: Use part (c) and then Problem 1.

Trivially,

$$\lim_{n\to\infty}\mathrm{MSE}_i[\hat{p}] = \lim_{n\to\infty}\frac{p_i(1-p_i)}{n} = 0$$

hence $\hat{p}_i^{\mathrm{MLE}}$ is consistent in mean square.

By Problem 1, this also implies that $\hat{p}_i^{\mathrm{MLE}}$ is consistent in probability.

(e) Define the total squared error of the whole pmf $\hat{p}^{\mathrm{MLE}}$ to be

$$\textstyle\sum_{i\in\mathcal{I}}(\hat{p}_i^{\mathrm{MLE}} - p_i)^2.$$

(This is similar to integrated squared error for pdfs, except that for pmfs we are using a sum, instead of an integral.) The total MSE is thus

$$\mathrm{MSE} = \mathbb{E}\left(\textstyle\sum_{i\in\mathcal{I}}(\hat{p}_i^{\mathrm{MLE}} - p_i)^2\right).$$

Show that

$$\mathrm{MSE} = \textstyle\sum_{i\in\mathcal{I}}\mathrm{MSE}_i,$$

and then use part (c) to show that

$$\mathrm{MSE} \le \frac{1}{n}$$

regardless of the true pmf p.

$$\begin{aligned}
\mathrm{MSE} &= \mathbb{E}\left[\sum_{i\in\mathcal{I}}(\hat{p}_i^{\mathrm{MLE}} - p_i)^2\right] = \sum_{i\in\mathcal{I}}\mathbb{E}[(\hat{p}_i^{\mathrm{MLE}} - p_i)^2] \\
&= \sum_{i\in\mathcal{I}}\mathrm{MSE}_i = \sum_{i\in\mathcal{I}}\frac{p_i(1-p_i)}{n} = \frac{1}{n}\sum_{i\in\mathcal{I}} p_i - p_i^2 \\
&\le \frac{1}{n}\sum_{i\in\mathcal{I}} p_i = \frac{1}{n}
\end{aligned}$$

(f) Show that the whole pmf $\hat{p}^{\mathrm{MLE}}$ is consistent for p in mean square in the sense that $\mathrm{MSE} \to 0$ as $n \to \infty$.

By dominated convergence, since $\mathrm{MSE} \le \frac{1}{n}$ and $\frac{1}{n} \to 0$ as $n \to \infty$, $\mathrm{MSE} \to 0$ as $n \to \infty$. Hence, $\hat{p}^{\mathrm{MLE}}$ is consistent in mean square.

6. **MLE in exponential families.** A gamma(α, β) random variable has pdf

$$f_{\alpha,\beta}(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x} \mathbb{1}\{x > 0\}$$

for parameters $\alpha, \beta > 0$.[5] (In this problem, as usual, $\log = \log_e$.)

(a) Express $f_{\alpha,\beta}$ as an exponential family and identify the sufficient statistics.

Let

$$\begin{aligned} \lambda &= (\alpha, \beta) \\ Z(\alpha, \beta) &= \frac{\Gamma(\alpha)}{\beta^\alpha} \\ h(x) &= x^{-1} \mathbb{1}\{x > 0\} \\ T(x) &= (\log x, -x) \end{aligned}$$

Then,

$$\begin{aligned} \frac{1}{Z(\alpha,\beta)} h(x) e^{\lambda \cdot T(x)} &= \frac{\beta^\alpha}{\Gamma(\alpha)} x^{-1} \mathbb{1}\{x > 0\} \exp(\alpha \log x - \beta x) \\ &= \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x} \mathbb{1}\{x > 0\} \\ &= f_{\alpha,\beta}(x) \end{aligned}$$

so $f_{\alpha,\beta}$ is an exponential family with sufficient statistics $\log x$ and $-x$.

(b) Suppose that I collect this data set of size 5:

$$1.45,\ 3.24,\ 1.07,\ 0.34,\ 2.29,$$

and that I model it as iid gamma(α, β) for unknown parameters α, β. It turns out that the MLE is[6]

$$(\hat{\alpha}^{\text{MLE}}, \hat{\beta}^{\text{MLE}}) = (2.1955, 1.3084).$$

Compute the value of the following integral, using nothing more than about ten operations on a calculator (or, say, in the Matlab Command window).

$$\int_0^\infty \log(x) \frac{1.3084^{2.1955}}{\Gamma(2.1955)} x^{2.1955-1} e^{-1.3084x} dx.$$

(The point is that you are not supposed to use something like Mathematica to symbolically integrate this, or Matlab to numerically integrate it. Instead, you are supposed to use theory we developed in class.)

$$\begin{aligned} \int_0^\infty \log x \frac{1.3084^{2.1955}}{\Gamma(2.1955)} x^{2.1955-1} e^{-1.3-84x}\ dx &= \int_0^\infty \log x \frac{(\hat{\beta}^{\text{MLE}})^{\hat{\alpha}^{\text{MLE}}}}{\Gamma(\hat{\alpha}^{\text{MLE}})} x^{\hat{\alpha}^{\text{MLE}}-1} e^{-\hat{\beta}^{\text{MLE}} x}\ dx \\ &= \int_0^\infty \log x f_{\hat{\alpha}^{\text{MLE}}, \hat{\beta}^{\text{MLE}}}(x)\ dx \\ &= \mathbb{E}_f[\log X] \\ &= \mathbb{E}_f[T_1(X)] \\ &= \overline{T_1(X)} \\ &= \frac{1}{5} \sum_{i=1}^5 T_1(x_i) \\ &= \frac{\log(1.45) + \log(3.24) + \log(1.07) + \log(0.34) + \log(2.29)}{5} \\ &\approx 0.2729 \end{aligned}$$

[5] Sometimes the gamma pdf is parameterized by α and $1/\beta$. If you are using software for gamma random variables, be sure you know which parameterization is being used. (You don't need software in this problem.) $\Gamma(\alpha) \triangleq \int_0^\infty x^{\alpha-1} e^{-x} dx$ is the *gamma function*.

[6] For the purposes of this problem, you can assume that this is the MLE to infinite numerical precision, although it is not really.

For 2610 or for extra credit:

7. **Modes of convergence, a brief introduction.** There are many ways in which a sequence of random variables (say $W_1, W_2, \ldots$) can converge to another random variable (say V). These "modes of convergence" can be conveniently discussed in the context of the special case in which $V = 0$, by simply replacing W_k with $W_k - V$ for each $k = 1, 2, \ldots$. Having made that simplification, consider these three particular modes of convergence:

   i. $\mathbb{P}\left(\lim_{k\to\infty} W_k = 0\right) = 1$, "almost sure convergence"

   ii. $\lim_{k\to\infty} \mathbb{P}\left(|W_k| > \epsilon\right) = 0$, for every $\epsilon > 0$, "convergence in probability"

   iii. $\lim_{k\to\infty} \mathbb{E}\left(|W_k|^2\right) = 0$, "mean square convergence"

   What, if any, relationships exist between different modes of convergence? For most pairs, the relationships are qualified, e.g. mode 'a' implies mode 'b' if the W_k's are uniformly bounded. Concerning the three modes listed above, there are two unqualified relationships: (1) almost sure convergence always implies convergence in probability, and (2) mean square convergence always implies convergence in probability. The first requires some elementary measure theory, and the second can be proven with the Markov inequality, as your saw in Problem 1.

   Oftentimes, the best way to understand the connections among the different modes is through specific counterexamples. Concerning the three modes defined here, there are six possible relationships (of the type 'a' implies 'b'). The unqualified relationships listed above address two of the six. None of remaining four are unqualified, and hence there are, in each case, counterexamples. Here we will construct a counterexample for each of the four.

   Let X have the uniform distribution on $[0, 1]$. Define a sequence of subsets of $[0, 1]$, $A_1, A_2, \ldots \subseteq [0, 1]$, as follows:

$$\begin{array}{lllllll}
A_1 = [0,1] & & & & & & \\
A_2 = [0,\frac{1}{2}] & A_3 = [\frac{1}{2},1] & & & & & \\
A_4 = [0,\frac{1}{4}] & A_5 = [\frac{1}{4},\frac{2}{4}] & A_6 = [\frac{2}{4},\frac{3}{4}] & A_7 = [\frac{3}{4},1] & & & \\
A_8 = [0,\frac{1}{8}] & A_9 = [\frac{1}{8},\frac{2}{8}] & A_{10} = [\frac{2}{8},\frac{3}{8}] & A_{11} = [\frac{3}{8},\frac{4}{8}] & \cdots & A_{15} = [\frac{7}{8},1] & \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \cdots
\end{array}$$

   For every $n = 0, 1, \ldots$ there are 2^n intervals of width $\frac{1}{2^n}$ (namely A_k for $k = 2^n, \ldots, 2^{n+1} - 1$) which share endpoints but otherwise partition $[0, 1]$.

   (a) Define $W_k = \mathbb{1}_{X \in A_k}$, $k = 1, 2, \ldots$. Show that W_k converges to zero in probability and mean square, but not almost surely.

   Assume $n = 2^k + j$ for some $j < 2^k$. Then we may write

$$W_n = \mathbb{1}\{X \in A_n\} = \mathbb{1}\{[\frac{j}{2^k}, \frac{j+1}{2^k}]\}$$

   Mean Square:

$$\text{MSE}[W_n] = \mathbb{E}\,|W_n|^2 = \mathbb{E}[\mathbb{1}\{[\frac{j}{2^k}, \frac{j+1}{2^k}]\}^2] = \mathbb{E}[\mathbb{1}\{[\frac{j}{2^k}, \frac{j+1}{2^k}]\}] = \mathbb{P}(X \in [\frac{j}{2^k}, \frac{j+1}{2^k}]) = \frac{1}{2^k} \to 0$$

   Probability: By Problem 1, $MSE[W_n] \to 0 \implies W_n \xrightarrow{\mathbb{P}} 0$.

   Almost Sure: Suppose $\mathbb{P}(\lim_{n\to\infty} W_n = 0) = 1$. But $\forall x_0$, $\exists k$ such that $x_0 \in [\frac{j}{2^k}, \frac{j+1}{2^k}]$ (nested interval property). But then $W_n(x_0) = 1$ on a set with measure $2^{-k} > 0$. Contradiction.

   (b) Define $W_k = 2^k \mathbb{1}_{X \in A_k}$, $k = 1, 2, \ldots$. Show that W_k converges to zero in probability but not in mean square and not almost surely.

   Mean Square:

$$\text{MSE}[W_n] = \mathbb{E}\,|W_n|^2 = \mathbb{E}[2^{2n}\mathbb{1}\{[\frac{j}{2^k}, \frac{j+1}{2^k}]\}] = 2^{2n}\mathbb{P}(X \in [\frac{j}{2^k}, \frac{j+1}{2^k}]) = 2^{2n-k} = 2^{2(2^k+j)-k} \not\to 0$$

   Probability:

$$\mathbb{P}(|[|\,W_n] > \varepsilon) = \mathbb{P}(|2^n\mathbb{1}\{X \in A_n\}| > \varepsilon) = \mathbb{P}(\mathbb{1}\{[\frac{j}{2^k}, \frac{j+1}{2^k}]\} > \frac{\varepsilon}{2^n}) = \mathbb{P}(X \in [\frac{j}{2^k}, \frac{j+1}{2^k}]) = \frac{1}{2^k} \to 0$$

Almost Sure:

$$\mathbb{P}(\lim_{n\to\infty} W_n = 0) = \mathbb{P}(\lim_{n\to\infty} 2^n \mathbb{1}\{X \in A_n\} = 0) = \mathbb{P}(X \notin A_n \quad \forall n \geq N)$$

but by same argument as for (a), this is not possible.

(c) Find a sequence of intervals $(A_1, A_2, \ldots \subseteq [0,1])$, and a corresponding sequence of numbers $(\alpha_1, \alpha_2, \ldots)$, such that if $W_k = \alpha_k \mathbb{1}_{X \in A_k}$ for all $k = 1, 2, \ldots$, then W_k converges to zero almost surely but not in mean square.

Let $\alpha_k = \sqrt{2^k}$ and $A_k = [0, \frac{1}{2^k}]$ so $W_k = \sqrt{2^k}\mathbb{1}\{X \in [0, \frac{1}{2^k}]\}$.

Almost Sure:

$$\mathbb{P}(\lim_{k\to\infty} W_k = 0) = \mathbb{P}(\lim_{k\to\infty} \sqrt{2^k}\mathbb{1}\{X \in [0, \frac{1}{2^k}]\} = 0) = \mathbb{P}(X \notin \lim_{k\to\infty} [0, \frac{1}{2^k}]) = 1 - \lim_{k\to\infty} \frac{1}{2^k} = 1 - 0 = 1$$

Mean Square:

$$\lim_{k\to\infty} \mathbb{E}\left[\left|\sqrt{2^k}\mathbb{1}\{X \in [0, \frac{1}{2^k}]\}\right|^2\right] = \lim_{k\to\infty} 2^k \cdot \mathbb{P}(X \in [0, \frac{1}{2^k}]) = \lim_{k\to\infty} 2^k \cdot \frac{1}{2^k} = 1 \neq 0$$

8. **Box-car kernel: pointwise consitency.** Consider the box-car kernel $\kappa(x) \triangleq \mathbb{1}\{|x| \leq 1/2\}$ and define the family of box-car kernels in the usual way via $\kappa_h(x) \triangleq \kappa(x/h)/h$, where $h > 0$ is the bandwidth. (In the lecture I assumed that the standard deviation of the kernel was one, but this was only for convenience.) Let $X_{1:n}$ be iid with pdf f and corresponding cdf F, and let $\widehat{f}_{n,h}$ be the kernel density estimator of f using κ_h.

(a) Fix x. Show that

$$\mathbb{E}\big(\widehat{f}_{n,h}(x)\big) = \frac{F(x+h/2) - F(x-h/2)}{h}$$

and that

$$\mathrm{Var}\big(\widehat{f}_{n,h}(x)\big) = \frac{F(x+h/2) - F(x-h/2) - \big(F(x+h/2) - F(x-h/2)\big)^2}{nh^2}.$$

Hint: $\widehat{f}_{n,h}(x)$ is the average of iid Bernoulli random variables.

$$\begin{aligned}
\hat{f}_{n,h}(x) &= \frac{1}{n}\sum_{i=1}^{n} \kappa_h(x - X_i) \\
&= \frac{1}{nh}\sum_{i=1}^{n} \kappa(\frac{x - X_i}{h}) \\
\mathbb{E}[\hat{f}_{n,h}(x)] &= \mathbb{E}\left[\frac{1}{nh}\sum_{i=1}^{n} \mathbb{1}\left\{|x - X_i| \leq \frac{h}{2}\right\}\right] \\
&= \frac{1}{nh}\sum_{i=1}^{n} \mathbb{E}\left[\mathbb{1}\{|x - X_i| \leq \frac{h}{2}\}\right] \\
&= \frac{1}{nh}\sum_{i=1}^{n} \mathbb{P}(|x - X_i| \leq \frac{h}{2}) \\
&= \frac{1}{nh}\sum_{i=1}^{n} \mathbb{P}(|X_i| \leq x + \frac{h}{2}) \\
&= \frac{F(x+\frac{h}{2}) - F(x-\frac{h}{2})}{h} \\
\mathrm{Var}[\hat{f}_{n,h}(x)] &= \mathrm{Var}\left[\frac{1}{nh}\sum_{i=1}^{n} \mathbb{1}\left\{|x - X_i| \leq \frac{h}{2}\right\}\right] \\
&= \frac{1}{n^2h^2}\sum_{i=1}^{n} \mathrm{Var}\left[\mathbb{1}\{|x - X_i| \leq \frac{h}{2}\}\right] \\
&= \frac{1}{n^2h^2}\sum_{i=1}^{n} \mathbb{E}\left[\mathbb{1}\{|x - X_i| \leq \frac{h}{2}\}^2\right] - \mathbb{E}\left[\mathbb{1}\{|x - X_i| \leq \frac{h}{2}\}\right]^2 \\
&= \frac{1}{n^2h^2}\sum_{i=1}^{n} \mathbb{P}(|X_i| \leq x + \frac{h}{2}) + \mathbb{P}(|X_i| \leq x + \frac{h}{2})^2 \\
&= \frac{F(x+\frac{h}{2}) - F(x-\frac{h}{2}) + (F(x+\frac{h}{2}) - F(x-\frac{h}{2}))^2}{nh^2}
\end{aligned}$$

(b) Fix an x for which f is continuous at x. Show that if $h_n \to 0$ and $nh_n \to \infty$ as $n \to \infty$, then $\widehat{f}_{n,h_n}(x)$ is consistent for $f(x)$ in probability.

By problem 1, it suffices to show that $\hat{f}_{n,h}(x)$ is consistent in MSE:

$$\mathrm{MSE}[\hat{f}_{n,h_n}] = \mathrm{Var}[\hat{f}_{n,h_n}] + (\mathbb{E}[\hat{f}_{n,h_n}] - f)^2$$

First, we calculate the bias:

$$\mathbb{E}[\hat{f}_{n,h_n}] - f = \frac{F(x+\frac{h_n}{2}) - F(x-\frac{h_n}{2})}{h_n} - f(x) \overset{h_n \to 0}{\longrightarrow} F'(x) - f(x) = f(x) - f(x) = 0$$

Then

$$\begin{aligned}
\text{Var}[\hat{f}_{n,h_n}] &= \frac{F(x+\frac{h_n}{2}) - F(x-\frac{h_n}{2}) + (F(x+\frac{h_n}{2}) - F(x-\frac{h_n}{2}))^2}{nh_n^2} \\
&\xrightarrow{h_n \to 0} \frac{f(x)}{nh_n} + \frac{f^2(x)}{n} \\
&\xrightarrow{n \to \infty} 0
\end{aligned}$$

9. **Estimating a pmf: strong consistency.** This is a continuation of Problem 5:

(a) For fixed i, use the strong law of large numbers to show that $\hat{p}_i^{\text{MLE}}$ is strongly consistent for p_i, meaning $\hat{p}_i^{\text{MLE}} \to p_i$ almost surely as $n \to \infty$.

From Problem 5, $\hat{p}_i^{\text{MLE}} = \frac{1}{n}\sum_{j=1}^{n} \mathbb{1}\{X_j = i\}$.

Let $Y_j = \mathbb{1}\{X_j = i\} \sim \text{Bernoulli}(p_i)$ for $j = 1 : n$.

By the Strong Law of Large numbers,

$$\frac{1}{n}\sum_{j=1}^{n} \mathbb{1}\{X_j = i\} = \bar{Y} \xrightarrow{a.s.} \mathbb{E}[Y_i] = p_i$$

so, in particular,

$$\hat{p}_i^{\text{MLE}} \xrightarrow{a.s.} p_i$$

(b) For a pmf q and a set $B \subseteq \mathcal{I}$ we define $q(B) \triangleq \sum_{i \in B} q_i$ to be the probability that the pmf assigns to the set B. For a fixed set B, show that $\hat{p}^{\text{MLE}}(B)$ is strongly consistent for $p(B)$ as $n \to \infty$.

$$\begin{aligned}
\hat{p}^{\text{MLE}}(B) &= \sum_{i \in B} \hat{p}_i^{\text{MLE}} \\
&= \sum_{i \in B} \frac{1}{n} \sum_{j=1}^{n} \mathbb{1}\{X_j = i\} \\
&= \frac{1}{n} \sum_{j=1}^{n} \sum_{i \in B} \mathbb{1}\{X_j = i\} \\
&\xrightarrow{a.s.} \mathbb{E}\left[\sum_{i \in B} \mathbb{1}\{X_j = i\}\right] \\
&= \sum_{i \in B} \mathbb{E}[\mathbb{1}\{X_j = i\}] \\
&= \sum_{i \in B} p_i = p(B)
\end{aligned}$$

(c) Show that $\hat{p}^{\text{MLE}}$ is strongly consistent for p in total variation distance, meaning that

$$\text{TV}(\hat{p}^{\text{MLE}}, p) \triangleq \sup_{B \subseteq \mathcal{I}} |\hat{p}^{\text{MLE}}(B) - p(B)| \to 0$$

almost surely as $n \to \infty$.

$$\begin{aligned}
\text{TV}(\hat{p}^{\text{MLE}}, p) &= \sup_{B \subseteq \mathcal{I}} \left|\hat{p}^{\text{MLE}}(B) - p(B)\right| \\
&\xrightarrow{a.s.} \sup_{B \subseteq \mathcal{I}} |p(B) - p(B)| \\
&= \sup_{B \subseteq \mathcal{I}} 0 = 0
\end{aligned}$$

10. **Another kind of cross validation and an introduction to order statistics.** Cross validation is a general approach to assessing the performance of an estimator without the necessity of collecting additional "out-of-sample" data. An example is the leave-one-out method for choosing a smoothing parameter, such as the bandwidth h in kernel density estimation. The idea is to ignore one of the samples, form the estimator from the remaining samples, and evaluate performance based on the unused sample. This can be repeated for every observation, and the results can be pooled.

If, for example, we are looking for a good bandwidth h, then the entire process can be repeated for each of a selection of bandwidths, and the empirically best bandwidth chosen. (Variations include the leave-k-out method, in which each evaluation is based on $k > 1$ samples that were removed in forming the estimator.)

This can work very well or very poorly. The MISE approach explored in previous problems (and supported by Stone's theorem) often works very well. Here we will look at a similar approach using likelihoods that often works very poorly.

Start with the n leave-one-out estimators for f used in Problem 4:

$$\left\{\hat{f}_h^{(i)}(x; {}_iX)\right\}_{i=1:n} \quad \text{where} \quad \hat{f}_h^{(i)}(x; {}_iX) = \frac{1}{n-1}\sum_{\substack{k=1:n\\k\neq i}} \kappa_h(x - X_k)$$

(For convenience, we will write $\hat{f}_h^{(i)}(x; {}_iX)$, and occasionally just $\hat{f}^{(i)}$, instead of $\hat{f}_{n-1,h}^{(i)}(x; {}_iX)$.) In problem 4, $\hat{f}^{(i)}$ is used to make an estimate of the ISE for a given value of h.

A compelling alternative is to use $\hat{f}^{(i)}$, instead, to estimate the likelihood:

$$\tilde{J}(h) = \prod_{l=1}^{n} \hat{f}_h^{(i)}(X_i; {}_iX)$$

and then search for h to *maximize* $\tilde{J}(h)$. After all, in many typical scenarios the maximum likelihood estimator has a number of excellent properties, and $\tilde{J}$ would appear to be a sensible substitute.

In this problem, we will assume that the true (but unknown) density is the exponential density with parameter one: $f(x) = e^{-x}$ on $x \geq 0$. The estimator is

$$\hat{f}_h(x; X_{1:n}) = \frac{1}{n}\sum_{k=1}^{n} \kappa_h(x - X_k)$$

where κ is the box-car kernel used in the previous homework, $\kappa(x) = \mathbb{1}_{|x|\leq 1/2}$, and hence $\kappa_h(x) = \frac{1}{h}\mathbb{1}_{|x|\leq h/2}$.

For each n, let $h_n^* = \text{argmax}\tilde{J}(h)$. The problem is to show that $\hat{f}_{h_n^*}$ *cannot* be a consistent estimator for f.

Here is one way to do this:

(a) Let $X_{(1)} < X_{(2)} < \cdots < X_{(n)}$ be the *order statistics* of $X_1, \ldots, X_n$, so that $X_{(1)} = \min_k X_k$, $X_{(2)} = \min_{k:X_k > X_{(1)}} X_k$, $X_{(3)} = \min_{k:X_k > X_{(2)}} X_k$, and so on, up to $X_{(n)} = \min_{k:X_k > X_{(n-1)}} X_k = \max_k X_k$. Derive the joint density of the n order statistics:

$$f_{X_{(1)},\ldots,X_{(n)}}(x_1, \ldots, x_n) = \begin{cases} n! \prod_{k=1}^{n} e^{-x_k} & \text{if } x_1 < x_2 < \ldots < x_n \\ 0 & \text{otherwise} \end{cases}$$

(b) Change variables: Let $Y_{(i)} = X_{(i)}$ for $i = 1, \ldots, n-1$ and $Y_{(n)} = X_{(n)} - X_{(n-1)}$, and compute the joint density of $Y_{(1)}, Y_{(2)}, \ldots, Y_{(n)}$.[7]

(c) Compute the marginal density on $Y_{(n)}$.

[7] The change-of-variables formula for densities is the same one you use for integrals (and for the same reasons): If the random vector $\vec{X} \in \mathbb{R}^n$ has density $\alpha(\vec{x})$, and if $\vec{Y} = m(\vec{X})$ for some smooth and one-to-one function $m : \mathbb{R}^n \to \mathbb{R}^n$, with smooth inverse m^{-1}, then $\vec{Y}$ has density

$$\beta(\vec{y}) = \alpha(m^{-1}(\vec{y}))\left|\frac{\partial x_{1:n}}{\partial y_{1:n}}\right|$$

where $\left|\frac{\partial x_{1:n}}{\partial y_{1:n}}\right|$ is the determinant of the Jacobian.

(d) Show that

$$h_n^* = \text{argmax}_h \prod_{i=1}^{n} \hat{f}_h^{(i)}(X_i; {}_iX)$$

does not converge to zero. (Hint: show that $\text{argmax}_h \prod_{i=1}^{n} \hat{f}_h^{(i)}(X_i; {}_iX) = 0$ whenever $h < 2Y_{(n)}$.)

In this formulation of cross validation, consistency requires a delicate balance between the tails of the kernel and the tails of the (unknown) density f. Basically, the tails of f cannot be "too heavy" relative to the tails of κ.