1. [20 points] Non-parametric density estimation: In class we discussed non-parametric regression. In non-parametric density estimation, you observe $X_1, \ldots, X_n \sim f$, and we want to estimate $f$. We assume that $f$ is Lipschitz smooth (as in Lecture on non-parametric regression) and use the estimator:

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x-X_i}{h}\right).$$

Consider the boxcar kernel:

$$K(x) = \begin{cases} 1 & \frac{-1}{2} < x < \frac{1}{2} \\ 0 & \text{otherwise.} \end{cases}$$

(a) Show that:

$$\mathbb{E}(\hat{f}(x)) = \frac{1}{h} \int_{x-(h/2)}^{x+(h/2)} f(y) dy.$$ 

Solution.

We can find the expectation as:

$$\mathbb{E}\left[\hat{f}(x)\right] = \mathbb{E}\left[\frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x-X_i}{h}\right)\right]$$

$$= \frac{1}{h} \mathbb{E}\left[K\left(\frac{x-X_l}{h}\right)\right]$$

$$= \frac{1}{h} \int_{-\infty}^{\infty} K\left(\frac{x-y}{h}\right) f(y) dy$$

$$= \frac{1}{h} \int_{-\infty}^{\infty} I\left(\frac{-1}{2} < \frac{x-y}{h} < \frac{1}{2}\right) f(y) dy$$

$$= \frac{1}{h} \int_{x-h/2}^{x+h/2} f(y) dy.$$
(b) Show also that:

\[ \text{Var}(\hat{f}(x)) = \frac{1}{nh^2} \left[ \int_{x-(h/2)}^{x+(h/2)} f(y) dy - \left( \int_{x-(h/2)}^{x+(h/2)} f(y) dy \right)^2 \right]. \]

**Solution.**

We can find the variance as:

\[
\text{Var}\left(\hat{f}(x)\right) = \text{Var}\left(\frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x-X_i}{h}\right)\right)
\]

\[
= \frac{1}{nh^2} \text{Var}\left(K\left(\frac{x-X_1}{h}\right)\right)
\]

\[
= \frac{1}{nh^2} \mathbb{P}_Y\left(-\frac{1}{2} < \frac{x-Y}{h} < \frac{1}{2}\right) \left[ 1 - \mathbb{P}_Y\left(-\frac{1}{2} < \frac{x-Y}{h} < \frac{1}{2}\right) \right]
\]

\[
= \frac{1}{nh^2} \left[ \int_{x-\frac{h}{2}}^{x+\frac{h}{2}} f(y) \, dy \left[ 1 - \int_{x-\frac{h}{2}}^{x+\frac{h}{2}} f(y) \, dy \right] \right]
\]

\[
= \frac{1}{nh^2} \left[ \int_{x-\frac{h}{2}}^{x+\frac{h}{2}} f(y) \, dy - \left( \int_{x-\frac{h}{2}}^{x+\frac{h}{2}} f(y) \, dy \right)^2 \right].
\]

(c) Show that if \( h \to 0 \) and \( nh \to \infty \), as \( n \to \infty \) then \( \hat{f}(x) \to f(x) \) in probability.

**Solution.**

Consider the bias-variance decomposition:

\[
\mathbb{E}\left[\left(\hat{f}(x) - f(x)\right)^2\right] = \left(\mathbb{E}\left[\hat{f}(x)\right] - f(x)\right)^2 + \text{Var}\left(\hat{f}(x)\right) = \text{Bias}^2 + \text{Variance}.
\]

To begin with the bias term, note that

\[
\frac{1}{h} \int_{x-\frac{h}{2}}^{x+\frac{h}{2}} f(y) dy - f(x) = \frac{1}{h} \int_{x-\frac{h}{2}}^{x+\frac{h}{2}} [f(y) - f(x)] dy
\]

\[
\leq \frac{1}{h} \int_{x-\frac{h}{2}}^{x+\frac{h}{2}} |f(y) - f(x)| dy
\]

\[
\leq \frac{1}{h} \int_{x-\frac{h}{2}}^{x+\frac{h}{2}} L|y - x| dy = \frac{Lh}{4} \tag{1}
\]
where (∗) uses the Lipschitz condition. In detail, by Taylor’s theorem, there exists ξ between y and x such that 

\[ f(y) = f(x) + f'(\xi)(y - x). \]

Thus

\[ |f(y) - f(x)| = |f'(\xi)(y - x)| \leq L|y - x|. \]

For the variance, we can simply bound

\[
\operatorname{Var}\{ \hat{f}(x) \} \leq \frac{1}{nh^2} \int_{x-h/2}^{x+h/2} f(y)dy = \frac{1}{nh} \left[ \frac{1}{h} \int_{x-h/2}^{x+h/2} f(y)dy + f(x) \right] \leq \frac{Lh}{4} + f(x) = \frac{1}{nh}f(x) + \frac{L}{4n},
\]

where (∗) uses the bound (1). Therefore

\[
E \left[ (\hat{f}(x) - f(x))^2 \right] \leq \frac{L^2h^2}{16} + \frac{1}{nh}f(x) + \frac{L}{4n} \to 0 \text{ as } h \to 0 \text{ and } nh \to \infty.
\]

Using Chebyshev’s inequality yields \( \hat{f}(x) \to f(x) \) in probability.

2. [20 points] Regression with Errors-in-variables: A common problem in regression is that we have errors in the observed covariates (a very similar situation arises when we have missing covariates). This can in some cases lead to a very difficult (statistically) regression problem, but we will consider a simple case.

Suppose that we have a 1D regression problem where \( y_i \) and \( X_i \) are linked via a linear model,

\[ y_i = \beta X_i + \epsilon_i, \]

where \( \epsilon_i \sim N(0, \sigma^2) \), and \( X_i \) are i.i.d., have finite second moment, and \( X_i \in \mathbb{R} \). Rather than observe \( X_i \) however we observe a noisy version \( W_i \) where,

\[ W_i = X_i + \delta_i, \]

and \( \delta_i \sim N(0, \tau^2) \), independently of everything else. So we observe \( n \) i.i.d samples of the form \( \{(y_1, W_1), \ldots, (y_n, W_n)\} \). Consider the least squares estimator:

\[
\hat{\beta} = \frac{\sum_{i=1}^{n} Y_i W_i}{\sum_{i=1}^{n} W_i^2},
\]

and show that the estimator is inconsistent, i.e. show that \( \hat{\beta} \not\to a\beta \) where \( a \neq 1 \). Find \( a \). Suppose that \( \tau \) was known to you. Construct a consistent estimator of \( \beta \).
We have $y_i = \beta x_i + \epsilon_i$, where $\epsilon_i \sim N(0, \sigma^2)$ and $W_i = X_i + \delta_i$, where $\delta_i \sim N(0, \tau^2)$. Therefore, the least squares estimator can be written as,

$$\hat{\beta} = \frac{\sum^n_{i=1} Y_i W_i}{\sum^n_{i=1} W_i^2}$$

$$= \frac{\sum^n_{i=1} (\beta X_i + \epsilon_i)(X_i + \delta_i)}{\sum^n_{i=1} (X_i + \delta_i)^2}$$

$$= \frac{\sum^n_{i=1} \beta X_i^2}{\sum^n_{i=1} (X_i + \delta_i)^2} + \frac{\sum^n_{i=1} X_i (\beta \delta_i + \epsilon_i)}{\sum^n_{i=1} (X_i + \delta_i)^2} + \frac{\sum^n_{i=1} \delta_i \epsilon_i}{\sum^n_{i=1} (X_i + \delta_i)^2}.$$  

Now notice that

$$E[(X_i + \delta_i)^2] = E[X_i^2] + E[\delta_i^2] = E[X_i^2] + \tau^2.$$ 

Therefore by weak law of large numbers and continuous mapping theorem,

$$\frac{\sum^n_{i=1} \beta X_i^2}{\sum^n_{i=1} (X_i + \delta_i)^2} \xrightarrow{p} \beta \frac{E[X_i^2]}{E[X_i^2] + \tau^2}.$$ 

Similarly notice that

$$E[X_i(\beta \delta_i + \epsilon_i)] = E[X_i]E[\beta \delta_i + \epsilon_i] = 0,$$

$$E[\delta_i \epsilon_i] = E[\delta_i]E[\epsilon_i] = 0.$$  

Therefore, by weak law of large numbers and continuous mapping theorem,

$$\frac{\sum^n_{i=1} X_i (\beta \delta_i + \epsilon_i)}{\sum^n_{i=1} (X_i + \delta_i)^2} \xrightarrow{p} 0 \text{ and } \frac{\sum^n_{i=1} \delta_i \epsilon_i}{\sum^n_{i=1} (X_i + \delta_i)^2} \xrightarrow{p} 0.$$ 

Hence,

$$\hat{\beta} \xrightarrow{p} \beta \frac{E[X_i^2]}{E[X_i^2] + \tau^2} = a \beta,$$

where

$$a = \frac{E[X_i^2]}{E[X_i^2] + \tau^2} \neq 1.$$  

If $\tau$ were known to us, we could instead consider the estimator

$$\hat{\beta}^* = \frac{\frac{1}{n} \sum^n_{i=1} X_i^2 + \tau^2}{\frac{1}{n} \sum^n_{i=1} X_i^2} \hat{\beta},$$

Then by weak law of large numbers and continuous mapping theorem again,

$$\frac{1}{n} \sum^n_{i=1} X_i^2 + \tau^2 \xrightarrow{p} \frac{E[X_i^2] + \tau^2}{E[X_i^2]} \text{ and thus } \hat{\beta}^* \xrightarrow{p} \beta$$

as desired.
3. [20 points] Model Selection: In our next lecture we will give the form of the AIC correction but will not provide much justification. In this question, we will try to explore its roots in a simple context.

Suppose we observe \( T = \{X_1, \ldots, X_n\} \sim \mathcal{N}(\theta, I) \), where \( \theta \in \mathbb{R}^d \) and we want to estimate \( \theta \), and want to understand the optimism of the log-likelihood of the training data. For our purposes the negative of the training log-likelihood (up to constants) is just given by:

\[
\hat{\ell}(\hat{\theta}) = \frac{1}{2n} \sum_{i=1}^{n} \|X_i - \hat{\theta}\|_2^2.
\]

Suppose that \( \hat{\theta} \) is the MLE, i.e. \( \hat{\theta} = \frac{1}{n} \sum_{i=1}^{n} X_i \). Show that,

\[
\mathbb{E}_T[\hat{\ell}(\hat{\theta})] = \frac{d}{2} - \frac{d}{2n}.
\]

On the other hand define the negative of the true log-likelihood as,

\[
\ell(\hat{\theta}) = \frac{1}{2} \mathbb{E}_{T,X} \|X - \hat{\theta}\|_2^2,
\]

where \( X \) is a new sample independent of \( T \). Show that,

\[
\ell(\hat{\theta}) = \frac{d}{2} + \frac{d}{2n}.
\]

This shows that the training log-likelihood has an upward bias (or the training loss on average appears lower than it really is) and that this bias depends on the complexity of the parameter we are estimating (in this case, we are estimating a \( d \)-dimensional mean). Use this to justify the AIC correction to the training log-likelihood in this setting.

Solution.

First we show that \( \mathbb{E}_T[\hat{\ell}(\hat{\theta})] = d/2 - d/(2n) \). Since \( \hat{\ell}(\hat{\theta}) \) is invariant to location transformations, we assume \( \theta = 0 \). Then

\[
(X_1 - \hat{\theta})^\top (X_1 - \hat{\theta}) = X_1^\top X_1 - 2X_1^\top \hat{\theta} + \hat{\theta}^\top \hat{\theta}
\]

\[
= X_1^\top X_1 - 2X_1^\top \left( \frac{1}{n} X_1 + \frac{1}{n} \sum_{i=2}^{n} X_i \right) + \hat{\theta}^\top \hat{\theta}.
\]

Note that \( \hat{\theta} \sim \mathcal{N}(0, n^{-1}I) \) and so

\[
\mathbb{E}_T[(X_1 - \hat{\theta})^\top (X_1 - \hat{\theta})] = d - \frac{2d}{n} + \frac{d}{n} = d - \frac{d}{n}.
\]
This gives
\[ E_T[\ell(\hat{\theta})] = \frac{1}{2n} \sum_{i=1}^{n} E \| X_i - \hat{\theta} \|^2 = \frac{d}{2} - \frac{d}{2n}. \]

Next we show that \( \ell(\hat{\theta}) = d/2 + d/(2n) \). Since \( X \) and \( \hat{\theta} \) are independent in the second case, we have
\[ \frac{1}{2} E_{T,X} \left[ (X - \hat{\theta})^\top (X - \hat{\theta}) \right] = \frac{1}{2} E_{T,X} \left[ X^\top X - 2X^\top \hat{\theta} + \hat{\theta}^\top \hat{\theta} \right] = \frac{d}{2} + \frac{d}{2n} \]
where we assumed \( \theta = 0 \) without loss of generality. This proves the second statement.

For this example where the dimension of the parameter space is \( d \), the AIC is given as
\[ AIC = 2n \left( \frac{1}{n} \sum_{i=1}^{n} \log p(X_i; \hat{\theta}) - \frac{d}{n} \right). \]

By taking the expectation and using the previous results, we have
\[ E[AIC] = 2n \left( \frac{1}{n} \sum_{i=1}^{n} \log p(X_i; \hat{\theta}) - \frac{d}{n} + C \right) = 2n \left( -\frac{d}{2} + \frac{d}{2n} - \frac{d}{n} + C \right) = 2n \left( -\ell(\hat{\theta}) + C \right) \]
for some constant \( C \). This result shows that the AIC removes the bias of the training log-likelihood and therefore provides justification of the AIC.

4. [20 points] Weak \( \ell_q \) sparsity: In lecture we showed that the hard-thresholding algorithm for the Gaussian Sequence Model achieves the rate,
\[ R(\hat{\theta}, \theta^*) \leq CR\sigma \sqrt{\log \frac{d}{n}}, \]
for estimating a vector \( \theta^* \) that is \( \ell_1 \)-sparse, i.e. satisfies \( \sum_{i=1}^{d} |\theta_i| \leq R \) (where \( C > 0 \) is some constant). Suppose that, \( \theta^* \) is instead \( \ell_q \) sparse for some \( q \in (0, 1] \), i.e.
\[ \sum_{i=1}^{d} |\theta_i|^q \leq R_q. \]
Then show that for some constant \( C > 0 \) the hard-thresholding estimator has risk:
\[ R(\hat{\theta}, \theta^*) \leq CR_q \left( \frac{\sigma^2 \log d}{n} \right)^{1-q/2}. \]
Notice that once again the estimator can have risk \( \rightarrow 0 \) even when \( d \gg n \). Furthermore, notice that the same hard-thresholding estimator works for a variety of different notions of sparsity.
Solution:

If we assume that for some $q \in (0, 1]$, $\theta^*$ satisfies

$$\sum_{i=1}^{d} |\theta_i|^q \leq R_q,$$

for some radius $R_q$, then the number of entries of the form $|\theta_i|^q$ larger than $R_q/k$ is at most $k$, for any $k$. Therefore the number of entries $|\theta_i|$ larger than $(R_q/k)^{1/q}$ is at most $k$, for any $k$. Then we can use the hard thresholding estimator’s bound in the lecture notes to obtain:

$$R(\hat{\theta}, \theta) \lesssim \sum_{i=1}^{d} \min \left\{ \theta_i^2, \frac{\sigma^2 \log(d)}{n} \right\}$$

$$\lesssim \sum_{i: \theta_i^2 \geq \frac{\sigma^2 \log(d)}{n}} \frac{\sigma^2 \log(d)}{n} + \sum_{i: \theta_i^2 \leq \frac{\sigma^2 \log(d)}{n}} \theta_i^2.$$ 

Here the symbol “$\lesssim$” means that the inequality holds up to a constant factor.

Since the number of entries, $|\theta_i|$ that can exceed $\sigma \sqrt{\log(d)/n}$ is at most $R_q \left( \sqrt{\frac{n}{\sigma^2 \log(d)}} \right)^q$, we obtain the bound that,

$$R(\hat{\theta}, \theta) \lesssim R_q \left( \sqrt{\frac{n}{\sigma^2 \log(d)}} \right)^q \frac{\sigma^2 \log(d)}{n} + \sum_{i: \theta_i^2 \leq \frac{\sigma^2 \log(d)}{n}} \theta_i^2.$$ 

5. [20 points] High-dimensional Estimation in $\ell_\infty$: Recall the Gaussian sequence model where we observe $\{y_1, \ldots, y_d\}$ where for each $i$,

$$y_i = \theta_i + \epsilon_i,$$

and $\epsilon_i \sim N(0, \sigma^2/n)$. For any vector $v \in \mathbb{R}^d$ we define its $\ell_\infty$-norm as:

$$\|v\|_\infty = \max_{i=1}^{d} |v_i|.$$
Suppose our goal is to estimate the vector $\theta$ well in the $\ell_\infty$ norm, i.e. we want an estimator $\hat{\theta}$ such that $\|\hat{\theta} - \theta\|_\infty$ is small (with high-probability).

(a) What is the MLE for $\theta$?

**Solution:**

First notice that the likelihood function is

$$L(\theta_1, \ldots, \theta_d; y_1, \ldots, y_d) = \prod_{i=1}^d \frac{1}{\sqrt{2\pi \sigma}} \frac{1}{\sqrt{n}} \exp \left( - \frac{n(y_i - \theta_i)^2}{2\sigma^2} \right),$$

which is maximized when $\theta_i = y_i$ for $i = 1, \ldots, d$. Hence the MLE for $\theta$ is $\hat{\theta}_{MLE,i} = y_i$ for $i = 1, \ldots, d$.

(b) Using the appropriate concentration inequality analyze, the MLE, i.e. give a reasonable bound on $\|\hat{\theta}_{MLE} - \theta\|_\infty$ that holds with probability at least $1 - \delta$.

**Solution:**

For $t > 0$, using the union bound yields

$$\mathbb{P}\left(\|\hat{\theta}_{MLE} - \theta\|_\infty > t\right) = \mathbb{P}\left(\max_{i=1}^d |y_i - \theta_i| > t\right) \leq \sum_{i=1}^d \mathbb{P}(|y_i - \theta_i| > t).$$

Since $y_i \sim N(\theta_i, \sigma^2/n)$, the two-sided Gaussian tail bound gives

$$\mathbb{P}(|y_i - \theta_i| > t) \leq 2 \exp\left(-nt^2/(2\sigma^2)\right).$$

Combining the two results, we have that

$$\mathbb{P}\left(\|\hat{\theta}_{MLE} - \theta\|_\infty > t\right) \leq 2d \exp\left(-nt^2/(2\sigma^2)\right).$$

By setting the right-hand side to be $\delta$ and rearranging the terms, we conclude that

$$\mathbb{P}\left(\|\hat{\theta}_{MLE} - \theta\|_\infty \leq \sqrt{\frac{2\sigma^2}{n} \log \left(\frac{2d}{\delta}\right)}\right) \geq 1 - \delta.$$

(c) Does the MLE make sense in high-dimensions? Concretely, suppose that $d \gg n$ (for simplicity, you can take $d = n^{100}$ so that $d$ grows as $n$ gets large), and fix a small value $\delta$, and analyze the error $\|\hat{\theta}_{MLE} - \theta\|_\infty$. Does this error become small as
$n$ grows (remember that $d$ is also getting large)? To be clear, we are not assuming sparsity or any other structure on $\theta$.

Solution:

For a fixed $\delta$, the upper bound for the $\ell_\infty$ norm converges to zero as $n, d \to \infty$ provided that $\log(d)/n \to 0$. For instance, if $d = n^{100}$, we see that

$$\sqrt{\frac{2\sigma^2}{n} \log \left( \frac{2d}{\delta} \right)} \lesssim \sqrt{\frac{\sigma^2 \log(n)}{n}} \to 0 \quad \text{as } n, d \to \infty.$$ 

Therefore, even if $d \gg n$, it is possible to consistently estimate $\theta$ in the $\ell_\infty$ norm without making an assumption on $\theta$. 
