Problem 1 [5 pts.]

(a) List all the independence statements implied by this graph.

(b) Find the causal distribution \( p(x_4 \mid \text{set } x_3 = s) \).

(c) Find the implied undirected graph for these random variables. Which independence statements get lost in the undirected graph (if any)?

Solution.

The graph can be visualized as follows:

```
  X1    X2
    /\    /\ 
   /  \  /  \ 
  X3   X3   X3
    \  /  \  /  \
   / \  / \  / \ 
  X4 X3 X4 X5
    \  /  /  /  \
   /  \  \  \  
  X5   X5   X5
```

(a) The independence statements implied are the following:

- \( X_1 \perp X_2 \);
- \( X_4 \perp \{X_1, X_2\} \mid X_3 \) and \( X_5 \perp \{X_1, X_2\} \mid X_3 \);
- \( X_5 \perp X_4 \mid X_3 \).

(b) Given that we set \( X_3 = x_3 \), then by the independence highlighted above \( X_1 \) and \( X_2 \) can be dropped from the graph. Hence we have that:

\[
p(x_4 \mid \text{set } x_3 = s) = \int p_4(x_4, x_5) dx_5 = \int p(x_4 \mid x_3 = s)p(x_5 \mid x_3 = s) dx_5 = p(x_4 \mid x_3 = s)
\]

(c) The moralized graph becomes:
We lose the (unconditional) independence between $X_1$ and $X_2$ during the moralization process, while all the others are retained.
Problem 2 [20 pts.]

Let $d \geq 2$, and let $X_1, \ldots, X_n \sim P$ where $X_i = (X_i(1), \ldots, X_i(d)) \in \mathbb{R}^d$. Assume that the coordinates of $X_i$ are independent. Further, assume that $X_i(j) \sim \text{Bernoulli}(p_j)$ where $0 < c \leq p_j \leq C < 1$. Let $\mathcal{P}$ be all such distributions. Let

$$R_n = \inf_{\hat{p}} \sup_{p \in \mathcal{P}} \mathbb{E}_p \|\hat{p} - p\|_\infty.$$ 

Find lower and upper bounds on the minimax risk.

Solution.

For **upper bound**, consider an estimator $\overline{X}$ for estimating $p$. Then $\overline{X} - p$ is sub-Gaussian with parameter $\sigma^2 = \frac{1}{4n}$. Hence, from the lemma below,

$$\mathbb{E}\|\overline{X} - p\|_\infty = \mathbb{E}\left[\max_{1 \leq i \leq d}\overline{X}_i - p_i\right] \leq \frac{1}{2\sqrt{n}} \sqrt{2\log(2d)}.$$

And since $d \geq 2$, we have $2\log(2d) \leq 4\log d$, so,

$$\inf_{\hat{p}} \sup_{p \in \mathcal{P}} \mathbb{E}_p \|\hat{p} - p\|_\infty \leq \mathbb{E}\|\overline{X} - p\|_\infty \leq \sqrt{\frac{\log d}{n}}.$$

For **lower bound**, let $\alpha = \sqrt{\frac{\log d}{10n}}$, $p^{(0)} = (c, \ldots, c) \in \mathbb{R}^d$.

And for $1 \leq j \leq d$, construct $d$-dimensional vector parameter

$$p_j := (p^{(i)}_j, i = 1, \cdots d) = (c, \cdots, c, c + \alpha, c, \cdots, c) \in \mathbb{R}^d,$$

then $\|p_j - p_k\|_\infty = \alpha$ for $j \neq k$, so

$$\min_{j \neq k} \|p_j - p_k\|_\infty = \alpha.$$

Let $P_j$ be the multivariate Bernoulli with parameter $p_j$, then it’s a product of uni-variate Bernoulli’s, $P_j = \prod_{i=1}^d P_j^{(i)}$, where $P_j^{(i)}$ is uni-variate Bernoulli with parameter $p^{(i)}_j$. Then

$$KL(P^{(j)}, P^{(k)}) = \sum_{i=1}^d KL(P_j^{(i)}, P_k^{(i)}) = KL(P_j^{(j)}, P_k^{(k)}) + KL(P_j^{(j)}, P_k^{(k)}).$$

since they only differ in two terms: $i = j$ or $i = k$. Let uni-variate Beroulli with parameter $c$ as $P_c$, then,

$$KL(P^{(j)}, P^{(k)}) = KL(P_j^{(j)}, P_k^{(k)}) + KL(P_j^{(j)}, P_k^{(k)})$$

$$= \sum_x P_{c+\alpha} \log\left(\frac{P_{c+\alpha}}{P_c}\right) + \sum_x P_c \log\left(\frac{P_c}{P_{c+\alpha}}\right)$$

$$= \sum_x (P_{c+\alpha} - P_c) \log\left(\frac{P_{c+\alpha}}{P_c}\right)$$

$$= \alpha \log \frac{(\alpha + c)(1-c)}{(1-\alpha - c)c}$$

$$\leq C\alpha^2,$$ by Taylor expansion.
(Fano’s method)

\[
\max_{j \neq k} KL(P_j, P_k) = \max_{j \neq k} \frac{1}{2} ||\mu_j - \mu_k||^2 = C\alpha^2
\]

\[
= \frac{\log d}{8n} \leq \frac{\log(d + 1)}{4n}.
\]

Hence by Corollary 13 in the minimax notes where \( N = d + 1 \),

\[
\inf \sup_{\hat{p}, P \in P} \mathbb{E}[p||\hat{p} - p||_{\infty} \geq \alpha / 4 = \sqrt{\frac{\log d}{256n}}.
\]

Note: following the same construction of \( p^{(i)} \) and KL distance bound, lower bounds for minimax with the same rate with respect to \( d \) and \( n \) can also be derived using Theorem 12 or Theorem 14.

**Lemma 1 (Maximal inequality for subgaussian random variables)**

Let \( \{X_i\}_{1 \leq i \leq n} \) be sub-Gaussian variables with parameter \( \sigma^2 \), then

\[
\mathbb{E}\left[\max_{1 \leq i \leq n} X_i\right] \leq \sigma \sqrt{2 \log n} \quad \text{and} \quad \mathbb{E}\left[\max_{1 \leq i \leq n} |X_i|\right] \leq \sigma \sqrt{2 \log(2n)}.
\]

It’s covered in Advanced stats.

Problem 3 [20 pts.]
Let \( \{ p_\theta : \theta \in \Theta \} \) where \( \Theta \subset \mathbb{R} \) be a parametric model. Suppose that the model satisfies the usual regularity conditions. In particular, the Fisher information \( I(\theta) \) is positive and smooth and the mle has the usual nice properties. Let the loss function be \( L(\hat{\theta}, \theta) = H(p_{\hat{\theta}}, p_\theta) \) where \( H \) denotes Hellinger distance. Find the minimax rate.

Solution.
Assume that the densities in \( \{ p_\theta : \theta \in \Theta \} \) is differentiable in quadratic mean (QMD) as equation (27) in the Minimax note, which says the Hellinger distance between \( p_{\theta + h}, p_\theta \) can be approximated by first order Taylor expansion on \( h \). In other words, the loss with Hellinger distance is then approximate equivalent with that with squared loss on \( \theta \),

\[
H^2(p_{\theta + h}, p_\theta) = \frac{1}{8} \| h \|^2 I(\theta) + o(\| h \|^2).
\]

So the minimax rate for Hellinger loss is the same as for squared risk

\[
\inf_{\hat{\theta}} \sup_{\theta \in \Theta} H^2(p_{\hat{\theta}}, p_\theta) = O\left( \frac{1}{n} \right),
\]

or equivalently,

\[
\inf_{\hat{\theta}} \sup_{\theta \in \Theta} H(p_{\hat{\theta}}, p_\theta) = O\left( \frac{n^{-1/2}}{n} \right).
\]
Alternative solution (thank Tim Barry for the ideas)

We first derive an upper bound for the KL distance between arbitrary $p_{\theta+h}, p_{\theta}$,

$$KL(p_{\theta+h}, p_{\theta}) = \int \log \frac{p_{\theta}}{p_{\theta+h}} dx$$

$$= \int \log p_{\theta} - \log p_{\theta+h} \frac{p_{\theta}}{p_{\theta+h}} dx$$

$$= \int \log p_{\theta} - \left( \log p_{\theta} + h \frac{\partial \log p_{\theta}}{\partial \theta} + h^2 \frac{\partial^2 \log p_{\theta}}{\partial \theta^2} + o(h^2) \right) p_{\theta} dx$$

$$= h^2 \left( - \int \frac{\partial^2 \log p_{\theta}}{\partial \theta^2} p_{\theta} dx \right) + o(h^2)$$

$$\leq CI(\theta) h^2,$$

where we assume good properties for the density to allow swapping the derivative and integral, so that,

$$\int \frac{\partial \log p_{\theta}}{\partial \theta} p_{\theta} dx = \int \frac{\partial p_{\theta}}{\partial \theta} p_{\theta} dx = \int \frac{\partial p_{\theta}}{\partial \theta} dx = \int \frac{\partial (\int p_{\theta} dx)}{\partial \theta} = 0.$$

For upper bound, consider MLE estimator,

$$H(p_{\hat{\theta}_{mle}}, p_{\theta}) \leq \sqrt{KL(p_{\hat{\theta}_{mle}}, p_{\theta})} \leq C(\hat{\theta}_{mle} - \theta)^2 I(\theta),$$

By MLE asymptotic distribution,

$$\hat{\theta}_{mle} - \theta \rightarrow N(0, I^{-1}(\theta)/n),$$

thus,

$$\inf_{\theta} \sup_{\theta \in \Theta} H(p_{\hat{\theta}}, p_{\theta}) \leq H(p_{\hat{\theta}_{mle}}, p_{\theta}) \leq \sqrt{CI(\theta)(Var(\hat{\theta}_{mle}) + bias^2)} \rightarrow O(n^{-1/2}).$$

For lower bound, consider two distribution $p_{\theta}$ and $p_{\theta+h}$, where $h = \sqrt{\frac{\log 2}{CI(\theta)n}}$, then by QMD condition,

$$H(p_{\theta+h}, p_{\theta}) = Ch = O(n^{-1/2}).$$

And the KL distance,

$$KL(p_{\theta+h}, p_{\theta}) \leq CI(\theta) h^2 \leq \frac{\log 2}{n}.$$ 

Hence by Corollary 5 in the minimax notes,

$$\inf_{\theta} \sup_{\theta \in \Theta} H(p_{\hat{\theta}}, p_{\theta}) = O(n^{-1/2}).$$
Problem 4 [15 pts.]

Let \( Y = (Y_1, \ldots, Y_d) \sim N(\theta, I) \) where \( \theta = (\theta_1, \ldots, \theta_d) \). Assume that \( \theta \in \Theta = \{ \theta \in \mathbb{R}^d : \|\theta\|_0 \leq 1 \} \). Let
\[
R_d = \inf_{\hat{\theta}} \sup_{\theta \in \Theta} \mathbb{E}\|\hat{\theta} - \theta\|^2.
\]

Show that \( c \log d \leq R_d \leq C \log d \) for some constants \( c \) and \( C \).

Solution.

For the upper bound, we first prove a high-probability bound lemma for the maximum of Gaussians (from 36-705, Lecture 27, Fall 2018):

**Lemma 2** Suppose that, \( \epsilon_1, \ldots, \epsilon_d \sim N(0, \sigma^2) \) then with probability at least \( 1 - \delta \),
\[
\max_{i=1}^d |\epsilon_i| \leq \sigma \sqrt{2 \log(2d/\delta)}.
\]

**Proof.** By the Gaussian tail bound, if \( \epsilon \sim N(0, \sigma^2) \):
\[
P(|\epsilon| \geq t) \leq 2 \exp(-t^2/(2\sigma^2)),
\]
By using the union bound:
\[
P(\max_{i=1}^d |\epsilon_i| \geq t) \leq 2d \exp(-t^2/(2\sigma^2)),
\]
By setting \( 2d \exp(-t^2/(2\sigma^2)) = \delta \) we obtain the lemma.

Now, we assume \( \hat{\theta} \) to be the hard-thresholding estimator, defined as:
\[
\hat{\theta}_i = y_i \mathbb{I}(|y_i| \geq t), \quad \forall \ i \in \{1, \ldots, d\},
\]
We have the following theorem (from 36-705, Lecture 27, Fall 2018):

**Theorem 3** Suppose we choose the threshold:
\[
t = 2\sigma \sqrt{2 \log(2d/\delta)},
\]
then with probability at least \( 1 - \delta \),
\[
\|\hat{\theta} - \theta\|^2 \leq 9 \sum_{i=1}^d \min \left\{ \theta_i^2, \frac{t^2}{4} \right\} \leq C t^2
\]

For some \( C_1 > 0 \in \mathbb{R} \).

**Proof.** We condition on the event from the previous lemma, i.e. that
\[
\max_{i=1}^d |\epsilon_i| \leq \sigma \sqrt{2 \log(2d/\delta)} \leq \frac{t}{2}.
\]
Now, observe that,
\[
\|\hat{\theta} - \theta\|^2 = \sum_{i=1}^d (\hat{\theta}_i - \theta_i)^2,
\]
so we can consider each coordinate separately. Let us consider some cases:
1. If for any co-ordinate \(|\theta_i| \leq \frac{t}{2}\) our estimate is 0, so our risk for that coordinate is simply \(\theta_i^2\).

2. If \(|\theta_i| \geq \frac{3t}{2}\) our estimate is simply \(\hat{\theta}_i = y_i\) so our risk is simply \(\epsilon_i^2 \leq \frac{t^2}{4}\).

3. If \(\frac{t}{2} \leq |\theta_i| \leq \frac{3t}{2}\), then our risk,

\[
(\hat{\theta}_i - \theta_i)^2 = (y_i \mathbb{I}(|y_i| \geq t) - \theta_i)^2 = \theta_i^2 \mathbb{I}(|y_i| < t) + \epsilon_i^2 \mathbb{I}(|y_i| \geq t) \leq \max \{\epsilon_i^2, \theta_i^2\} \leq \frac{9t^2}{4}.
\]

Putting these together we see that,

\[
||\hat{\theta} - \theta||_2 \leq 9 \sum_{i=1}^{d} \min \left\{ \theta_i^2, \frac{t^2}{4} \right\} = 9 \sum_{\theta_i \leq 0} \min \left\{ \theta_i^2, \frac{t^2}{4} \right\} + 9 \sum_{\theta_i \geq 0} \min \left\{ \theta_i^2, \frac{t^2}{4} \right\} \leq C_1 t^2
\]

We also have that, under the same assumptions of the theorems above:

\[
||\hat{\theta} - \theta||_2 = \sum_{i=1}^{d} (y_i \mathbb{I}(|y_i| \geq t) - \theta_i)^2 \leq C_2d \max_{i=1,...,d} (t - \theta_i)^2 \leq C_2 dt^2
\]

And so we have that:

\[
\mathbb{E} \left[ ||\hat{\theta} - \theta||_2 \right] = \int_0^\infty \mathbb{P} (||\hat{\theta} - \theta||_2 > x) dx = \int_0^{C_1 t^2} \mathbb{P} (||\hat{\theta} - \theta||_2 > x) dx + \int_{C_1 t^2}^{C_2 dt^2} \mathbb{P} (||\hat{\theta} - \theta||_2 > x) dx \leq \int_0^{C_1 t^2} (1 - \delta) dx + \int_{C_1 t^2}^{C_2 dt^2} \delta dx \leq C_1 t^2 (1 - \delta) + t^2 (C_2d - 1) \delta \leq C_3 t^2 \approx C_3 \log d
\]

For the lower bound, let \(P_j = N(\theta_j, I)\) where \(\theta_0 = (0, \ldots, 0)\) and \(\theta_j\) is the \(d\)-dimensional vector \((d \geq 8)\) where

\[
\theta_j(k) = \begin{cases} \sqrt{\frac{\log d}{32}} & k = j \\ 0 & k \neq j. \end{cases}
\]

\(P_0\) is absolutely continuous wrt each other distribution and for all \(j = 1, \ldots, d\), we claim

\[
KL(P_j, P_0) \leq \frac{\log d}{16}.
\]

The statement above actually works for any \(P_j, P_i\) with \(i \neq j\). Let \(P_i = N(\mu_i, I)\), then:

\[
KL(P_j, P_k) = \int \frac{1}{2} \left( ||x - \mu_k||_2^2 - ||x - \mu_j||_2^2 \right) P_j(x) dx = \frac{1}{2} ||\mu_j - \mu_k||_2^2.
\]

Which implies:

\[
\max_{j \neq k} KL(P_j, P_k) = \max_{j \neq k} \frac{1}{2} ||\mu_j - \mu_k||_2^2 = 2 \left( \sqrt{\frac{\log d}{32}} \right)^2 = \frac{\log d}{16}
\]

8
It then follows that:

\[ \frac{1}{d} \sum_{j=1}^{d} KL(P_j, P_0) \leq \frac{\log d}{16}. \]

Therefore, by Tsybakov’s bound,

\[
R_d \geq \frac{s}{16} = \max_{j \neq k} \| \theta_j - \theta_k \|^2 \frac{16}{16} = 2 \sqrt{\frac{\log d^2}{32}} \frac{16}{16} \geq c \log d.
\]
Problem 5 [20 pts.]

Let \( X_1, \ldots, X_n \sim F \) where \( F \) is some distribution on \( \mathbb{R} \). Suppose we put a Dirichlet process prior on \( F \):

\[
F \sim \text{DP}(\alpha, F_0).
\]

(a) (10 pts.) Recall the stick-breaking construction. Show that

\[
\mathbb{E} \left( \sum_{j=1}^{\infty} W_j \right) = 1.
\]

(b) (10 pts.) Simulate \( n = 10 \) data points from a \( \mathcal{N}(0,1) \). Try three values of \( \alpha \): namely, \( \alpha = .1 \), \( \alpha = 1 \) and \( \alpha = 10 \). Compute the 95 percent Bayesian confidence band and the 95 percent DKW band. Plot the results for one example. Now repeat the simulation 1,000 times and report the coverage probability for each confidence band.

Solution.

(a) We start by showing \( \mathbb{P}(\sum_{j=1}^{\infty} W_j = 1) = 1 \). The expectation will then follow easily. First we prove a series of lemmas.

**Lemma 4** For all \( n \in \mathbb{N} \),

\[
1 - \sum_{j=1}^{n} W_j = \prod_{j=1}^{n} (1 - V_j).
\]

**Proof.** (by induction)

**Base case.** \( k = 1 \)

\[
1 - W_1 = 1 - V_1. \quad \checkmark
\]

**Inductive hypothesis.** Now assume that

\[
1 - \sum_{j=1}^{n-1} W_j = \prod_{j=1}^{n-1} (1 - V_j).
\]

**Inductive step.**

\[
1 - \sum_{j=1}^{n} W_j = 1 - \sum_{j=1}^{n-1} W_j - W_n
\]

\[
= \prod_{j=1}^{n-1} (1 - V_j) - V_n \prod_{j=1}^{n-1} (1 - V_j)
\]

\[
= (1 - V_n) \prod_{j=1}^{n-1} (1 - V_j)
\]

\[
= \prod_{j=1}^{n} (1 - V_j). \quad \checkmark
\]

**Lemma 5** Let \( v_1, v_2, \ldots \) be a sequence such that \( 0 < v_j < 1 \) for all \( j \). Then

\[
\prod_{j=1}^{\infty} (1 - v_j) > 0 \text{ if and only if } \sum_{j=1}^{\infty} v_j < \infty.
\]
Proof. First notice that
\[- \log \prod_{j=1}^{\infty} (1 - v_j) = - \sum_{j=1}^{\infty} \log (1 - v_j),\]
and thus
\[\prod_{j=1}^{\infty} (1 - v_j) > 0 \iff - \sum_{j=1}^{\infty} \log (1 - v_j) < \infty.\]
We now have
\[\left\{- \log (1 - v_j) \right\}_{j \in \mathbb{N}} > 0 \quad \text{and} \quad \left\{ v_j \right\}_{j \in \mathbb{N}} > 0,
so we can use the Limit Comparison test to prove that
\[- \sum_{j=1}^{\infty} \log (1 - v_j) < \infty \iff \sum_{j=1}^{\infty} v_j < \infty.
Since both series diverge when \(v_j \to 0\), it suffices to consider only the sequences where \(v_j \to 0\).

\[
\lim_{v_j \to 0} \frac{- \log (1 - v_j)}{v_j} = \lim_{v_j \to 0} \frac{1}{1 - v_j} = 1.
\]
Hence,
\[\sum_{j=1}^{\infty} \log (1 - v_j) < \infty \quad \text{if and only if} \quad \sum_{j=1}^{\infty} v_j < \infty,
and thus,
\[\prod_{j=1}^{\infty} (1 - v_j) > 0 \quad \text{if and only if} \quad \sum_{j=1}^{\infty} v_j < \infty.
Corollary 6 Let \(v_1, v_2, \ldots\) be a sequence such that \(0 < v_j < 1\) for all \(j\). Then
\[\prod_{j=1}^{\infty} (1 - v_j) = 0 \quad \text{if and only if} \quad \sum_{j=1}^{\infty} v_j = \infty.
Lemma 7 (Borel-Cantelli) If \(\sum_{j=1}^{\infty} \mathbb{P}(V_j > \epsilon) = \infty\) for some \(\epsilon > 0\), then \(\mathbb{P}(V_j > \epsilon \ i.o.) = 1\).
Now since
\[V_j \sim \text{Beta}(1, \alpha), \ j = 1, 2, \ldots
we have
\[\mathbb{P}(V_j > \epsilon) > 0 \quad \text{for all} \ \epsilon \in (0, 1) \quad \text{and} \quad j = 1, 2, \ldots
(4)
since the beta distribution puts positive mass over its entire support \((0, 1)\), and now (4) implies
\[\sum_{j=1}^{\infty} \mathbb{P}(V_j > \epsilon) = \infty \quad \text{for all} \ \epsilon \in (0, 1).
(5)
So altogether,

\[
P\left(\sum_{j=1}^{\infty} W_j = 1\right) \overset{\text{Lemma 2}}{=} \mathbb{P}\left(\prod_{j=1}^{\infty} (1 - V_j) = 0\right) = \mathbb{P}\left(\sum_{j=1}^{\infty} V_j = \infty\right) \overset{\text{Cor. 4}}{\geq} \mathbb{P}(V_j > \epsilon \ i.o.) \overset{\text{Lemma 5}}{=} 1,
\]

where the inequality comes from the fact

\[
\left\{ \sum_{j=1}^{\infty} V_j = \infty \right\} \supset \left\{ V_j > \epsilon \ i.o. \right\}.
\]

Thus,

\[
P\left(\sum_{j=1}^{\infty} W_j = 1\right) = 1. \quad \blacksquare
\]

It follows that

\[
\mathbb{E}\left(\sum_{j=1}^{\infty} W_j\right) = \int_{\mathbb{R}} \sum_{j=1}^{\infty} W_j dF
\]

\[
= \mathbb{P}\left(\sum_{j=1}^{\infty} W_j = 1\right) \cdot 1 + 0
\]

\[
= 1,
\]

where \( F \) is the distribution function of the random variable \( \sum_{j=1}^{\infty} W_j \).

(b) Two good resources for such simulation are the \texttt{distr} package in R, which is showcased by this tutorial. For Python 3 this tutorial uses the \texttt{pyMC3} modules to provide a achieve a similar goal. We include R code for a single simulation, with parts taken from the tutorial indicated above.

```r
library(distr)
library(coda)
library(latex2exp)

# Setup
set.seed(7)
n <- 10
alpha_vec <- c(0.1, 1, 10)
x_grid <- seq(-3, 3, by=0.05)
signif_level <- 0.05

# Sample observations
x_pts <- rnorm(n)

# Generate DKW Band
x_ecdf <- ecdf(x_pts)
x_ecdf_error <- sqrt(log(2 / signif_level) / (2 * n))
```
dkw.lb <- pmax(x_ecdf(x_grid) - x_ecdf_error, 0)
dkw.ub <- pmin(x_ecdf(x_grid) + x_ecdf_error, 1)

### BAYESIAN BANDS
# Functions to generate Bayesian Credible Bands
sample_cdf <- function(F_hat, n){
  F_hat@r(n) # F_hat is a S6 class object from the distr package
}

# Sampling from the prior distribution
sample_prior <- function(F0, alpha, n){
  cdf_sample <- sample_cdf(F0, n)
  v <- rbeta(n, 1, alpha) # See 5a) for definitions
  w <- c(v[1], rep(0, n-1))
  for(ii in 2:n){
    w[ii] <- v[ii]*cumprod(1-v)[ii-1] # See 5a) for definitions
  }
  function(m){
    sample(cdf_sample, m, prob=w, replace=T)
  }
}

# Sampling from the posterior distribution
sample_posterior <- function(F0, alpha, data){
  n <- length(data)
  F_hat <- DiscreteDistribution(data) # distr function for empirical CDF
  F_for_post <- n/(n+alpha)*F_hat+alpha/(n+alpha)*F0
  sample_prior(F_for_post, alpha+n, n)
}

# Now simulate for all different alphas
list_out_bayes <- list()
for(alpha in alpha_vec){
  iters <- 100
  m <- 1000
  F0 <- DiscreteDistribution(rnorm(m))

  y <- matrix(nrow=length(x_grid), ncol=iters)
  for(iter in 1:iters){
    F_post <- sample_posterior(F0, alpha, x_pts)
    y[,iter] <- ecdf(F_post(m))(x_grid)
  }

  mean_post_sim <- rowMeans(y) # Posterior Mean
  cred_int <- apply(y, 1, function(row) HPDinterval(as.mcmc(row), prob=signif_level))
  # obtains 95% credible interval
  list_out_bayes[[as.character(alpha)]] <- list('cred_int' = cred_int,
                            'mean_post_sim'=mean_post_sim)
}

# Plot the results for each of the alpha
for (alpha_val in alpha_vec){
  plot(x_ecdf, xlim=c(-3, 3),
       main = TeX(sprintf("95\% DKW and Bayesian Credible Band ($\alpha = %s$)", as.character(alpha_val)))) # ECDF
lines(x_grid, dkw.lb, col="green") #DKW
lines(x_grid, dkw.ub, col="green")

points(x_grid, list_out_bayes[[as.character(alpha_val)]]$'mean_post_sim',
type='l', col="red")
points(x_grid, list_out_bayes[[as.character(alpha_val)]]$'cred_int'[2,], type='l',
col='blue')
points(x_grid, list_out_bayes[[as.character(alpha_val)]]$'cred_int'[1,], type='l',
col='blue')
curve(pnorm, xlim=c(-3, 3), add=TRUE, col="red", lwd=2)
legend("topleft", lty="solid", legend=c("True", "DKW", "Kolmogorov", "Posterior Mean"),
    col=c("black", "green", "blue", "red"))

Figure 1: 95% DKW and Bayesian credible bands at different alpha levels: 0.1, 1, 10 from upper left to bottom center respectively

For $n = 1,000$ simulations, one should replicate the above code and consider whether the full empirical CDF is captured between the bands for both the Bayesian confidence and DKW bands.
Problem 6 [20 pts.]

In this question we consider a nonparametric Bayesian estimator and compare to the minimax estimator. For \( i = 1, \ldots, n \) and \( j = 1, 2, \ldots \) let

\[
X_{ij} = \theta_j + \epsilon_{ij}
\]

where all the \( \epsilon_{ij}'s \) are independent \( N(0, 1) \). The parameter is \( \theta = (\theta_1, \theta_2, \ldots) \). Assume that \( \sum_j \theta_j^2 < \infty \). Due to sufficiency, we can reduce the problem to the sample means. Thus let

\[
Y_j = \frac{1}{n} \sum_{i=1}^{n} X_{ij}.
\]

So the model is

\[
Y_j \sim N(\theta_j, 1/n)
\]

for \( j = 1, 2, 3, \ldots \). We will put a prior \( \pi \) on \( \theta \) as follows. We take each \( \theta_j \) to be independent and we take

\[
\theta_j \sim N(0, \tau_j^2)
\]

(a) (5 pts.) Find the posterior for \( \theta \). Find the posterior mean \( \hat{\theta} \).

(b) (7 pts.) Suppose that \( \sum_j \tau_j^2 < \infty \). Show that \( \hat{\theta} \) is consistent, that is,

\[
\lim_{n \to \infty} \parallel \hat{\theta} - \theta \parallel^2 = 0.
\]

(c) (8 pts.) Now suppose that \( \theta \) is in the Sobolev ball

\[
\Theta = \left\{ \theta = (\theta_1, \theta_2, \ldots) : \sum_j j^{2p} \theta_j^2 \leq C^2 \right\}
\]

where \( p > 1/2 \). The minimax (for squared error loss) for this problem is \( R_n \propto n^{-2p/(2p+1)} \). Let \( \tau_j^2 = (1/j)^{2p} \). Find \( r \) so that the posterior mean achieves the minimax rate.

Solution.

(a) By Theorem 6 (in the appendix) we have

\[
\hat{\theta}_j = \frac{nY_j \tau_j^2}{1 + n\tau_j^2}.
\]

(b) For any \( \epsilon > 0 \),

\[
P\left( \parallel \hat{\theta} - \theta \parallel^2 > \epsilon \right) \leq \frac{\mathbb{E}[\parallel \hat{\theta} - \theta \parallel^2]}{\epsilon} \leq \frac{1}{\epsilon} \sum_{j=1}^{\infty} \mathbb{E}[\parallel \theta_j - \theta_j \parallel^2]
\]

by Markov’s inequality

\[
= \frac{1}{\epsilon} \sum_{j=1}^{\infty} \left( \mathbb{E}[\hat{\theta}_j - \theta_j]^2 + \sum_{j=1}^{\infty} \text{Var}(\hat{\theta}_j) \right)
\]

\[
= \frac{1}{\epsilon} \left( \sum_{j=1}^{\infty} \frac{\theta_j^2}{(1 + n\tau_j^2)^2} + \sum_{j=1}^{\infty} \frac{\tau_j^2}{1 + n\tau_j^2} \right)
\]

by Theorem 6

\[
\to 0,
\]

as \( n \to \infty \), since \( \sum_j \tau_j^2 < \infty \) and \( \sum_j \theta_j^2 < \infty \).

(c) See Shen and Wasserman (2001).
Theorem 8 Let \( X_1, \ldots, X_n \sim N(\mu, \sigma^2) \) where \( \sigma^2 \) is known. Let \( \mu \sim N(a, b^2) \). Then,

\[
E[\mu|X_1, \ldots, X_n] = \frac{a\sigma^2 + n\bar{X}b^2}{\sigma^2 + nb^2}.
\]

Proof.

\[
f_{X^n}(x^n|\mu) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{\sigma^2}(x_i - \mu)^2\right\} = (2\pi\sigma^2)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2\right\}
\]

\[
= (2\pi\sigma^2)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2} [(n-1)s^2 + n(\mu - \bar{X})^2]\right\} = (2\pi\sigma^2)^{-n/2} \exp\left\{-\frac{(n-1)s^2}{2\sigma^2}\right\} \exp\left\{-\frac{n(\mu - \bar{X})^2}{2\sigma^2}\right\}
\]

\[
\propto \exp\left\{-\frac{n(\mu - \bar{X})^2}{2\sigma^2}\right\}.
\]

\[
\pi(\mu) = \frac{1}{\sqrt{2\pi b^2}} \exp\left\{-\frac{1}{2b^2}(\mu - a)^2\right\} \propto \exp\left\{-\frac{1}{2b^2}(\mu - a)^2\right\}.
\]

Hence,

\[
\pi(\mu|X^n) \propto f_{X^n}(x^n|\mu) \pi(\mu) \propto \exp\left\{-\frac{n(\mu - \bar{X})^2}{2\sigma^2} - \frac{1}{2b^2}(\mu - a)^2\right\}
\]

\[
= \exp\left\{-\frac{n\mu^2 + 2n\mu \bar{X} - n\bar{X}^2}{2\sigma^2} - \frac{-\mu^2 + 2\mu a - a^2}{2b^2}\right\}
\]

\[
= \exp\left\{\mu^2\left(-\frac{1}{2b^2} - \frac{n}{2\sigma^2}\right) - 2\mu\left(-\frac{a}{2b^2} - \frac{n\bar{X}}{2\sigma^2}\right) - \left(\frac{a^2}{2b^2} + \frac{n\bar{X}^2}{2\sigma^2}\right)\right\}
\]

For simplicity, let

\[
U = \left(-\frac{1}{2b^2} - \frac{n}{2\sigma^2}\right) \quad \text{and} \quad V = \left(-\frac{a}{2b^2} - \frac{n\bar{X}}{2\sigma^2}\right).
\]

Then

\[
\pi(\mu|X^n) \propto \exp\{U\mu^2 - 2V\mu\} = \exp\left\{U(\mu^2 - 2\mu \frac{V}{U^2} + \frac{V^2}{U^2}) - \frac{V^2}{U}\right\}
\]

\[
\propto \exp\{U(\mu - \frac{V}{U})^2\} = \exp\left\{-\frac{1}{2(1/\sqrt{2U})^2}(\mu - \frac{V}{U})^2\right\} \propto N\left(\frac{V}{U}, \frac{1}{2U}\right).
\]

Therefore the mean of the posterior is,

\[
\bar{\mu} = E[\mu|X^n] = \frac{V}{U} = \frac{-a\sigma^2 - nb^2 \bar{X}}{-\sigma^2 - nb^2} = \frac{a\sigma^2 + n\bar{X}b^2}{\sigma^2 + nb^2}
\]