1. (Total variation and coupling) Let $P$ and $Q$ be probability measures on some discrete set $\mathcal{X}$. A joint distribution $P_{XY}$ is called a coupling of $P$ and $Q$ if $P_X = P$ and $P_Y = Q$. Recall the total variation distance $TV(P, Q) = \frac{1}{2} \sum_x |P(x) - Q(x)|$ satisfies the variational representation

$$TV(P, Q) \triangleq \sup_E (P[E] - Q[E]),$$

where the supremum is over all subset $E$ of $\mathcal{X}$. In fact we also have the following dual representation of the total variation

$$TV(P, Q) = \min_{P_{XY}} \{ P[X \neq Y] : P_X = P, P_Y = Q \}. \quad (1)$$

Let us prove this fact in the discrete case following the steps below (the method work in general as well):

(a) Show that for any coupling $P_{XY}$, $TV(P, Q) \leq P[X \neq Y]$.
(b) Show that $TV(P, Q) = 1 - \sum_x P(x) \wedge Q(x)$, where $a \wedge b \triangleq \min\{a, b\}$.
(c) Let $t = TV(P, Q)$. Assume that $0 < t < 1$. Define three probability measures: $R = \frac{P \wedge Q}{1-t}$, $P' = \frac{P - P \wedge Q}{t}$ and $Q' = \frac{Q - P \wedge Q}{t}$. Construct a coupling $P_{XY}$ as follows:

1) Generate $B \sim \text{Bernoulli}(t)$.
2) If $B = 0$, draw $Z \sim R$ and set $X = Y = Z$.
3) If $B = 1$, draw $X \sim P'$ and $Y \sim Q'$ independently.

Verify that this $P_{XY}$ is a coupling of $P$ and $Q$.
(d) Conclude that (1) holds (also verify the case where $TV = 0$ or $1$).
(e) Consider the case of finite $\mathcal{X}$. Derive the dual problem of (1) (linear program) and show it coincides with the sup-representation $TV(P, Q) = \frac{1}{2} \sup_{\|f\|_{\infty} \leq 1} E_P[f(X)] - E_Q[f(X)]$.

2. (Coin flips) Consider the experiment where we observe $X_1, \ldots, X_n \overset{i.i.d.}{\sim} \text{Bern}(\theta)$ with $\theta \in \Theta = [0, 1]$ and estimate the bias $\theta$. Consider the quadratic loss function $\ell(\theta, \hat{\theta}) = (\theta - \hat{\theta})^2$ and denote the minimax risk by $R^*$.

(a) Use the empirical frequency $\hat{\theta}_{\text{emp}} = \bar{X}$ to estimate $\theta$. Compute and plot the risk $R_{\theta}(\hat{\theta})$ and show that

$$R^* \leq \frac{1}{4n}.$$
(b) Compute the Fisher information of $P_\theta = \text{Bern}(\theta)^\otimes n$ and $Q_\theta = \text{Binom}(n, \theta)$. Explain why they are equal.

(c) Invoke the Bayesian Cramér-Rao lower bound to show that
\[
R^* = \frac{1 + o(1)}{4n}.
\]

(d) Notice that the risk of $\hat{\theta}_\text{emp}$ is maximized at 1/2 (fair coin), which suggests that it might be possible to hedge against this situation by the following randomized estimator
\[
\hat{\theta}_\text{rand} = \begin{cases} 
\hat{\theta}_\text{emp}, & \text{with probability } \delta \\
\frac{1}{2} & \text{with probability } 1 - \delta
\end{cases}
\]
Find the worst-case risk of $\hat{\theta}_\text{rand}$ as a function of $\delta$. Choose the best $\delta$ and show that this leads to a better upper bound:
\[
R^* \leq \frac{1}{4(n + 1)}.
\]

(e) Randomization is always improvable when the loss is convex; so we should always average out the randomness by considering the estimator
\[
\hat{\theta}^* = \mathbb{E}[\hat{\theta}_\text{rand}|X] = \bar{X}\delta + \frac{1}{2}(1 - \delta).
\]
Optimizing over $\delta$ to minimize the worst-case risk, find the resulting estimator $\hat{\theta}^*$ and its risk, show that it is constant (independent of $\theta$), and conclude
\[
R^* \leq \frac{1}{4(1 + \sqrt{n})^2}.
\]

(f) Next we show $\hat{\theta}^*$ found in part (e) is exactly minimax and hence
\[
R^* = \frac{1}{4(1 + \sqrt{n})^2}.
\]
Consider the following prior Beta($a, b$) with density:
\[
\pi(\theta) = \frac{\Gamma(a + b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1 - \theta)^{b-1}, \theta \in [0, 1],
\]
where $\Gamma(a) \triangleq \int_0^\infty x^{a-1}e^{-x}dx.$ Show that if $a = b = \frac{n}{2}$, $\hat{\theta}^*$ coincides with the Bayes estimator for this prior, which is therefore least favorable. (Hint: work with the sufficient statistic $S = X_1 + \ldots + X_n$.)

(g) Show that the least favorable prior is not unique; in fact, there is a continuum of them. (Hint: consider the Bayes estimator $\mathbb{E}[\theta|X]$ and show that it only depends on the first $n + 1$ moments of $\pi$.)

(h) (Nonparametric extension) Consider the following nonparametric model $\mathcal{P} = \mathcal{M}([0, 1])$, the set of all probability distributions on $[0, 1]$. The data are $X_1, \ldots, X_n \overset{i.i.d.}{\sim} P \in \mathcal{P}$ and the goal is to estimate the mean of $P$ under the quadratic loss. Show that the minimax risk is
\[
R^* = \frac{1}{4(1 + \sqrt{n})^2}.
\]
3. (More properties of $f$-divergences)

(a) (Invariance) For any one-to-one transformation $g : \mathcal{X} \to \mathcal{Y}$, show that

$$D_f(P_{g(X)} || Q_{g(X)}) = D_f(P_X || Q_X).$$

Hence $f$-divergences are invariant under translation, dilation or rotation.

(b) Show that

$$D_f(P_0 \otimes Q || P_1 \otimes Q) = D_f(P_0 || P_1).$$

(c) Show that

$$\text{TV} \left( \prod_{i=1}^k P_i, \prod_{i=1}^k Q_i \right) \leq \sum_{i=1}^k \text{TV}(P_i, Q_i).$$

(Hint: use the coupling characterization of TV from Problem 1).

4. ($f$-divergences for Gaussian distributions)

(a) Let $\Sigma$ be a $d \times d$ positive semidefinite matrix and $\theta \in \mathbb{R}^d$. Show that $\text{TV}(\mathcal{N}(\theta, \Sigma), \mathcal{N}(0, \Sigma)) = 1 - 2\Phi(\|\Sigma^{-1/2}\theta\|_2^2/2)$, where $\Phi(a) \triangleq \int_a^\infty \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx$ denotes standard normal tail probability. (Hint: first prove it for one dimension; for general dimension, apply whitening and use Problem 3(a) and 3(b).)

(b) Let $X, Y$ be standard normal random variable such that under $P$ they are correlated with correlation coefficient $\rho \in (-1, 1)$ and under $Q$ they are independent. Compute $D(P_{XY} || Q_{XY})$, $H^2(P_{XY}, Q_{XY})$, and $\chi^2(P_{XY} || Q_{XY})$ as a function of $\rho$.

5. (Joint range)

(a) Show that joint range of the KL divergence $D$ versus the $\chi^2$ divergence is given by the following inequality:

$$0 \leq D(P || Q) \leq \log(1 + \chi^2(P || Q)).$$

(b) Explain the absence of the upper boundary (i.e. the impossibility to bound $\chi^2$ by $D$) via an explicit construction.

(c) Exhibit distributions that attain the lower boundary.