Problem 1

(a) Define a projector operator to be the following:

\[
\text{Proj}(\theta) = \begin{cases} 
\theta, & \text{if } \theta \in \Theta, \\
\text{arg min}_{\theta' \in \Theta} \| \theta - \theta' \|^2, & \text{if } \theta \notin \Theta. 
\end{cases}
\]

Since \(\Theta\) is a convex compact set, the minimizer \(\text{arg min}_{\theta' \in \Theta} \| \theta - \theta' \|^2\) is unique, so that \(\text{Proj}\) operator is well defined.

Given an estimator \(\hat{\theta} : X \to \mathbb{R}^d\) such that \(P_{\theta}(\hat{\theta}(X) \notin \Theta^c) > \delta\), we take \(\tilde{\theta} = \text{Proj}(\hat{\theta})\). Then for any \(\theta \in \Theta\), we have

\[
\| \tilde{\theta}(x) - \theta \|^2 \leq \| \hat{\theta}(x) - \theta \|^2 - \eta \{ \hat{\theta}(x) \notin \Theta^c \},
\]

where

\[
\eta = \min_{\theta \in \Theta, \theta' \in \partial \Theta^c} \| \theta' - \theta \|^2 - \| \text{Proj}(\theta') - \theta \|^2.
\]

Since \(\Theta\) is a convex compact set, and \(\partial \Theta^c\) is a compact set, we have \(\eta > 0\).

As a result, we have for any \(\theta \in \Theta\),

\[
E_{\theta}[\| \hat{\theta}(X) - \theta \|^2] \leq E_{\theta}[\| \hat{\theta}(X) - \theta \|^2] - \eta P_{\theta}(\hat{\theta}(X) \notin \Theta^c) \leq E_{\theta}[\| \hat{\theta}(X) - \theta \|^2] - \eta \delta,
\]

and

\[
\sup_{\theta \in \Theta} E_{\theta}[\| \hat{\theta}(X) - \theta \|^2] \leq \sup_{\theta \in \Theta} E_{\theta}[\| \hat{\theta}(X) - \theta \|^2] - \eta \delta.
\]

That means \(\hat{\theta}\) has strictly better worst risk than \(\tilde{\theta}\), so that \(\hat{\theta}\) is not minimax optimal on \(\Theta\).

(b) First we consider the case when \(M \neq 0\). Since \(\Theta\) is a compact convex set, we take \(R\) large enough so that \(\Theta^c \subseteq B(0, R)\) for some small \(\varepsilon > 0\). The estimator \(\hat{\theta}(y) = My + \theta_0 = MD\theta + \theta_0 + \sigma Mg\), where \(g \sim N(0, I_n)\). Note \(\sigma Mg\) is not identically \(0\) when \(M \neq 0\), and \(\Theta\) is a compact set, we have

\[
\inf_{\theta \in \Theta} P_{\theta}(\|MDy + \theta_0\|_2 \geq R) \equiv \delta > 0.
\]

By problem (a), we conclude that \(\hat{\theta}\) cannot be minimax optimal on \(\Theta\).

Remark 1. To show \(\hat{\theta} = \theta_0\) is not minimax optimal, we need to make the additional assumption that \(D \in \mathbb{R}^{n \times d}\) has full column rank, otherwise this conclusion doesn’t hold. In the following, we prove this conclusion under this additional assumption.
Then we consider the case when \( M = 0 \). That means, \( \hat{\theta} = \theta_0 \). If \( \theta_0 \not\in \Theta \), it is obvious \( \hat{\theta} \) is not minimax optimal on \( \Theta \). Hence we consider the case when \( \hat{\theta} = \theta_0 \in \Theta \).

We claim that the \( \hat{\theta} = \theta_0 \) cannot be the Bayes estimator for any prior except the prior \( \delta(\theta_0) \). Suppose this claim holds, the Bayes risk \( R_B(\hat{\theta}, \delta(\theta_0)) = 0 \). Since \( \Theta \) contains at least two points, it is easy to see that the minimax risk should be larger than 0, hence \( \delta(\theta_0) \) is not the least favorable prior. By minimax theorem, the minimax estimator should be the Bayes estimator for least favorable prior. Therefore, \( \hat{\theta} = \theta_0 \) cannot be the minimax estimator.

Now suffice to show the claim above. Suppose \( Q \) is a prior probability distribution on \( \Theta \) and \( Q(\Theta \setminus \{\theta_0\}) > 0 \), then the Bayes estimator under prior \( Q \) and square loss should be the posterior expectation \( \hat{\theta} = \text{Proj}(\hat{\theta}) \). If \( \hat{\theta} \) does not satisfy the condition above, then we consider the case when \( \hat{\theta} \neq \theta_0 \). Hence we only need to consider the case when \( \hat{\theta} = \theta_0 \). For this estimator, suppose \( \hat{\theta} = \theta_0 \) does not satisfy the condition above, then we consider the case when \( \hat{\theta} \neq \theta_0 \).

The integration in the numerator above can be decomposed into the integration in \( \mathbb{B}(\theta_0, \delta) \) and the integration outside \( \mathbb{B}(\theta_0, \delta) \),

\[
\begin{align*}
\int_{\mathbb{B}(\theta_0, \delta)} (\theta - \theta_0, \theta - \theta_0) \varphi_n(\theta - \theta_0, \sigma) Q(d\theta) &\geq ||\theta - \theta_0||_2\varphi_n(\theta - \theta_0) - \frac{1}{(2\pi)^n/2} \exp\left(-||\theta - \theta_0||_2^2/2\right) \eta \\
&\geq ||\theta - \theta_0||_2\varphi_n(\theta - \theta_0) - \frac{1}{(2\pi)^n/2} \exp\left(-||\theta - \theta_0||_2^2/2\right) \eta,
\end{align*}
\]

where \( \varphi_n(\theta) \) is the standard Gaussian density function on \( \mathbb{R}^n \).

For \( \hat{\theta} = \theta_0 \) does not satisfy the condition above, then we consider the case when \( \hat{\theta} \neq \theta_0 \). Hence we only need to consider the case when \( \hat{\theta} = \theta_0 \). For this estimator, the minimax estimator should be the Bayes estimator for least favorable prior. Therefore, \( \hat{\theta} = \theta_0 \) cannot be the minimax estimator.
Since $L$ is strictly decreasing for $a < \theta$ and strictly increasing for $a > \theta$, and $\Theta$ and $\partial \Theta^c$ are compact sets, we have $\eta > 0$.

As a result, we have for any $\theta \in \Theta$,

$$R(\hat{\theta}, \theta) = \mathbb{E}_{\theta}[L(\hat{\theta}(X), \theta)] \leq \mathbb{E}_{\theta}[L(\hat{\theta}(X), \theta)] - \eta \mathbb{P}_{\theta}(\hat{\theta}(X) \notin \Theta^c) \leq R(\hat{\theta}, \theta) - \eta \delta.$$

Since $R(\hat{\theta}, \theta)$ is continuous in $\theta$, $R(\hat{\theta}, \theta)$ can attain the maximum, and we have

$$\sup_{\theta \in \Theta} R(\hat{\theta}, \theta) \leq \sup_{\theta \in \Theta} R(\hat{\theta}, \theta) - \eta \delta.$$

That means $\hat{\theta}$ is not minimax optimal on $\Theta$. 

3
Problem 2

(a) Let $\theta_1 = 1/2 - 1/(2\sqrt{2})$, $\theta_2 = 1/2 + 1/(2\sqrt{2})$, and let $q = 1/2$. Under the square loss, the Bayes optimal estimator for $Q$ is given by the conditional expectation

$$
\hat{\theta}_B(x) = \mathbb{E} [\theta | X = x]
$$

$$
= \begin{cases} 
\frac{\theta_1^2 + \theta_2^2}{\theta_1^2 + \theta_2^2 + \theta_1(1-\theta_1) + \theta_2(1-\theta_2)} & \text{if } x = 1 \\
\frac{3}{4} & \text{if } x = 0 \\
\frac{1}{4} & \text{if } x = 0 \\
\frac{x}{2} + \frac{1}{4} & \text{if } x = 0
\end{cases}
$$

The above implies that $\hat{\theta}_B(Q) = \hat{\theta}_{MM}$.

(b) As suggested in the hint, there exists an integer $m$, such that choosing $q_i \geq 0$ for $i = 0, 1, \cdots, m$ such that (here $Q$ is the measure induced by a Beta($\sqrt{\eta}/2$, $\sqrt{\eta}/2$) random variable)

$$
\sum_{i=0}^{m} q_i \left( \frac{i}{m} \right)^k = \int \theta^k Q(d\theta) \quad \text{for all } k = 0, 1, \cdots, n + 1.
$$

Then the above implies that, for any polynomial $p$ of degree at most $n + 1$, we have

$$
\sum_{i=0}^{m} q_i p \left( \frac{i}{m} \right) = \int p(\theta) Q(d\theta).
$$

Consider the prior distribution:

$$
Q_1 = \sum_{i=0}^{n+1} q_i \delta \left( \frac{i}{m} \right)
$$

The Bayes optimal estimator is given by the conditional expectation

$$
\hat{\theta}_{Q_1}(X) = \mathbb{E}_{Q_2} [\theta | X]
$$

$$
= \frac{\sum_{i=0}^{n+1} q_i (i/m)^{X+1}(1-i/m)^{n-X}}{\sum_{i=0}^{n+1} q_i (i/m)^X(1-i/m)^{n-X}}.
$$

On the other hand, the Bayes estimator with respect to Beta($\sqrt{\eta}/2$, $\sqrt{\eta}/2$) is given by

$$
\hat{\theta}_{MM}(X) = \frac{\sqrt{\eta}}{1 + \sqrt{\eta}} \cdot \frac{X}{n} + \frac{1}{1 + \sqrt{\eta}} \cdot \frac{1}{2}
$$

$$
= \mathbb{E}_{\theta} [\theta | X]
$$

$$
= \int \frac{\theta^{X+1}(1-\theta)^{n-X} Q(d\theta)}{\theta^{X}(1-\theta)^{n-X} Q(d\theta)}.
$$

Let $p_1(t; X) = t^{X+1}(1-t)^{n-X}$, $p_2(t; X) = t^X(1-t)^{n-X}$, then it clear that both $p_1$ and $p_2$ as a function of $t$ are polynomial of degree at most $n + 1$. Hence by (3) we have

$$
\hat{\theta}_{Q_1}(X) = \sum_{i=0}^{m} p_1 \left( \frac{i}{m}, X \right) \sum_{i=0}^{m} p_2 \left( \frac{i}{m}, X \right) q_i = \frac{1}{p_2(\theta, X)Q(d\theta)} = \hat{\theta}_{MM}(X).
$$

Therefore,

$$
\hat{\theta}_{Q_1}(X) = \hat{\theta}_{MM}(X) = \frac{\sqrt{\eta}}{1 + \sqrt{\eta}} \cdot \frac{X}{n} + \frac{1}{1 + \sqrt{\eta}} \cdot \frac{1}{2}
$$
Problem 3

(a)

Since $L$ is upper bounded by $L_0$, $R(A, \theta)$ is also bounded from above by $L_0$ for all $A \in \mathcal{A}$ and $\theta \in \Theta$. Given $Q$, for any statistical procedure $A$, we have

$$
R(A, Q) = \int_{\mathbb{R}^d} R(A, \theta) Q(d\theta) = \int_{\Theta} R(A, \theta) Q(d\theta) + \int_{\Theta^c} R(A, \theta) Q(d\theta)
$$

\[ \leq \sup_{\theta \in \Theta} R(A, \theta) + L_0 Q(\Theta^c). \tag{9} \]

Hence

$$
R_B(Q) - L_0 Q(\Theta^c) \leq R(A, Q) - L_0 Q(\Theta^c) \leq \sup_{\theta \in \Theta} R(A, \theta). \tag{10} \]

Since the above is true for all $A$, taking the infimum over $A \in \mathcal{A}$ gives

$$
R_M(\Theta) \geq R_B(Q) - L_0 Q(\Theta^c). \tag{11} \]

(b)

Let $\hat{\theta}$ be any estimator, and let $\hat{\theta}$ be the projection of $\hat{\theta}$ onto $B^d(0, M\sqrt{k})$. That is

$$
\hat{\theta} = \min \left\{ \frac{M\sqrt{k}}{||\theta||_2}, 1 \right\} \hat{\theta}. \tag{12} \]

Then it is clear that $L(\hat{\theta}, \theta) \leq L(\hat{\theta}, \theta)$ with probability 1 for all $\theta \in \Theta(d, k, M) \subset B^d(0, M\sqrt{k})$. Since $\hat{\theta} \in B^d(0, M\sqrt{k})$, it is sufficient to only consider estimators taking values in $B^d(0, M\sqrt{k})$. In this case, since both $\hat{\theta}$ and $\theta$ are in a ball with radius $M\sqrt{k}$, there distance square is upper bounded by the diameter square of the ball. That is, for all $\theta \in \Theta$ and $\theta$ in the above form, we have

$$
L(\hat{\theta}, \theta) \leq 4M^2 k. \tag{13} \]

Therefore it is also sufficient to replace the loss bound by $\tilde{L}(\hat{\theta}, \theta) = \min\{||\theta - \theta||^2_2, 4M^2 k\}$.

(c)

Let $G = \Pi_d \times \Sigma_d$ be a group, where $\Pi_d$ is the permutation group on $\{1, \ldots, d\}$, and $\Sigma_d = \{+1, -1\}^d$ is the sign changing group. For any $g = [\pi, \sigma] \in G$ ($\pi$ is a permutation), where $\{\pi(1), \ldots, \pi(d)\} = \{1, \ldots, d\}$ as a set; $\sigma = [\sigma_1, \ldots, \sigma_d]^T \in \{+1, -1\}^d$, the action of $\varphi_g$ on $x = (x_1, \ldots, x_d)^T \in \mathbb{R}^d$ gives $\varphi_g(x) = (\sigma_1 x_{\pi(1)}, \ldots, \sigma_d x_{\pi(d)})^T$. We would like to show our statistical model is invariant under this group. First we have $L(a, \theta) = ||a - \theta||^2_2 = ||\varphi_g(a) - \varphi_g(\theta)||^2_2 = L(\varphi_g(a), \varphi_g(\theta))$. Next we have $P_{\varphi_g(\theta)}(X \in S) = P_{Z \sim N(0, \sigma^2 I_d)}(\varphi_g(\theta) + Z \in S) = P_{Z \sim N(0, \sigma^2 I_d)}(\varphi_g(\theta) + \varphi_g(Z) \in S) = P_{Z \sim N(0, \sigma^2 I_d)}(\varphi_g(\theta + Z) \in S) = P_{\varphi_g(\theta)}(X \in S) = (\varphi_g(z) \neq P_{\varphi_g(\theta)}(X \in S)$. Hence our model is invariant under this group. Since minimax theorem holds for this model, there exists a least favorable prior. According to invariant least favorable prior theorem, there exists a least favorable prior that is invariant under the group action. This invariant least favorable prior can only be written in the form $Q = \sum_{i=0}^{k} p_i Q_i$.

(d)

By part (b) we know that $R_M(d, k; M) = \tilde{R}_M(d, k; M)$, and we can replace the loss $L$ by $\tilde{L}$, which is bounded from above by $4M^2 k$. By part (a) we have

$$
R_M(d, k; M) = \tilde{R}_M(d, k; M) \geq \tilde{R}_B(Q_{M, \varepsilon}) - 4M^2 k Q_{M, \varepsilon}(\Theta^c). \tag{14} \]

Let $X \in \mathbb{R}^d$ be a random variable whose induced measure is $Q_{M, \varepsilon}$, then it is clear that $Q_{M, \varepsilon}(\Theta^c)$ is equal to $P(||X||_0 > k)$. Since the coordinates of $X$ are independent and $1(X_i \neq 0)$ has Bernoulli($\varepsilon$) distribution, $||X||_0$ has Binomial($d, \varepsilon$) distribution. Therefore, (14) becomes

$$
R_M(d, k; M) \geq \tilde{R}_B(Q_{M, \varepsilon}) - 4M^2 k P(\text{Binomial}(d, \varepsilon) > k). \tag{15} \]
(e)

Note \( \theta = (\theta_1, \ldots, \theta_d) \sim Q_{M,\epsilon} = q_{M,\epsilon}^d \), and \( X \sim \mathcal{N}(\theta, \sigma^2 I_d) \). We have \((X_i, \theta_i)\) for \(i \in [d]\) are mutually independent. Hence the Bayes estimator which is the posterior mean gives

\[
(\hat{\theta}_B(x))_j = \mathbb{E}[\theta_j | X = x] = \mathbb{E}[\theta_j | X_j = x_j].
\]

Hence

\[
R_B(Q_{M,\epsilon}) = \mathbb{E}_{Q_{M,\epsilon}}[\|\hat{\theta}_B - \theta\|_2^2]
= \sum_{j \in [d]} \mathbb{E}_{Q_{M,\epsilon}}[(\hat{\theta}_B)_j - \theta_j]^2
= \sum_{j \in [d]} \mathbb{E}_{Q_{M,\epsilon}}[(\mathbb{E}(\theta_j | X_j) - \theta_j)^2]
= \sum_{j \in [d]} \mathbb{E}_{q_{M,\epsilon}}[(\mathbb{E}(\theta_j | X_j) - \theta_j)^2]
= dR_B(q_{M,\epsilon}).
\] (16)

Since we have

\[
P(\text{Binom}(d, \epsilon) > k) \leq e^{-kn^2/4},
\] (17)

which implies that \(kP(\text{Binom}(d, \epsilon) > k) = o_n(k)\), using (15), (16) and (7) in the question gives

\[
R_M(d, k; M) \geq dR_B(q_{M,\epsilon}) - (M^2 + 1)o_n(k) - 4M^2o_n(k).\] (18)

Since a constant times \(o_n(k)\) is still \(o_n(k)\), the \(-4M^2o_n(k)\) above can be merged with the first \(M^2o_n(k)\), so it can be simplifies to

\[
R_M(d, k; M) \geq dR_B(q_{M,\epsilon}) - (M^2 + 1)o_n(k).
\] (19)