1. BAYESIAN ESTIMATION

1.1

In coin experiment, i.i.d samples $X = (X1, X2, ..., Xn)$ is from the Bernoulli distribution with unknown parameter $p \in (0, 1)$.

Likelihood function:

$$L(p) = f(X_1, X_2, ..., X_n|p) = \prod_{i=1}^{n} f(X_i|p$$
$$= \prod_{i=1}^{n} p^{X_i}(1-p)^{1-X_i}$$

let $Y = \sum_{i=1}^{n} X_i$, then we have:

$$L(p) = p^Y(1-p)^{n-Y}$$

Then, the MAP estimator is:

$$\hat{p}_{MAP} = \arg\max_p L(p)\pi(p)$$

$\pi(p)$ is the priori distribution of $\hat{p}$, which is a beta distribution:

$$\pi(p) = \frac{1}{B(\alpha, \beta)} p^{\alpha-1}(1-p)^{\beta-1}$$

Finally,

$$\hat{p}_{MAP} = \arg\max_p p^Y(1-p)^{n-Y} p^{\alpha-1}(1-p)^{\beta-1}$$
$$= \arg\max_p p^{\alpha+Y-1}(1-p)^{\beta+n-Y-1}$$

Solving the equation above is equivalent to solve this equation:

$$\frac{\mathrm{d}}{\mathrm{d}p} \ln \hat{p}_{MAP} = 0$$

the result is:

$$\hat{p}_{MAP} = \frac{\alpha + Y - 1}{\alpha + \beta + n - 2}$$

1.2

The bias is:

$$B = E(\hat{p}_{MAP}) - p$$

$$E(\hat{p}_{MAP}) = E(\frac{\alpha + Y - 1}{\alpha + \beta + n - 2}) = \frac{\alpha + E(Y) - 1}{\alpha + \beta + n - 2}$$
$$= \frac{\alpha + np - 1}{\alpha + \beta + n - 2}$$

then,

$$B = \frac{\alpha + np - 1}{\alpha + \beta + n - 2} - p$$
$$= \frac{\alpha(1-p) - \beta p - 2p - 1}{\alpha + \beta + n - 2}$$

The variance is:

$$V = Var(\hat{p}_{MAP}) = \frac{Var(Y)}{(\alpha + \beta + n - 2)^2}$$
$$= \frac{np(1-p)}{(\alpha + \beta + n - 2)^2}$$

2. MINIMAX OPTIMALITY

2.1

$$MSE = E(\hat{\theta} - \theta)^2 = E(\hat{\theta} - E(\hat{\theta}) + E(\hat{\theta}) - \theta)^2$$
$$= E(\hat{\theta} - E(\hat{\theta}))^2 + 2E(\hat{\theta} - E(\hat{\theta}))(E(\hat{\theta}) - \theta) + (E(\hat{\theta}) - \theta)^2$$
$$= E(\hat{\theta} - E(\hat{\theta}))^2 + (E(\hat{\theta}) - \theta)^2$$
$$= V + B^2$$

2.2

the posterior risk is the inner integral, which is:

$$r(\hat{p}|X_1, X_2, ..., X_n) = \int_p \pi(p|X_1, X_2, ..., X_n)(\hat{p} - p)^2 \mathrm{d}p$$

take the derivative of $r(\hat{p}|X_1, X_2, ..., X_n)$ with respect to $\hat{p}$ and set it to be zero, which yields:

$$2\int_p \pi(p|X_1, X_2, ..., X_n)(\hat{p} - p)\mathrm{d}p = 0$$

solving this equation we have,

$$\hat{p}\int_p \pi(p|X_1, X_2, ..., X_n)\mathrm{d}p = \int_p \pi(p|X_1, X_2, ..., X_n)p\mathrm{d}p$$

the Bayes optimal estimator is:

$$\hat{p} = \int_p \pi(p|X_1, X_2, ..., X_n)p\mathrm{d}p = E(p|X_1, X_2, ..., X_n)$$

from the first question we know, the posterior distribution for $\hat{p}$ is:

$$\pi(\hat{p}|X_1, X_2, ..., X_n) = \hat{p}^{\alpha+Y-1}(1-\hat{p})^{\beta+n-Y-1}$$

this is actually a beta distribution with parameter $\alpha + Y$ and $\beta + n - Y$

thus,

$$\hat{p} = \frac{\alpha + Y}{\alpha + \beta + n}$$

2.3

generalize $MSE(\hat{p}, p)$ to be $R(\hat{p}, p)$, as the question indicates,

$$\sup_p R(\hat{p}, p) \leq B_\pi(\hat{p})$$

also, according to the definition, Bayes risk is an average of $R(\hat{p}, p)$

$$B_\pi(\hat{p}) \leq \sup_p R(\hat{p}, p)$$

combining the last two equations, we have,

$$B_\pi(\hat{p}) = \sup_p R(\hat{p}, p)$$

as a result, $\hat{p}$ is also a minimax optimal estimator.

2.4

because $B_\pi(\hat{p}) = \int_p R(\hat{p}, p)\pi(p)\mathrm{d}p$, if $B_\pi(\hat{p})$ is a constant, $R(\hat{p}, p) \leq B_\pi(\hat{p})$, apply it to the theorem in 2.3, next we only need to find the conditions for a constant MSE.

from 2.2, we have the Bayes optimal estimator, compute its $R(\hat{p}, p)$, namely $MSE(\hat{p}, p)$ by using the equation in 2.1,

$$\begin{aligned} MSE(\hat{p}, p) &= V + B^2 \\ &= \frac{np(1-p)}{(\alpha+\beta+n)^2} + (\frac{(1-p)\alpha - p\beta}{\alpha+\beta+n})^2 \\ &= \frac{np - np^2 + (\alpha+\beta)^2 p^2 - 2\alpha(\alpha+\beta)p + \alpha^2}{(\alpha+\beta+n)^2} \end{aligned}$$

in order to make it a constant, the coefficients of formulas containing p to be zero.

$$n = 2\alpha(\alpha+\beta)$$
$$n = (\alpha+\beta)^2$$

soving these two equations gives,

$$\alpha = \beta = \frac{\sqrt{n}}{2}$$

thus, the new estimator is:

$$\hat{p} = \frac{\sqrt{n}/2 + Y}{\sqrt{n} + n}$$

3. MINIMAX WITH 0/1 ERROR

for zero-one loss, the posterior risk of the estimator $\hat{p}$ is:

$$\begin{aligned} r(\hat{p}|X_1, X_2, ..., X_n) &= \sum_{p}^{c} \pi(p|X_1, X_2, ..., X_n) L(\hat{p}, p) \\ &= \sum_{\hat{p} \neq p}^{c} \pi(p|X_1, X_2, ..., X_n) \\ &= 1 - \pi(p'|X_1, X_2, ..., X_n) \end{aligned}$$

p' is the correct value for p

in order to minimize this risk function we need to choose a correct $\hat{p} = p$ that is the maximum of posterior distribution of p, namely, $\hat{p}_{MAP}$

given $p \in \{1/4, 3/4\}$, the prior distribution is a bernoulli distribution:

$$\pi(p) = k^{2p-\frac{1}{2}}(1-k)^{-2p+\frac{3}{2}}, p \in \{1/4, 3/4\}$$

then, the $\hat{p}_{MAP}$ is:

$$\hat{p}_{MAP} = \arg\max_{p} p^Y (1-p)^{n-Y} k^{2p-\frac{1}{2}} (1-k)^{-2p+\frac{3}{2}}$$

by taking derivative and set it to be zero

$$\frac{Y}{\hat{p}} + \frac{n-Y}{\hat{p}-1} + 2\ln k - 2\ln(1-k) = 0$$

as usual, next need to prove $B_\pi(\hat{p}) = c$, c is a constant. but this time it is equivalent to prove posterior risk is a constant which is $1 - \hat{p}_{MAP}$. if $\hat{p}_{MAP}$ is a constant, then we have $k = \frac{1}{2}$, then:

$$\frac{Y}{\hat{p}} + \frac{n-Y}{\hat{p}-1} = 0$$
$$\hat{p}_{MAP} = \frac{Y}{n}$$

because $\hat{p}_{MAP}$ is a constant, as proof above, it is also a minimax estimator

4. MLE FOR HIGH DIMENSIONAL DATA

4.1

the log likelihood function for normal distribution is:

$$l(\mu, \sigma^2|X_1, X_2, ..., X_n) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i - \mu)^2$$

the MLE estimator for μ is:

$$\hat{\mu}_{MLE} = \arg\max_{\mu} l(\mu, \sigma^2|X_1, X_2, ..., X_n)$$

take the partial derivative of the log likelihood function with respect to μ and set it to 0:

$$\frac{\partial}{\partial\mu} l(\mu, \sigma^2|X_1, X_2, ..., X_n) = 0$$
$$\frac{\partial}{\partial\mu}(-\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i - \mu)^2) = 0$$
$$\frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \mu) = 0$$

then we have,

$$\hat{\mu}_{MLE} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

the mean square error is:

$$\begin{aligned} MSE(\hat{\mu}, \mu) &= V + B^2 \\ &= (\frac{1}{n^2}\sum_{i=1}^{n} Var(X_i) + \sum_{i\neq j} Cov(X_i, X_j)) + (\frac{1}{n}\sum_{i=1}^{n} E(X_i) - \mu)^2 \\ &= \frac{1}{n}\sigma^2 \end{aligned}$$

4.2

compute $\hat{\theta}$ based on $\hat{\mu}$

$$\begin{aligned} \hat{\theta} &= \sum_{i=1}^{d} \hat{\mu}_i^2 \\ &= d(\bar{X})^2 \end{aligned}$$

the MSE of θ is:

$$\begin{aligned} MSE(\hat{\theta}, \theta) &= V + B^2 \\ &= d^2(Var(\bar{X}^2) + (E(\bar{X}^2) - \mu^2)^2) \\ &= d^2(E(\bar{X}^2) - E(\bar{X})^2) + (E(\bar{X}^2) - \mu^2)^2) \end{aligned}$$

$$\begin{aligned} E(\bar{X}^2) &= \frac{1}{n^2} E((\sum_{i=1}^{n} X_i)^2) \\ &= \frac{1}{n^2}(\sum_{i=1}^{n} E(X_i^2) + \sum_{i\neq j} E(X_i X_j)) \\ &= \frac{1}{n^2}(\sum_{i=1}^{n} (\mu^2 + \sigma^2) + \sum_{i\neq j} \mu^2) \\ &= \mu^2 + \frac{1}{n}\sigma^2 \end{aligned}$$

then, the result is:

$$MSE(\hat{\theta},\theta) = d^2(\mu^2 + \frac{1}{n}\sigma^2 - \mu^2) + (\mu^2 + \frac{1}{n}\sigma^2 - \mu^2)^2)$$
$$= \frac{d^2}{n^2}\sigma^4 + \frac{d^2}{n}\sigma^2$$

the error will go to 0 as $n \to \infty$

4.3

if d also increases, the error will not go to 0 as $n \to \infty$
set $\theta_{new} = \sqrt{\|\mu\|^2}$

$$\hat{\theta_{new}} = \sqrt{\sum_{i=1}^{d}\hat{\mu}_i^2}$$
$$= \sqrt{d}\bar{X}$$

$$MSE(\hat{\theta}_{new},\theta) = V + B^2$$
$$= d(Var(\bar{X}) + (E(\bar{X}) - \mu)^2)$$
$$= \frac{d}{n}\sigma^2$$

the new error will go to 0

5. BAYESIAN ESTIMATION FOR MULTINOMIAL DISTRIBUTION

5.1

the likelihood function is:

$$L(\vec{p}) = f(X_1, X_2, ..., X_n|\vec{p}) = \prod_{i=1}^{n}\prod_{j=1}^{k} p_j^{I(X_i=j)}$$
$$= \prod_{j=1}^{k} p_j^{N_j}$$

where $N_j = \sum_{i=1}^{n} I(X_i = j)$ and the log function is:

$$l(\vec{p}) = \sum_{j=1}^{k} N_j \ln p_j$$

finding the maximum of $l(\vec{p})$ is constrained by $\sum_{j=1}^{k} p_j = 1$, so we need to introduce Lagrange multiplier λ, then the Lagrangian log function is:

$$\tilde{l}(\vec{p}) = l(\vec{p}) + \lambda(1 - \sum_{j=1}^{k} p_j)$$

solving $\frac{\partial}{\partial p_j}\tilde{l}(\vec{p}) = 0$ yields:

$$N_j = \lambda p_j$$

sum all N_j yields:

$$N = \lambda\sum_{j=1}^{k} p_j = \lambda$$
$$\hat{p_j}_{MLE} = \frac{N_j}{N}$$

thus,

$$\hat{\vec{p}}_{MLE} = (\frac{N_1}{N}, \frac{N_2}{N}, ..., \frac{N_k}{N})$$

5.2

since there is no loss function specified, I consider Bayes estimator to be maximum a posteriori estimator just as in question 1, the conjugate prior is Dirichlet distribution:

$$\pi(\vec{p}) = \frac{1}{B(\alpha)}\prod_{j=1}^{k} p_j^{\alpha_j - 1}$$

then,

$$\hat{\vec{p}}_{MAP} = \arg\max_{\vec{p}} L(\vec{p})\pi(\vec{p})$$
$$= \arg\max_{\vec{p}} \prod_{j=1}^{k} p_j^{N_j + \alpha_j - 1}$$

repeat the same process of 5.1 the Lagrangian log function is:

$$\tilde{l}(\vec{p}) = \sum_{j=1}^{k}(N_j + \alpha_j - 1)\ln p_j + \lambda(1 - \sum_{j=1}^{k} p_j)$$

solving $\frac{\partial}{\partial p_j}\tilde{l}(\vec{p}) = 0$ yields:

$$N_j + \alpha_j - 1 = \lambda p_j$$

because all α_j are the same and set their value to be α, summing last equation,

$$N + k\alpha - k = \lambda$$
$$\hat{p_j}_{MAP} = \frac{N_j + \alpha - 1}{N + k\alpha - k}$$
$$\hat{\vec{p}}_{MAP} = (\frac{N_1 + \alpha - 1}{N + k\alpha - k}, \frac{N_2 + \alpha - 1}{N + k\alpha - k}, ..., \frac{N_k + \alpha - 1}{N + k\alpha - k})$$

6. NAIVE BAYES CLASSIFIER

6.1

see code

6.2

see code

6.3

when $\alpha = 1.02$, the accuracy is the highest.

6.4

Bayesian estimate does better because it considers our prior knowledge