6262 HOMEWORK 2 + SOLUTIONS

Problem 1. (1) *A sample $X_1, X_2, \ldots, X_n$ is drawn from a distribution $f(x;\theta)$ with parameter $\theta \in \mathbb{R}$. If the loss function is the $0-1$ loss, i.e.*

$$L(\hat{\theta}, \theta) = \begin{cases} 1, & \hat{\theta} \neq \theta \\ 0, & \hat{\theta} = \theta \end{cases},$$

show that any estimator is a minimax estimator. Does this make sense? Can you interpret it?

(2) *If the sample is drawn from $Bernoulli(p)$ with $p \in (0,1)$ and the loss function is*

$$L(\hat{p}, p) = \frac{(\hat{p}-p)^2}{p^3}$$

show that any estimator is a minimax estimator.

Proof. (1) By definition,

$$R(\hat{\theta}, \theta) = \mathbb{E}[L(\hat{\theta}, \theta)] = \int L(\hat{\theta}(x), \theta) f(x;\theta) dx = \mathbb{P}(\hat{\theta} \neq \theta)$$

so we obtain in the first place that

$$R(\hat{\theta}, \theta) \leq 1 \text{ for any choice of } \theta.$$

This means that

$$\bar{R}(\hat{\theta}) \leq 1$$

for any $\hat{\theta}$. In fact this is got to be equal to 1 for any $\hat{\theta}$, because if not, then we can find an $\epsilon > 0$, such that

$$\mathbb{P}(\hat{\theta} \neq \theta) = R(\hat{\theta}, \theta) \leq 1 - \epsilon$$

for any θ or alternatively, this means that

$$\mathbb{P}(\hat{\theta} = \theta) \geq \epsilon$$

for any θ. This is not possible because, if we choose $\theta = 1, 2, 3$, we have the sets $\{\hat{\theta} = \theta\}$ are going to be disjoint and this is not possible because we would have

$$1 \geq \mathbb{P}(\hat{\theta} = 1) + \mathbb{P}(\hat{\theta} = 2) + \cdots \geq \infty \times \epsilon = \infty$$

an obvious contradiction. Therefore, for any estimator $\hat{\theta}$, we have that $\bar{R}(\theta) = 1$, and thus any estimator is a minimax.

This makes sense because if θ takes infinitely many values, with the loss function $0-1$ it is hard to get a handle on θ. The reason, for this is that we look for the worst case scenario, which is the largest possible risk. This turns out to be 1 in this case for any estimator. Things become easier if we look at the complement, namely the probability that we estimate exactly θ with a given estimator $\hat{\theta}$. The smallest such value is 0 when we run the parameter over all possible values, but this is the math we delivered above.

(2) Take any estimator $\hat{p}$. Then the risk is

$$R(\hat{p}, p) = \mathbb{E}[L(\hat{p}, p)] = \frac{\mathbb{E}[(\hat{p}-p)^2]}{p^3} = \frac{\mathbb{E}[(\hat{p}-p)^2]}{p^3} \geq \frac{(E[\hat{p}]-p)^2}{p^3}$$

where we use the Cauchy's inequality to pass to the last term. Now, we need to maximize this over all $1 > p > 0$. If $\mathbb{E}[\hat{p}] = 0$, then

$$R(\hat{p}, p) \geq \frac{1}{p}$$

while for $\mathbb{E}[\hat{p}] > 0$, we obtain that

$$R(\hat{p}, p) \xrightarrow[p \to 0]{} \infty.$$

Thus we obtain that

$$\bar{R}(\hat{p}) = \infty$$

for any choice of the estimator $\hat{p}$. This means that any estimator is minimax.

□

Problem 2. *Describe and interpret in your own words what the setup of a minimax and Bayesian estimator means. Why and how is this relevant from the statistical point of view?*

Proof. We have a sample $X_1, X_2, \ldots, X_n$ from a sample $f(x; \theta)$. The purpose is to find a good estimator of θ. The loss function $L(\hat{\theta}, \theta)$ measures how far we are from the true parameter. The next element is the risk function which is defined as

$$R(\hat{\theta}, \theta) = \int L(\hat{\theta}(\vec{x}), \theta) f(\vec{x}; theta) dx$$

this is probably better written in the form

$$R(\hat{\theta}, \theta) = \mathbb{E}[L(\hat{\theta}(X_1, X_2, \ldots, X_n), \theta)]$$

which is the AVERAGE loss associated to each parameter. The next step is to take the worst possible such risk over all parameters θ. Thus

$$\bar{R}(\hat{\theta}) = \sup_{\theta} R(\hat{\theta}, \theta)$$

The notion of minimax is the estimator which minimizes this overall worst average error. Thus the key is that this is to some extent the best one can do for all possible values of θ.

The Bayesian estimator changes slightly, so that instead of taking the maximum risk, we take the average risk with respect to some density f assigned to θ. Thus we need to look at

$$r(f; \hat{\theta}) = \int R(\hat{\theta}, \theta) f(\theta) d\theta.$$

The Bayes estimator is the one which minimizes this expression. This is easier to estimate than the minimax.

□

Problem 3. *Assume that $X_1, X_2, \ldots, X_n$ is a sample from a distribution $f(x; \theta)$ and the loss function L. Assume that for a given prior f we know that the Bayesian estimator $\hat{\theta}^f$ satisfies*

$$R(\hat{\theta}^f, \theta) \leq r(f; \hat{\theta}^f) \text{ for any } \theta.$$

Prove that $\hat{\theta}^f$ is a minimax.

Proof. This one of the key result. The proof is rather simple. Assume that $\hat{\theta}^f$ is not minimax. Then there must exist another estimator $\tilde{\theta}$ such that

$$\sup_{\theta} R(\tilde{\theta}, \theta) < \sup_{\theta} R(\hat{\theta}^f, \theta).$$

Since we have always

$$r(f;\tilde{\theta}) \leq \sup_{\theta} R(\tilde{\theta},\theta)$$

and by the hypothesis we also get

$$\sup_{\theta} R(\hat{\theta}^f,\theta) \leq r(f,\hat{\theta}^f)$$

we get in the end that

$$r(f;\tilde{\theta}) \leq \sup_{\theta} R(\tilde{\theta},\theta) < \sup_{\theta} R(\hat{\theta}^f,\theta) \leq r(f,\hat{\theta}^f)$$

Thus

$$r(f;\tilde{\theta}) < r(f,\hat{\theta}^f)$$

which is a contradiction with the definition of $\hat{\theta}^f$ (this is the minimum of $r(f,\hat{\theta})$).

□

Problem 4. *Assume we have a parameter θ taking only two values, $\{1,2\}$ and for each of these we have distributions given by*

t	-1	1
$p(t;1)$	$1/4$	$3/4$
$p(t;2)$	$3/4$	$1/4$

(1) *We draw a single sample from this distribution. Assume the loss function L is given by the $0-1$ loss.*
   (a) *Find a minimax estimator.*
   (b) *Assume we put a prior distribution on the space of parameters, namely we consider $f(1)=\lambda$ and $f(2)=1-\lambda$. For this prior distribution find the Bayesian estimator of θ.*
   (c) *Using the Bayesian estimator above, can you find a minimax? Is this the same as the one found above with direct methods?*
   (d) *If we observe the sample $t=1$, what is the estimation of θ?*

(2) *Now, assume the square loss function $L(\hat{\theta},\theta)=(\hat{\theta}-\theta)^2$,*
   (a) *Find the risk function associated to any estimator $\hat{\theta}(X)$.*
   (b) *Find a minimax estimator and also a Bayesian estimator for the prior $f(1)=\lambda$, $f(2)=1-\lambda$.*
   (c) *If we observe the sample $t=1$, what is the estimation of θ?*

(3) *Do the same for the distributions*

t	-1	1
$p(t;1)$	$1/2$	$1/2$
$p(t;2)$	$3/4$	$1/4$

Proof. (1) (a) An estimator is completely determined by $\hat{\theta}(-1)=x$ and $\hat{\theta}=y$. The risk associated to this is given by $R(\hat{\theta},1)=L(x,1)/4+3L(y,1)/4$ while $R(\hat{\theta},2)=3L(x,2)/4+L(y,2)/4$ which gives $R(x,y)=\max\{R(\hat{\theta},1),R(\hat{\theta},2)\}$

	$x,y=1$	$x=1,y=2$	$x=2,y=1$	$x,y=2$	$x=1,y\neq 1,2$	$x=2,y\neq 1,2$	$y=1,x\neq 1,2$	$y=2,x\neq 1,2$	$x,y\neq 1,2$
$\theta=1$	0	3/4	1/4	1	3/4	1	1/4	1	1
$\theta=2$	1	3/4	1/4	0	1	1/4	1	3/4	1
$\bar{R}(\hat{\theta})$	1	3/4	1/4	1	1	1	1	1	1

Consequently, the values of x,y which minimizes $\bar{R}$ is $x=2,y=1$ which gives $\hat{\theta}(-1)=2$ and $\hat{\theta}(1)=1$.

(b) If we take the prior then, denoting as above $\hat{\theta}(-1) = x$ and $\hat{\theta}(1) = y$, we obtain

$$r(f, \hat{\theta}) = \lambda(L(x,1)/4 + 3L(y,1)/4) + (1-\lambda)(3L(x,2)/4 + L(y,2)/4)$$

To deal with we use a similar table

	$x, y = 1$	$x = 1, y = 2$	$x = 2, y = 1$	$x, y = 2$	$x = 1, y \neq 1, 2$	$x = 2, y \neq 1, 2$	$y = 1, x \neq 1, 2$	$y = 2, x \neq 1, 2$	$x, y \neq 1, 2$
$\theta = 1$	0	$3/4$	$1/4$	1	$3/4$	1	$1/4$	1	1
$\theta = 2$	1	$3/4$	$1/4$	0	1	$1/4$	1	$3/4$	1
$r(f, \hat{\theta})$	$1 - \lambda$	$3/4$	$1/4$	λ	$1 - \lambda/4$	$1/4 + 3\lambda/4$	$1 - 3\lambda/4$	$3/4 + \lambda/4$	1

(c) At any rate, the smallest value is attained for $x = 2, y = 1$ for any $\lambda \in (1/4, 3/4)$ which provides the same estimator as the minimax.
However for the case of $\lambda = 1/4$ we have another estimator which is given by $x = y = 2$ and in the case of $\lambda < 1/4$ this is the *only* Bayes estimator. As one can see this is in agreement with the prior, namely, $\lambda < 1/4$ strongly favors $\theta = 2$ and this means in this case that the Bayes estimator is geared toward 2.
On the other hand, in the case of $\lambda = 3/4$, we have a different estimator, as $x = y = 1$ and for $\lambda > 3/4$, this is the only estimator. As one can see this obviously favors $\theta = 1$ over $\theta = 2$.
For $\lambda = 0$, we obtain another estimator which is $x = 2$ and $y \neq 1, 2$. Notice that for any $y \neq 1, 2$ we obtain another estimator. This is Bayesian for the prior concentrated at $\theta = 2$ and gives $\hat{\theta}(-1) = 2$ and $\hat{\theta}(1) = y$. In particular we can see that there are infinitely many such estimators (for each y we get a different one).
For $\lambda = 1$, we also get another estimator which is $y = 1, x \neq 1, 2$. This is Bayesian for the prior concentrated at $\theta = 1$ and gives that $\hat{\theta}(-1) = x \neq 1, 2$, while $\hat{\theta}(1) = 1$.
At any rate, the only Bayesian estimator which achieves the equal risks is given by $\hat{\theta}(-1) = 2$, while $\hat{\theta}(1) = 1$, the same as the minimax rule.

(d) If we observe $t = -1$, then we estimate $\hat{\theta} = 2$, while if $t = 1$, then $\hat{\theta} = 1$ for both the minimax and also for any Bayesian estimator with $\lambda \neq 0$. For $\lambda = 0$, we correctly predict $\hat{\theta}(-1) = 2$ but not $\hat{\theta}(1) = y$. This means that if we start with the belief that $\theta = 2$ is the correct value, then the Bayesian estimator predicts the value of 1 arbitrarly with with any value of $y \neq 1, 2$.

(2) (a) The risk function associated to this is

$$R(\hat{\theta}, \theta) = \mathbb{E}[(\hat{\theta}(X) - \theta)^2]$$

To deal with this we notice that the estimator $\hat{\theta}$ is entirely determined by $x = \hat{\theta}(-1)$ and $y = \hat{\theta}$. Then the risk above becomes

$$R(\hat{\theta}, \theta) = \begin{cases} (1/4)(x-1)^2 + (3/4)(y-1)^2, & \theta = 1 \\ (3/4)(x-2)^2 + (1/4)(y-2)^2, & \theta = 2 \end{cases}$$

(b) Based on this we compute $\bar{R}$ as

$$\bar{R}(\hat{\theta}) = \bar{R}(x, y) = \max\{(1/4)(x-1)^2 + (3/4)(y-1)^2, (3/4)(x-2)^2 + (1/4)(y-2)^2\}$$

We have to find the region where the first one is larger than the second one. Taking the difference, we get

$$\begin{aligned}(1/4)(x-1)^2 + (3/4)(y-1)^2 - ((3/4)(x-2)^2 + (1/4)(y-2)^2) &= \frac{1}{2}\left(-x^2 + 5x + y^2 - y - 6\right) \\ &= (y - 1/2)^2 - (x - 5/2)^2\end{aligned}$$

If this expression is positive, which amounts to $|y-1/2| \geq |x-5/2|$ the maximum is given by $(1/4)(x-1)^2+(3/4)(y-1)^2$, while if this is negative the maximum is given by $(3/4)(x-1)^2+(1/4)(y-1)^2$.
To find the minimax, we need to minimize $\bar{R}$ this over x, y. This is usually difficult. Imagine that this becomes a lot more difficult if $p(x;\theta)$ is defined for x having 4 values. Add to this the case where the sample size is much larger. In that case we need to use some numerical software to get the calculation done.
At any rate, take the two cases separately, namely the minimization of

$$(1/4)(x-1)^2+(3/4)(y-1)^2 \text{ and } (3/4)(x-2)^2+(1/4)(y-2)^2$$

have the minima at $(x,y)=(1,1)$ and the other has the minima at $(x,y)=(2,2)$. Unfortunately neither $(x,y)=(1,1)$ nor $(x,y)=(2,2)$ are in the allowed regions. Therefore what this means is that the minimum of $\bar{R}(x,y)$ is not attained at the minimum of each of the components.
However it seems that we can actually minimize over the intersection curve. This means we need to minimize either of the components, say for example, $(1/4)(x-2)^2/4+(3/4)(y-2)^2$ over the range of $|y-1/2|=|x-5/2|$ or $y=x-2$ or $y=3-x$. In each case we have

$$3(x-2)^2/4+(1/4)(y-2)^2 = 3(x-2)^2/4+(x-4)^2/4 \text{ or } 3(x-2)^2/4+(1/4)(y-2)^2 = 3(x-2)^2/4+(1-x)^2/4.$$

The first function yields the minimum $3/4$ attained at $x=5/2$, while the second attains the minimum of $3/16$ are $x=7/4$. Thus the minimum is given by $x=7/4, y=5/4$. The other case of minimizing the $(x-1)^2/4+3(y-1)^2/4$ yields the same minimum at $x=7/4$ and $y=5/4$.
Thus the minimax estimator is $\hat{\theta}(-1)=7/4$ and $\hat{\theta}(1)=5/4$. The moral is that if we observe the value of $t=-1$, then an estimate of θ is $7/4$ (closer to 2), while if we observe the value of 1, an estimate of θ is $5/4$ (closer to 1).
For the bias estimator, we have

$$\begin{aligned} r(f,\hat{\theta}) &= \lambda(x-1)^2/4+3\lambda(y-1)^2/4+3(1-\lambda)(x-2)^2/4+(1-\lambda)(y-2)^2 \\ &= \frac{1}{4}\left((3-2\lambda)x^2-(10\lambda-12)x+(2\lambda+1)y^2+(-2\lambda-4)y+16-12\lambda\right). \end{aligned}$$

For any $\lambda \in [0,1]$ we can minimize this much simpler. In fact we can minimize separately in x and y to yield $x=\frac{6-5\lambda}{3-2\lambda}$ and $y=\frac{\lambda+2}{2\lambda+1}$. The calculations here for the risk function become pretty complicated and takes the form

$$R(x,y,\theta)=\begin{cases} \frac{3(\lambda-1)^2\left(4\lambda^2+3\right)}{(3-2\lambda)^2(2\lambda+1)^2}, & \theta=1 \\ \frac{3\lambda^2\left(4\lambda^2-8\lambda+7\right)}{(3-2\lambda)^2(2\lambda+1)^2}, & \theta=2. \end{cases}$$

A calculation shows that the λ for which this becomes constant in θ (both branches of $R(x,y,\theta)$ are equal) is $\lambda=1/2$. This yields $x=7/4$ and $y=7/4$

(3) For the given function we can do the same as above, the only issue is that the computations become a little worse. Thus we start with

$$R(\hat{\theta},\theta)=\begin{cases} (1/2)(x-1)^2+(1/2)(y-1)^2, & \theta=1 \\ (3/4)(x-2)^2+(1/4)(y-2)^2, & \theta=2. \end{cases}$$

To find the minimax we need to minimize the function

$$\bar{R}(x,y)=\max\{(1/2)(x-1)^2+(1/2)(y-1)^2, (3/4)(x-2)^2+(1/4)(y-2)^2\}$$

Unfortunately, as opposed to the previous case this turns out to be a little more complicated. We first have to figure out the region where $(1/2)(x-1)^2+(1/2)(y-1)^2 \geq (3/4)(x-2)^2+(1/4)(y-2)^2$ versus the region where $(1/2)(x-1)^2+(1/2)(y-1)^2 \leq (3/4)(x-1)^2+(1/4)(y-1)^2$. The boundary of this is $(1/2)(x-1)^2+(1/2)(y-1)^2 = (3/4)(x-2)^2+(1/4)(y-2)^2$ which gives that $-x^2+8x+y^2-12=0$, or $x=4\pm\sqrt{y^2+4}$. Following an analysis similar to the one above leads to complications because x and y have a more complicated relation.

Taking any of the branches of $\bar{R}(x,y)$ and plugging in the intersection curve, nameyl $x=4\pm\sqrt{4+y^2}$, we get

$$\bar{R}(x,y)=\begin{cases} y^2+3\sqrt{y^2+4}-y+7, & x=4+\sqrt{4+y^2} \\ y^2-3\sqrt{y^2+4}-y+7, & x=4-\sqrt{4+y^2}. \end{cases}$$

Minimizing numerically the first one gives the minimum of 12.8568 attained at $y=0.286961$, while the second branch has minimum of 0.233339 for $y=1.33107$. Thus the minimum is the second one and is attained at $x=1.59755$ and $y=1.33107$.

Thus the minimax statistic in this case is $\hat{\theta}(-1)=1.59755$ and $\hat{\theta}=1.33107$.

If we follow the Bayesian approach we put the prior λ on 1 and $1-\lambda$ on 2 and then the Bayesian score of the estimator is

$$r(x,y)=\lambda((1/2)(x-1)^2+(1/2)(y-1)^2)+(1-\lambda)((3/4)(x-2)^2+(1/4)(y-2)^2)$$

Thus we need to minimize this over x,y and this is easy to be done because it is simply a quadratic function in x and y separately. We can actually separate these as

$$r(x,y)=(1/4)x^2(3-\lambda)+x(-3+2\lambda)+(1/4)y^2(1+\lambda)-y+4-3\lambda$$

which is minimized at $x_\lambda=-\frac{2(3-2\lambda)}{\lambda-3}$ and $y_\lambda=\frac{2}{\lambda+1}$. Thus the Bayesian statistic is given by

$$\hat{\theta}(t)=\begin{cases} -\frac{2(3-2\lambda)}{\lambda-3}, & t=-1 \\ \frac{2}{\lambda+1}, & t=1. \end{cases}$$

To find the minimax, we need to look at the risks for each of these branches adn that gives

$$R(x_\lambda,y_\lambda,\theta)=\begin{cases} \frac{(\lambda-1)^2\left(5\lambda^2+6\lambda+9\right)}{(\lambda-3)^2(\lambda+1)^2}, & \theta=1 \\ \frac{4\lambda^2\left(\lambda^2+3\right)}{(\lambda-3)^2(\lambda+1)^2}, & \theta=2 \end{cases}$$

Equating the two branches gives a minimax equation, which boils down in this case to the following condition on λ

$$0=\lambda^4-4\lambda^3-10\lambda^2-12\lambda+9$$

This has the roots $-\sqrt{5}-i\sqrt{2\sqrt{5}-3}+1, -\sqrt{5}+i\sqrt{2\sqrt{5}-3}+1, -\sqrt{5}-\sqrt{2\sqrt{5}-3}+1, -\sqrt{5}+\sqrt{2\sqrt{5}-3}+1$ or in numerical form $-1.23607-1.21332i, -1.23607+1.21332i, 0.502547, 5.96959$. As one can see, only one is in $(0,1)$ and this provides also the minimax estimator, which yields the same estimator as above, because $x_0.502547=1.59755$ and $y_0.502547=1.33107$. □

Remark 5. *Problem 4 is a prototype for how one computes in the case the parameter takes a discrete number of values and for each value we have a distribution assigned. This procedure can be extended to the general case, though in that case the problem becomes an optimization problem which in general is hard to solve by hand. For instance, if have multiple values of t in the distribution of $p(x;\theta)$, and we also have many sample points, we need to consider many variables and then the minimization becomes much*

more complicated. As a message is that from the computational side the Bayesian method is much much simpler and yields faster results, because for a fixed prior we can very easily minimize the average loss (at least in the case of a square loss).

Problem 6. *Assume we take a sample $X_1, X_2, \dots, X_n$ from $N(\theta, \sigma^2)$. Assume that a prior distribution on θ is $N(a, b^2)$.*

(1) *Find the posterior distribution of θ.*
(2) *Compute the expectation of this posterior distribution and denote it $\hat{\theta}$.*
(3) *Find the risk function associated to this estimator with respect to the square loss function $L(\hat{\theta}, \theta) = (\hat{\theta} - \theta)^2$.*
(4) *Find the maximum of this risk over all θ. For which values of a, b is this finite?*
(5) *Try the same thing, this time with respect to the loss function $L(\hat{\theta}, \theta) = (\hat{\theta} - \theta)^4$.*

Proof. (1) The posterior distribution of θ is given by

$$\begin{aligned} f(\theta|x) &= C(x) \exp\left(-\frac{\sum_{i=1}^n (x_i - \theta)^2}{2\sigma^2}\right) \exp\left(-\frac{(\theta - a)^2}{2b^2}\right) \\ &= C(x) \exp\left(-\frac{1}{2}\left(\frac{n}{\sigma^2} + \frac{1}{b^2}\right)\left(\theta - \frac{n\bar{x}/\sigma^2 + a/b^2}{n/\sigma^2 + 1/b^2}\right)^2\right) \end{aligned}$$

where $C(x)$ is a constant which depends on x. Thus

$$\theta|\vec{x} \sim N(\frac{n\bar{x}/\sigma^2 + a/b^2}{n/\sigma^2 + 1/b^2}, \frac{1}{\left(\frac{n}{\sigma^2} + \frac{1}{b^2}\right)}) = N(\frac{n\bar{x}b^2}{nb^2 + \sigma^2} + \frac{a\sigma^2}{nb^2 + \sigma^2}, \frac{b^2\sigma^2}{nb^2 + \sigma^2}).$$

(2) This in particular implies that the expectation of the posterior distribution is $\frac{n\bar{x}b^2}{nb^2+\sigma^2} + \frac{a}{nb^2+\sigma^2}$. Notice that this is not unbiased.
(3) The risk is the integral of the loss function against $f(x|\theta)$ which gives

$$\begin{aligned} R(\hat{\theta}, \theta) &= E\left[\left(\frac{n\bar{X}b^2 + a\sigma^2}{nb^2 + \sigma^2} - \theta\right)^2\right] \\ &= \frac{1}{(nb^2 + \sigma^2)^2}\mathbb{E}[(n\bar{X}b^2 + a\sigma^2 - \theta(nb^2 + \sigma^2))^2] \\ &= \frac{1}{(nb^2 + \sigma^2)^2}\mathbb{E}[(n(\bar{X} - \theta)b^2 + (a - \theta)\sigma^2)^2] \\ &= \frac{\sigma^2 b^4/n + (a - \theta)^2\sigma^4}{(nb^2 + \sigma^2)^2} \end{aligned}$$

where we used the fact that $\mathbb{E}[\bar{X}] = \theta$ and $\mathbb{E}[(\bar{X} - \theta)^2] = \sigma^2/n$. Obviously this is unbounded as we slide θ to infinity no matter what value of a and σ we take (except possibly $\sigma = 0$ which is in fact not allowed). This means that the standard Bayesian scheme of finding a minimax by figuring out the parameters a, b such that the risk is constant does not work.

(4) For the quartic case we have

$$\begin{aligned} R(\hat{\theta},\theta) &= E\left[\left(\frac{n\bar{X}b^2+a\sigma^2}{nb^2+\sigma^2}-\theta\right)^4\right] \\ &= \frac{1}{(nb^2+\sigma^2)^4}\mathbb{E}[(n(\bar{X}-\theta)b^2+(a-\theta)\sigma^2)^4] \\ &= \frac{3\sigma^4b^8/n^2+6\sigma^8b^4(a-\theta)^2/n+(a-\theta)^4\sigma^8}{(nb^2+\sigma^2)^4} \end{aligned}$$

and again this is not bounded no matter what a and b ($b \neq 0$, otherwise the prior is degenerate).

□

Problem 7. *Assume that $X \sim N(\theta,1)$ and take the estimator $\hat{\theta}_c = cX$ for some constant. Given a loss function L, we say that the estimator $\hat{\theta}$ is better than $\bar{\theta}$ if $R(\hat{\theta},\theta) \leq R(\bar{\theta},\theta)$ for all choices of $\theta \in \mathbb{R}$.*

(1) *Find the risk associated to $\hat{\theta}$ for the square loss function. Show that θ_1 is better than θ_c for any $c > 1$.*
(2) *Is $\hat{\theta}_{1/2}$ better than θ_1? Comment?*

Proof. (1) For the risk associated to $\hat{\theta}$, we need to compute $R(\hat{\theta},\theta) = \mathbb{E}[(cX-\theta)^2] = c^2\mathbb{E}[X^2] - 2c\theta\mathbb{E}[X] + \theta^2 = c^2(\theta^2+1) - 2c\theta^2 + \theta^2 = c^2 + (c-1)^2\theta^2$. To compare this with $\hat{\theta}_1$, we need to look at the difference

$$R(\hat{\theta}_c,\theta) - R(\hat{\theta}_1,\theta) = c^2 - 1 + (c-1)^2\theta^2$$

which is indeed nonnegative for $c \geq 1$, thus $\hat{\theta}_1$ is definitely better than $\hat{\theta}_c$.

(2) Obviously with the same computation as above we have that

$$R(\hat{\theta}_{1/2},\theta) - R(\hat{\theta}_1,\theta) = -3/4 + (1/4)\theta^2$$

which is not non-negative for all values of θ. For θ small, this is negative. This means that for small values of θ, $\hat{\theta}_{1/2}$ is better than $\hat{\theta}_1$ but this is not better for larger values of θ.

□

Problem 8. *If we take the loss function of the form*

$$L(\hat{\theta},\theta) = w(\theta)(\hat{\theta}-\theta)^2 \text{ with } w(\theta) > 0 \text{ for all } \theta$$

and consider a sample $X_1, X_2, \ldots, X_n$ from the distribution $f(x|\theta)$ and θ is assumed to follow the distribution $f(\theta)$, show that the Bayes estimator is given by

$$\hat{\theta}^f(\vec{x}) = \frac{\int \theta w(\theta) f(\theta|\vec{x})d\theta}{\int w(\theta)f(\theta|\vec{x})d\theta} = \frac{\int \theta w(\theta) f(\vec{x}|\theta)f(\theta)d\theta}{\int w(\theta)f(\vec{x}|\theta)f(\theta)d\theta} \tag{0.1}$$

where $\vec{x} = (x_1, x_2, \ldots, x_n)$ while

$$f(\theta|\vec{x}) = \frac{f(\vec{x}|\theta)f(\theta)}{m(\vec{x})} \text{ with } m(\vec{x}) = \int f(\vec{x}|\theta)f(\theta)d\theta \text{ and } f(\vec{x}|\theta) = f(x_1|\theta)f(x_2|\theta)\ldots f(x_n|\theta)f(\theta).$$

Proof. This is entirely base on the same argument as the conditional expectation. We have from the main theorem we discussed in class, that the Bayes estimator is defined as

$$\hat{\theta}^f = argmin_{z\in\mathbb{R}} \int L(z,\theta)f(\theta|x)d\theta.$$

To minimize

$$\int L(z,\theta)f(\theta|x)d\theta = \int w(\theta)(z-\theta)^2 f(\theta|x)d\theta.$$

Taking the derivative with respect to z and setting this to be equal to 0 yields that

$$\int (z-\theta)w(\theta)f(\theta|x)d\theta = 0$$

and this solves as

$$\hat{\theta}^f(\vec{x}) = z = \frac{\int \theta w(\theta)f(\theta|\vec{x})d\theta}{\int w(\theta)f(\theta|\vec{x})d\theta} = \frac{\int \theta w(\theta)f(\vec{x}|\theta)f(\theta)d\theta}{\int w(\theta)f(\vec{x}|\theta)f(\theta)d\theta}$$

which gives the claim.

□

Problem 9. (1) *Assume $X_1, X_2, \ldots, X_n$ is a sample from a Bernoulli random variable with parameter p. Under the assumption that the prior of p is a $Beta(\alpha,\beta)$, find the posteriori distribution of p. Show that this estimator is a Bayes estimator for the square loss function.*

(2) *Find the risk function for the above estimator and a minimax estimator in this case.*

(3) *Under the same assumptions, assume that the loss function is $L(\hat{p},p) = \frac{(\hat{p}-p)^2}{p(1-p)}$. Find a Bayes rule for this loss function.*

(4) *Can you do the same for the loss function $L(\hat{\theta},\theta) = \frac{(\hat{p}-p)^2}{p^a(1-p)^b}$ with $a,b \geq 0$ integer numbers?*

Proof. This uses Problem 8, particularly (0.1) for the computation of the Bayesian estimator.

(1) We did this in class. The estimator is (denote here $s = \sum_{i=1}^n x_i$)

$$\hat{p}_{\alpha,\beta}(x) = \int_0^1 pf(p|\vec{x})dp = \frac{\int_0^1 p^s(1-p)^{n-s}p^{\alpha-1}(1-p)^{\beta-1}dp}{\int_0^1 pp^s(1-p)^{n-s}p^{\alpha-1}(1-p)^{\beta-1}dp} = \frac{B(\alpha+s+1,\beta+n-s)}{B(\alpha+s,\beta+n-s)} = \frac{s+\alpha}{\alpha+\beta+n}.$$

where in the last step we used (0.3).

(2) The risk function for the above estimator is given by (here $S = \sum_{i=1}^n X_i$)

$$\begin{aligned} R(\alpha,\beta,\theta) &= \mathbb{E}[(\hat{p}-p)^2] = \mathbb{E}[(\frac{S+\alpha}{\alpha+\beta+n}-p)^2] = \frac{\mathbb{E}[(S+\alpha-(\alpha+\beta+n)p)^2]}{(\alpha+\beta+n)^2} \\ &= \frac{\mathbb{E}[S^2]+2(\alpha-p(\alpha+\beta+n))\mathbb{E}[S]+(\alpha-p(\alpha+\beta+n))^2)}{(\alpha+\beta+n)^2} \\ &= \frac{np(1-p)+n^2p^2+2(\alpha-p(\alpha+\beta+n))np+(\alpha-p(\alpha+\beta+n))^2}{(\alpha+\beta+n)^2} \\ &= \frac{\alpha^2+(n-2\alpha(\alpha+\beta))p+((\alpha+\beta)^2-n)p^2}{(\alpha+\beta+n)^2} \end{aligned}$$

Now, to make sure this is independent of p, we need to equate the coefficients of p and p^2 to 0. This way we will solve for α and β. This leads to $\alpha = \beta = \sqrt{n}/2$ and thus

$$R(\hat{p},p) = \frac{n/2}{(n+\sqrt{n})^2} = \frac{1}{2(\sqrt{n}+1)^2}.$$

Notice that this converges to 0 as $n \to \infty$. This means that $\frac{S+\sqrt{n}/2}{n+\sqrt{n}}$ is the minimax estimator.

(3–4) We will do the more general case for

$$L(\hat{p},p) = \frac{(\hat{p}-p)^2}{p^a(1-p)^b}.$$

In this case the Bayesian estimator is given by

$$\hat{p}_{a,b,\alpha,\beta}(x) = \frac{\int_0^1 pp^{-a}(1-p)^{-b}p^s(1-p)^{n-s}p^{\alpha-1}(1-p)^{\beta-1}dp}{\int_0^1 p^{-a}(1-p)^{-b}p^s(1-p)^{n-s}p^{\alpha-1}(1-p)^{\beta-1}dp}$$
$$= \frac{B(s+\alpha-a+1,\beta-b+n-s)}{B(s+\alpha-a,\beta-b+n-s)} = \frac{s+\alpha-a}{\alpha+\beta-a-b+n}$$

In the same spirit as above, we have that the risk function is given by

$$\mathbb{E}[\frac{(\hat{p}-p)^2}{p^a(1-p)^b}] = \frac{\mathbb{E}[(S+\alpha-a-(\alpha+\beta-a-b+n)p)^2]}{(\alpha+\beta-a-b+n)^2p^a(1-p)^b}$$
$$= \frac{(\alpha-a)^2+(n-2(\alpha-a)(\alpha+\beta-a-b))p+((\alpha+\beta-a-b)^2-n)p^2}{(\alpha+\beta-a-b+n)^2p^a(1-p)^b}. \tag{0.2}$$

To make this independent of p, we need to take $a+b \leq 2$, which means that $(a,b) = (0,0),(0,1),(1,0),(1,1),(2,0),(0,2)$. The case $a,b=0$ is the first case. The case of $a=1$ and $b=0$ leads to the conclusion that we need to take $\alpha = a = 1$ and $\beta = \sqrt{n}$, from which then

$$\hat{p}_{1,0,1,\sqrt{n}}(x) = \frac{s}{n+\sqrt{n}}.$$

In a similar fashion we get that for $a=0$, $b=1$, $\alpha+\beta-1=\sqrt{n}$ and we must have then $\alpha^2 = -n+2\alpha(\alpha+\beta-1)$ which solves as $\alpha=\sqrt{n}$ and $\beta=1$, thus the estimator is

$$\hat{p}_{0,1,\sqrt{n},1}(x) = \frac{s+\sqrt{n}}{n+\sqrt{n}}.$$

For the case of $a=b=1$, we need to impose the condition that $\alpha=1$ and $n=n-(\beta-1)^2$, or $\beta=1$. Thus in this case the estimator becomes

$$\hat{p}_{1,1,1,1}(x) = \frac{s}{n} = \bar{x}.$$

In the remaining cases of $a=2,b=0$, we must have $\alpha=2$ and $n=0$ (from the coefficient of p in the numerator of (0.2)) which is not possible. Thus we can not determine an estimator using this method for the case of $a=2,b=0$.

For the other case, $a=0,b=2$, we must impose the condition that $\alpha^2 = (\alpha+\beta-2)^2 - n = -(n-2\alpha(\alpha+\beta-2))/2$ which does not have a solution. Thus the only cases we can solve through this method for the minimax, is the case of $a=b=0$, $a=1,b=0$, $a=0,b=1$ and $a=b=1$.

□

Problem 10. *Let X be a Bernoulli with parameter p and assume the prior is uniform on $[0,1]$.*

(1) *Find the posterior distribution of p and then the Bayes estimator for the square loss function.*
(2) *If in addition, we know that $p \notin (1/3,2/3)$ and use the prior to be the uniform on the $(0,1/3)\cup(2/3,1)$, find the Bayes estimator now, again with respect to the square loss function.*

Proof. (1) The first part is already done in class and also in Problem 9.

(2) If the prior is the uniform distribution on $(0,1/3)\cup(2/3,1)$, then the Bayes posterior is

$$\hat{p}(x) = \int pf(p|x)dp = \int_{(0,1/3)\cup(2/3,1)} p^{x+1}(1-p)^{1-x}\frac{dp}{2/3} = \begin{cases} 7/54, & x=0 \\ 10/27, & x=1. \end{cases}$$

□

Problem 11. *This problem is about the* Γ *and* $Beta$ *distributions. For* $\alpha, \beta > 0$*, we set*

$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x} dx$$

while

$$B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}.$$

(1) $\Gamma(\alpha+1) = \alpha\Gamma(\alpha)$ *and* $\Gamma(1) = 1$, $\Gamma(n) = (n-1)!$.
(2) *Show that*

$$f_{\alpha,\beta}(x) = \begin{cases} \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x} dx, \text{ for } x > 0 \\ 0, \text{ for } x < 0 \end{cases}$$

is a density. It is the density of a Gamma distribution with parameters $\alpha, \beta > 0$*, in short* $\Gamma(\alpha, \beta)$.
(3) *Compute the mean and the variance of* $\Gamma(\alpha, \beta)$.
(4) *Show that*

$$B(\alpha+1, \beta) = B(\alpha, \beta)\frac{\alpha}{\alpha+\beta} \tag{0.3}$$

(5) *Show that the following*

$$f_{\alpha,\beta}(x) = \begin{cases} \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)}, 0 < x < 1 \\ 0, \text{ otherwise} \end{cases}$$

is a density. This is called the $Beta(\alpha, \beta)$*. Compute the mean and the variance of* $Beta(\alpha, \beta)$.

Proof. (1) Use an integration by part to get $\Gamma(\alpha+1) = \alpha\Gamma(\alpha)$. Also observe that $\Gamma(1) = 1$ and then use the formula recursively to get $\Gamma(n+1) = n!$.
(2) This follows easily from a simple change of variable $x = y/\beta$.
(3) The mean of $\Gamma(\alpha, \beta)$ is α/β and the variance is α/β^2.
(4) The Beta relation is a consequence of the first item.
(5) To show that $f_{\alpha,\beta}$ is a density we need to compute the integral

$$\int_0^1 x^{\alpha-1}(1-x)^{\beta-1} dx.$$

This can be found at https://en.wikipedia.org/wiki/Beta_function.
The mean is $\alpha/(\alpha+\beta)$ and the variance is $\alpha\beta/((\alpha+\beta)^2(\alpha+\beta+1))$.

□

Problem 12. *Assume we have a sample* $X_1, X_2, \ldots, X_n$ *from a Poisson distribution* $f(x;\theta)$ *with* $\theta > 0$*, i.e.* $f(x;\theta) = e^{-\theta}\frac{\theta^x}{x!}$ *for* $x = 0, 1, 2, 3 \ldots$*. Assume now that the prior distribution of* θ *is* $\Gamma(\alpha, \beta)$.

(1) *Find the Bayesian estimate of* θ.
(2) *Can you use this for some parameters* $\alpha, \beta > 0$ *to find a minimax estimator?*

Proof. (1) The posterior is given by (here $s = \sum_{i=1}^n x_i$)

$$f(\theta|\vec{x}) = C(\vec{x})e^{n\theta}\theta^s\theta^{\alpha-1}e^{-\beta\theta} = C(\vec{x})\theta^{\alpha+s-1}e^{-(\beta-n)\theta}$$

where the constant $C(\vec{x})$ is obtained by integrating so this becomes a density, i.e.

$$C(\vec{x})\int_0^\infty \theta^{\alpha+s-1}e^{-(\beta-n)\theta} = 1$$

which gives

$$C(\vec{x})\Gamma(\alpha+s)/(\beta-n)^{\alpha+s}=1 \text{ so } C(\vec{x})=\frac{(\beta-n)^{\alpha+s}}{\Gamma(\alpha+s)}$$

The mean of the posterior distribution of θ gives $\hat{\theta}$ is given by (assuming here that $\beta > n$ so that things are integrable)

$$\hat{\theta}(\vec{x})=C(\vec{x})\int_0^\infty \theta^{\alpha+s+1-1}e^{-(\beta-n)\theta}=\frac{(\beta-n)^{\alpha+s}}{\Gamma(\alpha+s)}\frac{\Gamma(\alpha+s+1)}{(\beta-n)^{\alpha+s+1}}=\frac{(\alpha+s)}{\beta-n}.$$

(2) Is we take the square loss we want to compute the risk associated to this estimator. We get (using here that s is Poisson with parameter $n\theta$)

$$\begin{aligned}R(\hat{\theta},\theta)=\mathbb{E}[(\frac{\alpha+s}{\beta-n}-\theta)^2]&=\frac{\mathbb{E}[(s+\alpha-\theta\beta+n\theta)^2]}{(\beta-n)^2}\\&=\frac{\mathbb{E}[(s-n\theta+(\alpha-(\beta-2n)\theta)^2]}{(\beta-n)^2}\\&=\frac{n\theta+(\alpha-(\beta-2n)\theta)^2}{(\beta-n)^2}\end{aligned}$$

Thus, no matter how we try this be independent of θ is not going to be constant for any choices of α, β. This method does not produce a minimax rule.

□

Problem 13 (optional)**.** *Assume that we have a sample $X_1, X_2, \ldots, X_n$ from a given distribution $f(x;\theta)$ for which there exists a sufficient statistic $T = u(X_1, X_2, \ldots, X_n)$.*

(1) *Show that for any prior f, the Bayesian estimator $\hat{\theta}^f$ is a function of T.*
(2) *Show that if the loss function is convex and $\hat{\theta}$ is an estimator, then $\tilde{\theta} = \mathbb{E}[\hat{\theta}|T] = \psi(T)$ has a smaller risk for any value of θ. (Hint: You need here Jensen's inequality for the conditional expectation).*
(3) *Show that if the loss function $L(\hat{\theta},\theta)$ is convex in $\hat{\theta}$, then any minimax estimator is a function of T.*