ORFE 524: Statistical Theory and Methods
Homework 3

Zachary Hervieux-Moore

Friday 28th October, 2016

Exercise 1: Suppose $\theta \in \mathbb{R}^d$. Show twidehat the following loss functions are convex.

1) $\mathcal{L}(a) = \|a - \theta\|_p$, $p \geq 1$,

2) $\mathcal{L}(a) = \|a - \theta\|_p^q$, $p, q \geq 1$,

Answer:

1) We have from Minkowski's inequality,

$$\begin{aligned}
\mathcal{L}(\lambda x + (1-\lambda)y) &= \|\lambda x + (1-\lambda)y - \theta\|_p \\
&= \|\lambda x + (1-\lambda)y - (1-\lambda)\theta - \lambda\theta\|_p \\
&\leq \|\lambda x - \lambda\theta\|_p + \|(1-\lambda)y - (1-\lambda)\theta\|_p \\
&= \lambda\|x - \theta\|_p + (1-\lambda)\|y - \theta\|_p \\
&= \lambda\mathcal{L}(x) + (1-\lambda)\mathcal{L}(y)
\end{aligned}$$

Thus, $\mathcal{L}(a)$ is convex for $p \geq 1$ since Minkowski's inequality holds for all $p \geq 1$.

2) Following the same set of steps as before,

$$\begin{aligned}
\mathcal{L}(\lambda x + (1-\lambda)y) &= \|\lambda x + (1-\lambda)y - \theta\|_p^q \\
&= \lambda^q\|x - \theta\|_p^q + (1-\lambda)^q\|y - \theta\|_p^q
\end{aligned}$$

Since $\lambda \in [0, 1]$, and $q \geq 1$, $\lambda^q \leq \lambda$,

$$\begin{aligned}
&\leq \lambda\|x - \theta\|_p^q + (1-\lambda)\|y - \theta\|_p^q \\
&= \lambda\mathcal{L}(x) + (1-\lambda)\mathcal{L}(y)
\end{aligned}$$

Again, we get $\mathcal{L}(a)$ is convex if $p, q \geq 1$.

Exercise 2: Let $X \sim Ber^n(p)$ where $p \in (0,1)$. Consider the naive estimator $\widehat{p} = X_1$, and let $\tilde{p} = \mathbb{E}[\widehat{p}|T]$, where $T(X) = \sum_{i=1}^{n} X_i$.

1) Derive $\tilde{p}$.

2) Compute and compare $\mathbb{E}[(\widehat{p} - p)^2]$ and $\mathbb{E}[(\tilde{p} - p)^2]$

Answer:

1) We have,

$$\tilde{p} = \mathbb{E}[\widehat{p}|T] = 0 \cdot \frac{n-T}{n} + 1 \cdot \frac{T}{n} = \frac{T}{n}$$

2) $\widehat{p}$ is $Bernoulli(p)$, so,

$$\mathbb{E}[(\widehat{p} - p)^2] = p(1-p)$$

From before, $\tilde{p} = \frac{T}{n}$,

$$\mathbb{E}[(\tilde{p} - p)^2] = \mathbb{E}[(T/n)^2] - 2p\mathbb{E}[T/n] + p^2 = \frac{\mathbb{E}[T^2]}{n^2} - 2p\frac{\mathbb{E}[T]}{n} + p^2$$

Where T is $Binormial(n,p)$,

$$= \frac{np(1-p+np)}{n^2} - 2p\frac{np}{n} + p^2 = \frac{p(1-p+np)}{n} - p^2 = \frac{p(1-p)}{n}$$

Thus, $\mathbb{E}[(\widehat{p}-p)^2]/n = \mathbb{E}[(\tilde{p}-p)^2]$, so the bias of $\tilde{p}$ disappears with large sample size but $\widehat{p}$ remains constantly biased.

Exercise 3: Let us prove Rao-Blackwell theorem in a different way. Denote the ℓ_2 loss by $R(\widehat{\theta}) = \mathbb{E}[\|\widehat{\theta} - \theta\|^2]$.

1) Show that for any two random variables X, Y defnied on the same space $(\Omega, \Sigma, \mathbb{P})$,

$$Var(X) = Var(\mathbb{E}[X|Y]) + \mathbb{E}[Var(X|Y)]$$

Recall that by definition, $\mathbb{E}[X|Y]$ is a function of Y, and $Var(X|Y)$ is defined to be $\mathbb{E}\left[(X - \mathbb{E}[X|Y])^2|Y\right]$. Hint: use the fact that $\mathbb{E}[X] = \mathbb{E}[\mathbb{E}[X|Y]]$.

2) Use the decomposition of variance in 1) to prove Rao-Blackwell theorem.

Answer:

1) We begin with $Var(X)$,

$$\begin{aligned} Var(X) &= \mathbb{E}[X^2] - \mathbb{E}[X]^2 = \mathbb{E}[\mathbb{E}[X^2|Y]] - \mathbb{E}[\mathbb{E}[X|Y]]^2 \\ &= \mathbb{E}[Var(X|Y) + \mathbb{E}[X|Y]^2] - \mathbb{E}[\mathbb{E}[X|Y]]^2 \\ &= \mathbb{E}[Var(X|Y)] + \mathbb{E}[\mathbb{E}[X|Y]^2] - \mathbb{E}[\mathbb{E}[X|Y]]^2 \\ &= \mathbb{E}[Var(X|Y)] + Var(\mathbb{E}[X|Y]) \end{aligned}$$

Thus, $Var(X) = Var(\mathbb{E}[X|Y]) + \mathbb{E}[Var(X|Y)]$

2) We begin with an unbiased estimator $\widehat{\theta}$ and $\tilde{\theta} = \mathbb{E}[\widehat{\theta}|T]$. Note that $\tilde{\theta}$ is also unbiased,

$$\mathbb{E}[\tilde{\theta}] = \mathbb{E}[\mathbb{E}[\widehat{\theta}|T]] = \mathbb{E}[\widehat{\theta}] = \theta$$

Now we work with the loss,

$$\begin{aligned} R(\widehat{\theta}) &= \mathbb{E}[\|\widehat{\theta} - \theta\|^2] = \mathbb{E}[(\widehat{\theta} - \theta)((\widehat{\theta} - \theta))^T] \\ &= \mathbb{E}[(\widehat{\theta} - \mathbb{E}[\widehat{\theta}])((\widehat{\theta} - \mathbb{E}[\widehat{\theta}]))^T] = Var(\widehat{\theta}) \end{aligned}$$

The same reasoning yields,

$$R(\tilde{\theta}) = Var(\tilde{\theta}) = Var(\mathbb{E}[\widehat{\theta}|T])$$

So using our variance decomposition,

$$\begin{aligned} R(\widehat{\theta}) &= Var(\widehat{\theta}) = Var(\mathbb{E}[\widehat{\theta}|T]) + \mathbb{E}[Var(\widehat{\theta}|T)] \\ &= Var(\tilde{\theta}) + \mathbb{E}[Var(\widehat{\theta}|T)] \geq Var(\tilde{\theta}) = R(\tilde{\theta}) \end{aligned}$$

Which is the Rao-Blackwell theorem.

Exercise 4: In K-means algorithm, we do the following steps:

- Step 1: start with $\{c_j\}^0$, each with K initial centers in $\mathbb{R}^d$.
- Step 2: for $l = 0, 1, \ldots$
  - Assign each x_j to the nearest center $c_j^l \in \{c_j\}^l$. If a tie appears, x_j is assigned to one of its nearest centers in an arbitrary way.
  - Updating the centers: compute the mean of points in class C_j^l, where C_j^l consists of points that are assigned to the center c_j^l. Let c_j^{l+1} be that mean.

Show that K-means algorithm converges (not necessarily to its global minimizer), i.e. $\lim_{l\to\infty} \phi(\{c_j\}^l)$ exists, where

$$\phi(\{c_j\}^l) = \sum_{j=1}^{K} \sum_{x_i \in C_j^l} \|x_i - c_j^l\|^2$$

Answer: We start with,

$$\phi(\{c_j\}^l) = \sum_{j=1}^{K} \sum_{x_i \in C_j^l} \|x_i - c_j^l\|^2$$

Note that c_j^{l+1} is the mean for C_j^l. That is, c_j^{l+1} minimizes $\|\cdot\|^2$ for C_j^l. So,

$$\sum_{j=1}^{K} \sum_{x_i \in C_j^l} \|x_i - c_j^l\|^2 \geq \sum_{j=1}^{K} \sum_{x_i \in C_j^l} \|x_i - c_j^{l+1}\|^2$$

Also, the reassignment of points to closer centers in Step 1 implies,

$$\sum_{j=1}^{K} \sum_{x_i \in C_j^l} \|x_i - c_j^{l+1}\|^2 \geq \sum_{j=1}^{K} \sum_{x_i \in C_j^{l+1}} \|x_i - c_j^{l+1}\|^2 = \phi(\{c_j\}^{l+1})$$

$$\iff \phi(\{c_j\}^l) \geq \phi(\{c_j\}^{l+1})$$

So we have a decreasing sequence which is lower bounded by 0. Hence, it converges and $\lim_{l\to\infty} \phi(\{c_j\}^l)$ exists.

Exercise 5: Suppose that $\theta \in \mathbb{R}^p$, and a statistic $T(X)$ is also in $\mathbb{R}^d$, with $\mathbb{E}_\theta[T(X)] = g(\theta)$ for some function $g : \mathbb{R}^d \to \mathbb{R}^d$. Cramer-Rao Theorem states that under the same regularity conditions as in the univariate case, we have

$$Cov(T) \succeq \nabla_\theta g(\theta)^T I(\theta)^{-1} \nabla_\theta g(\theta),$$

where $A \succeq B$ means $A - B$ is a positive semi-definite matrix, $(\nabla_\theta g(\theta))_{ij} = \frac{\partial}{\partial \theta_i} g_j(\theta)$, and

$$I(\theta) = \mathbb{E}[(\nabla_\theta \log f_\theta(x))(\nabla_\theta \log f_\theta(x))^T]$$

Where $I(\theta)$ is assumed to exist and be invertible. You can prove this theorem using the following ideas.

1) Let $a(x) = \nabla_\theta \log f_\theta(x)$. Derive $Cov((T, a)^T)$ in terms of $Cov(T)$, $\nabla_\theta g(\theta)$, and $I(\theta)$.

2) Find a matrix B such that

$$B^T Cov \begin{pmatrix} T \\ a \end{pmatrix} B = Cov(T) - \nabla_\theta g(\theta)^T I(\theta)^{-1} \nabla_\theta g(\theta)$$

3) Conclude the proof of Cramer-Rao Theorem

Answer: 1) Following the suggestion, let $a(x) = \nabla_\theta \log f_\theta(x)$. Now,

$$\begin{aligned} Cov((T, a)^T) &= \mathbb{E}[(T, a)^T (T, a)] - \mathbb{E}[(T, a)^T]\mathbb{E}[(T, a)^T]^T \\ &= \mathbb{E} \begin{bmatrix} TT^T & Ta^T \\ aT^T & aa^T \end{bmatrix} - \begin{bmatrix} \mathbb{E}[T]\mathbb{E}[T]^T & \mathbb{E}[T]\mathbb{E}[a]^T \\ \mathbb{E}[a]\mathbb{E}[T]^T & \mathbb{E}[a]\mathbb{E}[a]^T \end{bmatrix} \end{aligned}$$

But,

$$\begin{aligned} \mathbb{E}[a] = \mathbb{E}[\nabla_\theta \log f_\theta(x)] &= \mathbb{E}[\frac{\nabla_\theta f_\theta(x)}{f_\theta(x)}] \\ &= \int \frac{\nabla_\theta f_\theta(x)}{f_\theta(x)} f_\theta(x) dx \\ = \int \nabla_\theta f_\theta(x) dx = \nabla_\theta \int f_\theta(x) dx &= \nabla_\theta 1 = 0 \end{aligned}$$

Where I can take out the gradient because of regularity. So,

$$Cov((T,a)^T) = \mathbb{E}\begin{bmatrix} TT^T & Ta^T \\ aT^T & aa^T \end{bmatrix} - \begin{bmatrix} \mathbb{E}[T]\mathbb{E}[T]^T & 0 \\ 0 & 0 \end{bmatrix}$$
$$\begin{bmatrix} \mathbb{E}[TT^T] - \mathbb{E}[T]\mathbb{E}[T]^T & \mathbb{E}[Ta^T] \\ \mathbb{E}[aT^T] & \mathbb{E}[aa^T] \end{bmatrix}$$

Note that,

$$\mathbb{E}[aa^T] = \mathbb{E}[\nabla_\theta \log f_\theta(x)(\nabla_\theta \log f_\theta(x))^T] = I(\theta)$$
$$\mathbb{E}[TT^T] - \mathbb{E}[T]\mathbb{E}[T]^T = Cov(T)$$
$$\mathbb{E}[Ta^T] = \int T(x)\frac{\nabla_\theta f_\theta(x)}{f_\theta(x)} f_\theta(x)dx$$
$$\mathbb{E}[Ta^T] = \nabla_\theta \int T(x) f_\theta(x)dx = \nabla_\theta g(\theta)$$

Finally,

$$Cov((T,a)^T) = \begin{bmatrix} Cov(T) & \nabla_\theta g(\theta) \\ \nabla_\theta g(\theta)^T & I(\theta) \end{bmatrix}$$

2) With the benefit of hindsight, I choose B to be

$$B = \begin{bmatrix} I_p \\ -I^{-1}(\theta)\nabla_\theta g(\theta) \end{bmatrix}$$

Then,

$$B^T Cov\begin{pmatrix} T \\ a \end{pmatrix} B = Cov(T) - 2\nabla_\theta g(\theta)^T I^{-1}(\theta)\nabla_\theta g(\theta) + \nabla_\theta g(\theta)^T I^{-1}(\theta)\nabla_\theta g(\theta)$$
$$= Cov(T) - \nabla_\theta g(\theta)^T I^{-1}(\theta)\nabla_\theta g(\theta)$$

3) The above $B^T Cov\begin{pmatrix} T \\ a \end{pmatrix} B$ is PSD since all covariance matrices are symmetric. So we can decompose this matrix into something like $B^T C^T CB$ and so for all vectors x, $x^T B^T C^T CBx = \|CBx\|_2^2 \geq 0$. Since the LHS is PSD, so is the RHS, so $Cov(T) \succeq \nabla_\theta g(\theta)^T I(\theta)^{-1}\nabla_\theta g(\theta)$.

Exercise 6: Suppose that $f_\theta(x)$ is the density function of $P_\theta \in \mathcal{P}$, where $\theta \in \mathbb{R}^k$. We assume that $f_\theta(x)$ is twice differentiable in θ, satisfying the regular condition, meaning that

$$\nabla_\theta \int h_\theta(x)dx = \int \nabla_\theta h_\theta(x)dx$$

holds for $h_\theta = f_\theta$ and $h_\theta = \nabla_\theta f_\theta$. Show that

$$I(\theta) = -\mathbb{E}[\nabla_\theta^2 \log f_\theta(X)]$$

Answer: Begin with the inside of the expectation,

$$\begin{aligned}\nabla_\theta^2 \log f_\theta(X) &= \frac{\nabla_\theta^2 f_\theta(X)}{f_\theta(X)} - \left(\frac{\nabla_\theta f_\theta(X)}{f_\theta(X)}\right)^T \left(\frac{\nabla_\theta f_\theta(X)}{f_\theta(X)}\right) \\ &= \frac{\nabla_\theta^2 f_\theta(X)}{f_\theta(X)} - (\nabla_\theta \log f_\theta(X))^T (\nabla_\theta \log f_\theta(X))\end{aligned}$$

Now take expectation of the first term

$$\mathbb{E}\left[\frac{\nabla_\theta^2 f_\theta(X)}{f_\theta(X)}\right] = \nabla_\theta^2 \int f_\theta(X)dx$$

where the equality comes from regularity conditions,

$$= \nabla_\theta^2 \int f_\theta(X)dx = \nabla_\theta^2 1 = 0$$

So, we have,

$$\mathbb{E}[\nabla_\theta^2 \log f_\theta(X)] = -\mathbb{E}[(\nabla_\theta \log f_\theta(X))^T (\nabla_\theta \log f_\theta(X))]$$

Which was what we wished to show.

Exercise 7: Let $f_\alpha(x) = h(x)l(\alpha)e^{\alpha T(x)}$, $x \in \mathbb{R}$, be a density function of probability measure P_α in the exponential family $\mathcal{P} = \{P_\alpha : \alpha \in \mathcal{A}\}$. Suppose that $\mathcal{A}$ is an open set in $\mathbb{R}$.

1) Show that

$$\left|\frac{e^{az}-1}{z}\right| \leq \frac{e^{\delta|a|}}{\delta}$$

holds for $|z| \leq \delta$

2) Use the Dominated Convergence Theorem to show the following. Let g be a Borel function such that $\mathbb{E}[g] < \infty$. Show that $\frac{d}{d\alpha}\mathbb{E}[g] = \int g(x)\frac{d}{d\alpha}f_\alpha(x)dx$ (you may assume the l.h.s. is differentiable).

Answer:

1) Expand the numerator using the Taylor series for e^x,

$$\begin{aligned}\left|\frac{e^{az}-1}{z}\right| = \frac{|e^{az}-1|}{|z|} &= \frac{|\sum_{n=0}^{\infty}\frac{(az)^n}{n!}-1|}{|z|} = \frac{|\sum_{n=1}^{\infty}\frac{(az)^n}{n!}|}{|z|} \\ &\leq \frac{\sum_{n=1}^{\infty}\frac{|a|^n|z|^n}{n!}}{|z|} = \sum_{n=1}^{\infty}\frac{|a|^n|z|^{n-1}}{n!} \leq \sum_{n=1}^{\infty}\frac{|a|^n\delta^{n-1}}{n!} \\ &= \frac{\sum_{n=1}^{\infty}\frac{|a|^n\delta^n}{n!}}{\delta} = \frac{e^{\delta|a|}}{\delta}\end{aligned}$$

2) Recall that f_α is exponential,

$$\mathbb{E}[g] = \int g(x)h(x)l(\alpha)e^{\alpha T(x)}dx$$

Remember that $\mathcal{A}$ is open, so we can find a sequence $\epsilon_n \to 0$ such that for sufficiently large N, all $n \geq N$, $\epsilon_n \in \mathcal{A}$. So we work with this sequence that is inside $\mathcal{A}$. Also, $l(\alpha)$ can be removed from the integral and so we ignore it for now. We now write the definition of differentiation,

$$\begin{aligned}&\lim_{n\to\infty}\frac{\int g(x)h(x)(e^{(\alpha+\epsilon_n)T(x)}-e^{\alpha T(x)})dx}{\epsilon_n} \\ &= \lim_{n\to\infty}\int g(x)h(x)e^{\alpha T(x)}\frac{(e^{\epsilon_n T(x)}-1)}{\epsilon_n}dx\end{aligned}$$

However, we showed that $\frac{(e^{\epsilon_n T(x)}-1)}{\epsilon_n}$ is dominated and the resulting integral is,

$$\leq \int g(x)h(x)e^{\alpha T(x)}\frac{e^{\delta|T(x)|}}{\delta}dx$$

Which is integrable since it is of the exponential family. Thus, we can take the limit inside by Dominated Convergence Theorem and we get differentiation. The last detail that needs to be dealt with is the $l(\alpha)$ sitting outside of the integral. Last homework we showed,

$$l(\alpha) = \frac{1}{\int h(x)e^{\alpha T(x)}dx}$$

Thus, since we can differentiate the reciprocal ($g(x) = 1$), we can differentiate this using quotient rule. By combining this with product rule, we can differentiate the entire RHS.

Exercise 8: Consider the fixed design model $y = X\beta + \eta$, where $y \in \mathbb{R}^n$, $\beta \in \mathbb{R}^d$, and $X \in \mathbb{R}^{n\times d}$. Here 'fixed design' means that X is a deterministic matrix. Suppose that $n \geq d$, X^TX is invertible, and $\eta \sim N(0, \sigma^2 I_n)$, where σ^2 is known.

1) Show that $\widehat{\beta} = (X^TX)^{-1}X^Ty$ is a UMVUE of β.

2) Derive the mean square error $R_2(\widehat{\beta})$ in terms of σ^2 and $(X^TX)^{-1}$ only. What would $R_2(\widehat{\beta})$ be if $n^{-1}X^TX$ is equal to the identity matrix?

Answer:

1) First, we show that $\widehat{\beta}$ is unbiased,

$$\widehat{\beta} = \mathbb{E}[(X^TX)^{-1}X^Ty] = (X^TX)^{-1}X^T\mathbb{E}[y]$$
$$= (X^TX)^{-1}X^T\mathbb{E}[X\beta + \eta] = (X^TX)^{-1}X^TX\beta = \beta$$

Where η disappears as it has mean 0 and the X's are constant (fixed design setting). We now show that $\widehat{\beta}$ achieves the Cramer-Rao Lower bound and hence is a UMVUE. Computing the Fisher information, start with log likelihood,

$$\log \mathcal{L}(\beta, \sigma^2 I_n) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{(y - X\beta)^T(y - X\beta)}{2\sigma^2}$$

Now differentiate w.r.t. β,

$$\frac{\partial \log \mathcal{L}(\beta, \sigma^2 I_n)}{\partial \beta} = -\frac{1}{2\sigma^2}(-2X^Ty + 2X^TX\beta)$$

Again, differentiate,

$$\frac{\partial^2 \log \mathcal{L}(\beta, \sigma^2 I_n)}{\partial \beta^2} = -\frac{X^TX}{\sigma^2}$$

By exercise 6,

$$I(\beta) = -\mathbb{E}[\nabla^2_\theta \log f_\theta(X)] = -\mathbb{E}[-\frac{X^TX}{\sigma^2}]$$

But X's and σ^2 is assumed to be known,

$$I(\beta) = \frac{X^TX}{\sigma^2}$$

The Cramer-Rao Lower bound is the inverse so the bound is,

$$I^{-1}(\beta) = \sigma^2(X^TX)^{-1}$$

We now show that the variance of $\widehat{\beta}$ achieves this.

$$\begin{aligned}
Var(\widehat{\beta}) &= \mathbb{E}[(\widehat{\beta} - \mathbb{E}[\widehat{\beta}])(\widehat{\beta} - \mathbb{E}[\widehat{\beta}])^T] \\
&= \mathbb{E}[(\widehat{\beta} - \beta)(\widehat{\beta} - \beta)^T] \\
&= \mathbb{E}[((X^TX)^{-1}X^Ty - \beta)((X^TX)^{-1}X^Ty - \beta)^T] \\
&= \mathbb{E}\big[((X^TX)^{-1}X^T(X\beta + \eta) - \beta)((X^TX)^{-1}X^T(X\beta + \eta) - \beta) - \beta)^T\big] \\
&= \mathbb{E}[((X^TX)^{-1}X^T\eta)((X^TX)^{-1}X^T\eta)^T] = (X^TX)^{-1}X^T\mathbb{E}[\eta\eta^T]X(X^TX)^{-1} \\
&= (X^TX)^{-1}X^T\sigma^2 I_n X(X^TX)^{-1} = \sigma^2(X^TX)^{-1}X^TX(X^TX)^{-1} = \sigma^2(X^TX)^{-1}
\end{aligned}$$

Which matches the Cramer-Rao lower bound so we have that $\widehat{\beta}$ is an UMVUE of β.

2) We have the decomposition of MSE and the fact that UMVUE is unbiased,

$$R_2(\widehat{\beta}) = tr(Var(\widehat{\beta})) + bias^2(\widehat{\beta}) = tr(Var(\widehat{\beta})) = \sigma^2 tr((X^TX)^{-1})$$

So if we have,

$$I_n = \frac{X^TX}{n} \iff nI_n = X^TX \iff \frac{1}{n}I_n(X^TX)^{-1}$$

So the mean squared error becomes,

$$R_2(\widehat{\beta}) = \sigma^2 tr(\frac{1}{n}I_n) = n \cdot \frac{\sigma^2}{n} = \sigma^2$$

Exercise 9: In a Bayesian setting where the prior distribution θ is π. The Bayesian risk is defined as $\bar{R}_{\mathcal{L}}(\widehat{\theta}) = \mathbb{E}_{\theta}[R_{\mathcal{L}}(\widehat{\theta}, \theta)]$. Recall that the risk function $R_{\mathcal{L}}(\widehat{\theta}, \theta)$ is defined as $R_{\mathcal{L}}(\widehat{\theta}, \theta) = \mathbb{E}_{X|\theta}[\mathcal{L}(\widehat{\theta}(X), \theta)]$. We assume that all integrals exist. Answer the following questions.

1) Let X be a continuous random variable with distribution P_X. Show that $\mathbb{E}[|X - c|]$ is minimized at $c = \text{Median}(P_X)$, where $\text{Median}(P_X)$ is the median of P_X. Assume for simplicity that the cdf is strictly increasing, in which case the median is unique and is just $F_X^{-1}(1/2)$.

2) Let θ, X be jointly continuous, where $\theta \in \mathbb{R}$, and $X \in \mathbb{R}^d$. Find a minimizer $\widehat{\theta}$ for $\bar{R}_{\mathcal{L}}(\widehat{\theta})$ where
$$\mathcal{L}(\widehat{\theta}(X), \theta) = |\widehat{\theta}(X) - \theta|$$

3) Let $\theta \in \Theta$ and $X \in \mathbb{R}^d$, where $\Theta = \{1, 2, \ldots, K\}$. Find a minimizer $\widehat{\theta}$ for $\bar{R}_{\mathcal{L}}(\widehat{\theta})$ where
$$\mathcal{L}(\widehat{\theta}(X), \theta) = 1_{\{\widehat{\theta}(X) \neq \theta\}}$$

4) Let $\theta \in \mathbb{R}^k$, and $X \in \mathbb{R}^d$. Find a minimizer $\widehat{\theta}$ for $\bar{R}_{\mathcal{L}}(\widehat{\theta})$ where $\mathcal{L}(\widehat{\theta}(X), \theta) = \|\widehat{\theta}(X) - \theta\|_2^2$

Answer:

1) We write the definition of $\mathbb{E}[|X - c|]$,
$$\begin{aligned}
\mathbb{E}[|X - c|] &= \int_{X-c<0} c - X dP + \int_{X-c\geq 0} X - c dP \\
&= \int_{-\infty}^{c} c - X dP + \int_{c}^{\infty} X - c dP \\
&= c\int_{-\infty}^{c} dP - \int_{-\infty}^{c} X dP + \int_{c}^{\infty} X dP - c\int_{c}^{\infty} dP \\
&= cF(c) - c(1 - F(c)) - \int_{-\infty}^{c} X dP + \int_{c}^{\infty} X dP
\end{aligned}$$
Now we minimize this expression by differentiating by c and setting to 0,
$$\frac{\partial}{\partial c}\left(cF(c) - c(1 - F(c)) - \int_{-\infty}^{c} X dP + \int_{c}^{\infty} X dP\right)$$
$$= F(c) + cP_X(c) - 1 + F(c) + cP_X(c) - cP_X(c) - cP_X(c) = 2F(c) - 1$$

So, $F(c) = 1/2$ minimizes this since the second derivative is postive ($P_X(c) > 0$). So we conclude that $c = F_X^{-1}(1/2)$.

2) Now, $\mathcal{L}(\widehat{\theta}(X), \theta) = |\widehat{\theta}(X) - \theta|$ and,

$$\begin{aligned} R_{\mathcal{L}}(\widehat{\theta}, \theta) &= \mathbb{E}_{X|\theta}[\mathcal{L}(\widehat{\theta}(X), \theta)] \\ &= \int_{\mathbb{R}^d} \mathcal{L}(\widehat{\theta}(X), \theta) p(x|\theta) dx \end{aligned}$$

Where $p(x|\theta)$ is defined to be $\frac{\pi(\theta|x)p(x)}{\pi(\theta)}$. So,

$$\begin{aligned} \bar{R}_{\mathcal{L}}(\widehat{\theta}) &= \mathbb{E}_{\theta}[R_{\mathcal{L}}(\widehat{\theta}, \theta)] \\ &= \int_{\mathbb{R}} \int_{\mathbb{R}^d} \mathcal{L}(\widehat{\theta}(X), \theta) \frac{\pi(\theta|x)p(x)}{\pi(\theta)} \pi(\theta) dx d\theta \\ = \int_{\mathbb{R}^d} p(x) \int_{\mathbb{R}} \mathcal{L}(\widehat{\theta}(X), \theta) \pi(\theta|x) d\theta dx &= \int_{\mathbb{R}^d} p(x) \int_{\mathbb{R}} |\widehat{\theta} - \theta| \pi(\theta|x) d\theta dx \end{aligned}$$

We drop the outside integral since we'll minimize the inside for each x. Remove the absolute value by breaking it into two integrals,

$$= \int_{-\infty}^{\widehat{\theta}} (\widehat{\theta} - \theta) \pi(\theta|x) d\theta + \int_{\widehat{\theta}}^{\infty} (\theta - \widehat{\theta}) \pi(\theta|x) d\theta$$

Differentiate w.r.t. $\widehat{\theta}$ and set to 0,

$$\implies \int_{-\infty}^{\widehat{\theta}} \pi(\theta|x) d\theta = \int_{\widehat{\theta}}^{\infty} \pi(\theta|x) d\theta$$

That is, $\widehat{\theta}$ is the median for the posterior distribution.

3) Repeating the same steps as the previous example, we get to the point where we wish to minimize the following integral for all x,

$$\begin{aligned} &= \int_{\Theta} 1_{\{\widehat{\theta}(X) \neq \theta\}} \pi(\theta|x) d\theta \\ &= \int_{\Theta} (1 - 1_{\{\widehat{\theta}(X) = \theta\}}) \pi(\theta|x) d\theta \\ &= \int_{\Theta} \pi(\theta|x) d\theta - \int_{\Theta} 1_{\{\widehat{\theta}(X) = \theta\}} \pi(\theta|x) d\theta \end{aligned}$$

By defition of probabilities,

$$= 1 - \pi(\widehat{\theta}(X) = \theta|x)$$

Since the second term is negative, if we minimize the whole expression, we wish to maximize the second term. Thus, we choose $\widehat{\theta}$ that maximizes the posterior distribution, in other words, $\widehat{\theta}$ is the MAP.

4) Finally, we consider $\mathcal{L}(\widehat{\theta}(X), \theta) = \|\widehat{\theta}(X) - \theta\|_2^2$. We decompose it as follows,

$$\bar{R}_{\mathcal{L}}(\widehat{\theta}) = \mathbb{E}_{X|\theta}[\|\widehat{\theta}(X) - \theta\|_2^2] = Var(\theta|X) + \|\widehat{\theta} - \mathbb{E}_{\theta|X}[\theta]\|_2^2$$

Since the first term has no dependence on $\widehat{\theta}$, we can just modify the second term. We can make the second term minimal by setting $\widehat{\theta} = \mathbb{E}_{\theta|X}[\theta]$. That is $\widehat{\theta}$ is the posterior mean.

Exercise 10: Let $\mathcal{P} = \{\mathbb{P}_\theta = Uniform[0,\theta], \theta > 0\}$, that is, $\mathcal{P}$ is the family of uniform distributions. Let $X = \{X_i\}_{i=1}^n$ be n i.i.d. realizations of some $\mathbb{P}_\theta \in \mathcal{P}$. Show that

$$T(X) = \max_{1 \le i \le n} X_i$$

is both sufficient and complete. Hint: consider $\frac{d}{d\theta}\mathbb{E}[h(T)]$ for some Borel measurable function h.

Answer: Last homework, we derived,

$$f(x|\theta) = \frac{nx^{n-1}}{\theta^n} \cdot 1_{\{x \in [0,\theta]\}}$$

It is easy to see from this that this can be factorized into $g(\theta, T(X)) \cdot h(X)$. So it is sufficient . Recall that if T is complete, then if $\mathbb{E}[h(T)] = 0$ then $h(\cdot) = 0$. We now use the hint and consider $\frac{d}{d\theta}\mathbb{E}[h(T)]$ for some Borel measurable function h. Furhtermore, assume $\mathbb{E}[h(T)] = 0$ and so $\frac{d}{d\theta}\mathbb{E}[h(T)] = 0$. We then differentiate according to the Lebesgue differentiation theorem,

$$\begin{aligned}
\frac{d}{d\theta}\mathbb{E}[h(T)] = \frac{d}{d\theta}\int_0^\theta h(t)\frac{nx^{n-1}}{\theta^n}dx = - \\
= \frac{d}{d\theta}\int_0^\theta h(t)nx^{n-1}dx = 0 \\
= \frac{1}{\theta^n}nh(\theta)\theta^{n-1} = 0 \\
\implies h(\theta) = 0 \implies h(T) = 0 \text{ a.s.}
\end{aligned}$$

Thus, since $h(T) = 0$ when $\mathbb{E}[h(T)] = 0$, we must have that T is complete

Exercise 11: An ancillary statistic S for a family $\mathcal{P} = \{\mathbb{P}\}$ is one that has no information on $\mathbb{P}$. That is, the distribution of $S(X)$, when $X \sim \mathbb{P}$, is the same for all $\mathbb{P} \in \mathcal{P}$. Show that if T is complete and sufficient for $\mathcal{P}$, then T and any ancillary statistic S are uncorrelated. Assume that both S and T are in $\mathbb{R}^d$. Note: As an example of ancillary statistic we consider

$$\mathcal{P} = \{N(\mu, \sigma^2), \mu \in \mathbb{R}, \sigma^2 \text{ fixed and known}\}$$

Then statistic

$$T(X) = \frac{n - 1S^2_{n-1}}{\sigma^2} = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{\sigma^2} \sim \chi^2_{n-1}$$

which is an ancillary statistic. Hint: Consider any statistic S' that is an unbiased estimator of zero vector, i.e., $\mathbb{E}[S'] = 0 \in \mathbb{R}^d$ for any $\mathbb{P} \in \mathcal{P}$. First show that $Cov(S', T) = \mathbb{E}[(S' - [S'])(T - \mathbb{T})^T] = 0$. Now noticing that $\mathbb{E}[S]$ is a constant for any $\mathbb{P} \in \mathcal{P}$, conclude that $Cov(S, T) = 0$.

Answer: Consider an ancillary statistic such that $\mathbb{E}[S] = c$. We also have by completeness of T that if $\mathbb{E}[f(T)] = c$ then $f(T) = c$ a.s. for all $\mathbb{P} \in \mathcal{P}$. We now compute $Cov(S, T)$,

$$Cov(S, T) = \mathbb{E}[ST] - \mathbb{E}[S]\mathbb{E}[T]$$

Now,

$$\mathbb{E}[ST] = \mathbb{E}[\mathbb{E}[ST|T]] = \mathbb{E}[T\mathbb{E}[S|T]]$$

Note that $\mathbb{E}[S|T]$ is a measurable function of T and so by completeness of T, if $\mathbb{E}[\mathbb{E}[S|T]] = c$ then $\mathbb{E}[S|T] = c$. We have that $\mathbb{E}[\mathbb{E}[S|T]] = \mathbb{E}[S] = c$. So $\mathbb{E}[S|T] = c$. We sub this back in,

$$\mathbb{E}[ST] = \mathbb{E}[Tc] = c\mathbb{E}[T]$$

So,

$$Cov(S, T) = c\mathbb{E}[T] - \mathbb{E}[S]\mathbb{E}[T] = c\mathbb{E}[T] - c\mathbb{E}[T] = 0$$

So S and T are uncorrelated.