ORF524 - Problem Set 3

Bachir EL KHADIR

October 30, 2015

Question 1

1.

$$
\begin{aligned}
\mathcal{L}\lambda a + (1-\lambda)b &= ||\lambda(a-\theta) + (1-\lambda)(b-\theta)||_p && (1)\\
&= ||\lambda(a-\theta)||_p + ||(1-\lambda)(b-\theta)||_p && \text{By Minkowsky} \quad (2)\\
&= \lambda\mathcal{L}(a) + (1-\lambda)(b) && (3)
\end{aligned}
$$

2. for $q > 1$, let's denote $f_q : x \to x^q$ for $x > 0$. f_q is convexe and non decreasing because $f_q'(x) = qx^{q-1} > 0$ and $f_q''(x) = q(q-1)x^{q-2} > 0$.

$$
\begin{aligned}
\mathcal{L}\lambda a + (1-\lambda)b &= f_q(||\lambda(a-\theta) + (1-\lambda)(b-\theta)||_p) && (4)\\
&\le f_q(\lambda||(a-\theta)|| + (1-\lambda)||(b-\theta)||_p) && \text{because } f_q \text{ non decreasing} \quad (5)\\
&\le \lambda f_q(||(a-\theta)||) + (1-\lambda)f_q(||(b-\theta)||_p) && \text{because } f_q \text{ convexe} \quad (6)
\end{aligned}
$$

Question 2

1. The X_i have the same distribution and play symetric roles, so:

$$\tilde{p} = E[\hat{p}|T(X)] = E[X_1|\sum_i X_i] = \frac{1}{n}E[\sum X_i|\sum_i X_i] = \frac{T(X)}{n}$$

2.

$$E[(\hat{p}-p)^2] = \text{Var}(X_1) = p(1-p)$$

the X_i being iid:

$$E[(\tilde{p}-p)^2] = E[(\sum \frac{T(X_i)-p}{n})^2] = \text{Var}(\sum \frac{X_i - p}{n}) = \sum_i \frac{Var(X_i)}{n^2} = \frac{Var(X_1)}{n} = \frac{p(1-p)}{n}$$

$\hat{p}$ has better variance than $\tilde{p}$, because it uses the information from all the X_i.

Question 3

-

$$Var(E[X|Y]) = E[(E[X|Y] - E[E[X|Y]])^2] = E[(E[X|Y] - E[X])^2]$$

$$E[Var(X|Y)] = E[E[(X - E[X|Y])^2|Y]] = E[(X - E[X|Y])^2]$$

By summing:

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

because

$$\begin{aligned}E[(X - E[X|Y])(E[X] - E[X|Y])] &= E[E[XE[X] - E[X|Y](E[X] + X) + E[X|Y]^2|Y]] \quad (7)\\ &= E[X]^2 - E[X]^2 + E[E[X|Y]]^2 + E[E[X|Y]^2] \quad (8)\\ &= 0 \quad (9)\end{aligned}$$

- Let $g(X)$ be an estimator of θ and $T(X)$ a sufficient statistics. The bias of $g(X)$ and $g(X)|T(X)$ are the same because of the law of iterated expectation. We can assume that the bias is 0 without loss of generality by substracting it from both variables.

  Then

  $E[(E[g(X)|T(X)] - \theta)^2] = E[(E[g(X) - \theta|T(X)])^2] = E[Var(g(X)|T(X))] \leq Var(g(X))$ (because $Var \geq 0$) ie:

$$E[(E[g(X)|T(X)] - \theta)^2] \leq E[(g(X) - \theta)^2]$$

$$Var[E[g(X)|T(X)]] \leq Var[g(X)]$$

Question 4

Let's prove that $\phi(\{c_j\}^l$ is non-increasing at each step.

By assigning each x_i to the nearest $c_j'^{l+1}$, each quantity $||x_i - c_j^l||^2$ in the sum above is replaced by a smaller (or equal) quantity $||x_i - c_j'^{l+1}||^2$. The new clusters are C_j^{l+1}

For every $j = 1..K$, $\sum_{x_i \in C_j^{l+1}} ||x_i - c_j'^{l+1}||^2 \leq \sum_{x_i \in C_j^{l+1}} ||x_i - c_j^{l+1}||^2$ because the mean of the point $x_i \in C_j^l$, minimizes the quantity $\mu \to \sum_{x_i \in C_j^l} ||x_i - \mu||^2$. (By taking the the first and the second derivative, the function being quadratic)

So: $\phi(\{c_j\}^l) = \sum_j \sum_{x_i \in C_j^l} ||x_i - c_j^l||^2 \leq \sum_j \sum_{x_i \in C_j^{l+1}} ||x_i - c_j^{l+1}||^2 = \phi(\{c_j\}^{l+1})$

Therefore $\phi(\{c_j\}^l$ is non negative non-increasing, and the limit exsits.

Question 5

1.

$$Cov((T, a)^T) = \mathbb{E}(T, a)(T, a)^T - (\mathbb{E}(T, a))(\mathbb{E}(T, a))^T$$

$$\begin{aligned}&= \mathbb{E}\begin{bmatrix} TT^T & T^T a \\ aT^T & aa^T \end{bmatrix} - \begin{bmatrix} \mathbb{E}(T)\mathbb{E}(T)^T & \mathbb{E}(T)^T\mathbb{E}(a) \\ \mathbb{E}(a)\mathbb{E}(T)^T & \mathbb{E}(a)\mathbb{E}(a)^T \end{bmatrix} \\ &= \begin{bmatrix} Cov(T) & Cov(T, a) \\ Cov(a, T) & Cov(a) \end{bmatrix} \\ &= \begin{bmatrix} Cov(T) & \nabla_\theta g(\theta) \\ \nabla_\theta g(\theta)^T & I(\theta) \end{bmatrix}\end{aligned}$$

Because:

- $\mathbb{E}(a) = \int \nabla_\theta \log f_\theta(x) f_\theta(x) \mathrm{dx} = \int \frac{\nabla_\theta \mathrm{f}_\theta(\mathrm{x})}{\mathrm{f}_\theta(\mathrm{x})} \mathrm{f}_\theta(\mathrm{x}) \mathrm{dx} = \nabla_\theta 1 = 0$
- $Cov(a) = \mathbb{E}(aa^T) = I(\theta)$
- $$\begin{aligned} Cov(T, a) &= \mathbb{E}(T^T a) \\ &= \int T(x) \nabla_\theta \log f_\theta(x) f_\theta(x) \mathrm{dx} \\ &= \int T(x) \frac{\nabla_\theta f_\theta(x)}{f_\theta(x)} f_\theta(x) \mathrm{dx} \\ &= \nabla_\theta \int T(x) f_\theta(x) \mathrm{dx} && \text{(By regularity condition)} \\ &= \nabla_\theta g(\theta) \end{aligned}$$

2. $$B = \begin{pmatrix} -I_p & , \nabla_\theta g(\theta)^T I(\theta)^{-1} \end{pmatrix}^T$$

$$\begin{aligned} B^T Cov(T, a)^T B &= Cov(T) - \nabla_\theta g(\theta)^T I(\theta)^{-1} \nabla_\theta g(\theta) - \nabla_\theta g(\theta)^T I(\theta)^{-1} \nabla_\theta g(\theta) && (10) \\ &\quad + (\nabla_\theta g(\theta)^T I(\theta)^{-1}) I(\theta) (I(\theta)^{-1} \nabla_\theta g(\theta)) && (11) \\ &= Cov(T) - \nabla_\theta g(\theta)^T I(\theta)^{-1} \nabla_\theta g(\theta) && (12) \end{aligned}$$

3. $Cov(T) - \nabla_\theta g(\theta) I(\theta) \nabla_\theta g(\theta) = B^T Cov(T, a)^T B = Cov(B(T, a)^T) \geq 0$

Question 6

In the following we write f instead of $f_\theta(x)$ or $f_\theta(X)$.

$$\nabla_\theta^2 \log f = \nabla_\theta (\frac{\nabla_\theta f}{f}) = \frac{\nabla_\theta^2 f}{f} - \frac{\nabla_\theta f \nabla_\theta f^T}{f^2} = \frac{\nabla_\theta^2 f}{f} - \nabla_\theta \log f \nabla_\theta \log f^T$$

But $\mathbb{E}(\frac{\nabla_\theta^2 f}{f}) = \int \frac{\nabla_\theta^2 f}{f} f \mathrm{dx} = \nabla_\theta^2 \int \mathrm{fdx} = 0$, so

$$I(\theta) = \mathbb{E}(\nabla_\theta f_\theta(x) \nabla_\theta f_\theta(x)^T) = -\mathbb{E}(\nabla_\theta^2 f)$$

Question 7

1. By the series expansion of exponential:

$$|\frac{e^{az} - 1}{z}| = |\sum_{k=1}^{\infty} \frac{a^k z^{k-1}}{k!}| \leq \sum_{k=1}^{\infty} \frac{|a|^k |z|^{k-1}}{k!} \leq \sum_{k=1}^{\infty} \frac{1}{\delta} \frac{|a|^k |\delta|^k}{k!} = \frac{e^{|a\delta|}}{|\delta|}$$

2. Let $|\delta| \leq \varepsilon$, so that $\alpha \pm \varepsilon \in \mathcal{A}$.

   Because $\forall \alpha \int f_\alpha = 1$, we have that:

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

   But

$$|h(x) \frac{e^{(\alpha+\delta)T(x)} - e^{\alpha T(x)}}{\delta}| \leq h(x) e^{\alpha T(x)} \frac{e^{|\delta||T(x)|}}{\delta} \leq \max(f_{\alpha-\varepsilon}(x), f_{\alpha+\varepsilon}(x)) \in L_1$$

By dominated convergence theorem: $\lim_0 \frac{l(\alpha+\delta)-l(\alpha)}{\delta}$ exists and is equal to $\int h(x)e^{\alpha T(x)}T(x)dx$

As a result, l, and therefore $\alpha \to f_\alpha(x)$ are differentiable and $E[T] = l(\alpha)\int h(x)e^{\alpha T(x)}T(x)dx < \infty$.

We have that

- l is differentiable, so $\frac{l'}{l}$ is continuous and bounded by a constant M on $[\alpha-\epsilon, \alpha+\epsilon]$.
-
$$|f_{\alpha+\delta}(x)| < \max(f_{\alpha+\varepsilon}(x), f_{\alpha-\varepsilon}(x))$$

So:

$$\begin{aligned}
|\frac{df_{\alpha+\delta}(x)}{d\alpha}| &= |f_{\alpha+\delta}(x)(T(x)+\frac{l'(\alpha+\delta)}{l(\alpha+\delta)})| && (13)\\
&\leq \max(f_{\alpha+\varepsilon}(x)|T(x)|, f_{\alpha-\varepsilon}(x)|T(x)|) + M\max(f_{\alpha+\varepsilon}(x), f_{\alpha-\varepsilon}(x)) := H(x) \in L_1 && (14)
\end{aligned}$$

And since $\forall |\delta| < \varepsilon E^{\alpha+\delta}[g] < \infty$, e $|g(x)\frac{df_{\alpha+\delta}(x)}{d\alpha}|$ is bounded uniformly by an integrable function $g(x)H(x)$.

By dominated convergence theorem, we have the result to the question.

Question 8

1.
$$\hat{\beta} = (X^TX)^{-1}X^Ty = (X^TX)^{-1}X^TX\beta + (X^TX)^{-1}X^T\eta = \beta + (X^TX)^{-1}X^T\eta$$

   $\mathbb{E}(\hat{\beta}) = \beta + (X^TX)^{-1}X^T\mathbb{E}(\eta) = \beta$. So $\hat{\beta}$ is unbiased.

   The family of distributions for y ($\{f_\beta(y) = \text{cte } e^{||y||^2+||X\beta||^2-2(X^Ty)^T\beta}\}$) is a full exponential family for which $T(y) = X^Ty$. $\hat{\beta}$ begin linear in $T(y)$ and unbiased, it is an UMVUE.

2. $R_2(\hat{\beta}) = Var(\hat{\beta}) = Var((X^TX)^{-1}X^T\eta) = (X^TX)^{-1}X^TVar(\eta)((X^TX)^{-1}X^T)^T = \sigma^2(X^TX)^{-1}X^TX(X^TX)^{-1}$ $\sigma^2(X^TX)^{-1}$

   If $X^TX = I_p$, $R_2(\hat{\beta}) = \sigma^2 I_p$

Question 9

1.
$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{dc}}\mathbb{E}(|X-c|) &= \frac{\mathrm{d}}{\mathrm{dc}}\int^c (c-x)f(x)\mathrm{dx} + \int_{\mathrm{c}} (\mathrm{x}-\mathrm{c})\mathrm{f}(\mathrm{x})\mathrm{dx}\\
&= \frac{\mathrm{d}}{\mathrm{dc}}c(F(c)-(1-F(c)) + \int^c -xf(x)\mathrm{dx} + \int_{\mathrm{c}} \mathrm{xf}(\mathrm{x})\mathrm{dx}\\
&= \frac{\mathrm{d}}{\mathrm{dc}}c(2F(c)-1) - 2\int^c xf(x)\mathrm{dx} + \int_{\mathbb{R}} \mathrm{xf}(\mathrm{x})\mathrm{dx}\\
&= 2F(c) - 1 + 2cf(c) - 2cf(c)\\
&= 2F(c) - 1
\end{aligned}$$

   (To justify the existence of the derivative, one can start from the bottom to top)

   The derivative is increasing so the function is strictly convexe, and therefore it attains its minimum when the derivative is 0, or $F(c) = \frac{1}{2}$, or $c = \text{median}(P_X)$

2. By Fubini-Tonelli

$$\bar{\mathcal{L}}_{\mathcal{R}}(\hat{\theta}) = E_X[E_{\theta|X}[|\hat{\theta}(X) - \theta|]]$$

By the previous question, taking $\hat{\theta}(X) = \operatorname{median} F_{\theta|X} = F_{\theta|X}^{-1}(\frac{1}{2})$ minimizes the quantity inside the expectation pointwise, and therefore in average.

3. For Y a discrete variable, the c that minimizes $E[1_{Y \neq c}]$ is $c = \arg\max_y P(Y = y)$

$$\bar{\mathcal{L}}_{\mathcal{R}}(\hat{\theta}) = E_X[E_{\theta|X}[1_{\hat{\theta}(X) \neq \theta}]]$$

$$\hat{\theta} = \arg\max_{k=1 \ldots K} P(\theta = k|X)$$

4. The c that minimizes $E[||Y - c||_2^2]$ is $c = E[Y]$, using the same justification as above:

$$\hat{\theta} = E[\theta|X]$$