**36-752, Spring 2018**
**Homework 3**

Due Thu March 22, by 5:00pm in Jisu's mailbox.

1. Assume $X$ and $Y$ are integrable random variables. Prove that, for each $r > 0$,
$$\mathbb{E}|X+Y|^r \leq C_r \left(\mathbb{E}|X|^r + \mathbb{E}|Y|^r\right),$$
where $C_r = 1$ if $r \in (0,1]$ and $C_r = 2^{r-1}$ for $r > 1$.
*Hint: for $r > 1$ use Jensen's inequality. For $r \in (0,1]$ use the fact that $(1+x)^r \leq 1+x^r$ for $x \geq 0$.*

2. Prove the following generalization of Hölder inequality. Let $p_1, \ldots, p_k$ positive number such that $\sum_{i=1}^k \frac{1}{p_i} = 1$ and let $X_1, \ldots, X_k$ random variables such that $\|X_i\|_{p_i} < \infty$ for all $i$. Then,
$$\mathbb{E}\left[\left|\prod_{i=1}^k X_i\right|\right] \leq \prod_{i=1}^k \|X_i\|_{p_i}.$$
*Hint: apply the standard version of Hölder's inequality recursively.*

3. Prove Paley-Zygmund's inequality: let $X$ be a non-negative random variable with finite variance. Then, for ay $\lambda > 0$,
$$\mathbb{P}(X \geq \lambda) \geq \frac{\left[(\mathbb{E}[X]-\lambda)+\right]^2}{\mathbb{E}[X^2]}.$$
If $X$ is non-negative and bounded – that is, $0 \leq X \leq b$ almost surely for some $b > 0$ – prove that, for all $\lambda \in (0, \mathbb{E}[X])$,
$$\mathbb{P}(X \geq \lambda) \geq \frac{\mathbb{E}[X]-\lambda}{b-\lambda}.$$

4. Let $X_1, \ldots, X_k \overset{i.i.d}{\sim} \text{Uniform}(0,\theta)$, for some $\theta > 0$. Show that $T = \max_i X_i$ is a sufficient statistic for $\theta$ by proving that the conditional distribution of the $X_i$'s given $T$ is independent of $\theta$. In this case $\sigma(T)$ is referred to as the sufficient $\sigma$-field.[1]

5. Let $X$ and $Y$ be random variables over the probability space $(\Omega, \mathcal{F}, P)$. Assume that the range of $Y$ is a countable subset $\mathcal{Y}$ of $\mathbb{R}$ such that $P\left(Y^{-1}(\{y\})\right) > 0$ for all $y \in \mathcal{Y}$. Show that the conditional expectation of $X$ given $Y$ is the random variable $g(Y)$, where the function $g\colon \mathbb{R} \to \mathbb{R}$ is given by
$$y \mapsto \frac{1}{P\left(Y^{-1}(\{y\})\right)} \int_{Y^{-1}(\{y\})} X dP.$$

[1] There is much more that could be said about sufficiency from the measure theoretic standpoint, including a nice derivation of the Fisher-Neyman factorization theorem. For more details, see Billingsley (1995), Probability and Measure, Wiley, page 450.

In particular, if $Y = 1_A$ for some $A \in \mathcal{F}$ we may speak of the conditional expectation of $X$ given $A$ when referring to $\mathbb{E}[X|Y]$. This is what "conditioning on an event" means.[2] (Special thanks to Matteo and Pratik for suggesting the problem...).

6. If $X$ and $Y$ are independent random variables with finite expectations on a common probability space $(\Omega, \mathcal{F}, P)$, show that $\mathbb{E}(X|Y) = \mathbb{E}[X]$, a.e. $[P]$.
   *This can be proved in many ways, some simpler than others. You should try to provide a measure-theoretic proof of the following, more general result: if $\mathcal{C}$ and $\sigma(X)$ are independent $\sigma$-fields contained in $\mathcal{F}$, then $\mathbb{E}[X|\mathcal{C}] = \mathbb{E}[X]$, a.e. $[P]$.*

7. Let $X$ be a random variable on $(\Omega, \mathcal{F}, P)$ and $\mathcal{C} \subset \mathcal{F}$ a $\sigma$-field. Show that, for each $p \geq 1$,
   $$\mathbb{E}\left[|\mathbb{E}[X|\mathcal{C}]|^p\right] \leq \mathbb{E}|X|^p.$$
   That is, the condition expectation is a contraction on the $L_p$ space of random variables on $(\Omega, \mathcal{F}, P)$ with finite $p$-th moment. In particular, show that the variance of $\mathbb{E}[X|\mathcal{C}]$ is smaller than the variance of $X$. This is a way of formalizing the intuition that conditioning (which can be thought of as extra information) reduces uncertainty.

8. **Exponential families**.
   Below, for two vectors $x = (x_1, \ldots, x_k)$ and $y = (y_1, \ldots, y_k)$ in $\mathbb{R}^k$, we let $x \cdot y$ denote their inner product $\sum_{i=1}^k x_i y_i$. Let $\mu$ be a $\sigma$-finite measure on $(\mathbb{R}^k, \mathcal{B}^k)$ and let
   $$\Theta = \{\theta \in \mathbb{R}^k \colon \int_{\mathbb{R}^k} e^{x \cdot \theta} d\mu(x) < \infty\}.$$
   For any $\theta \in \Theta$, let
   $$\psi(\theta) = \log\left(\int_{\mathbb{R}^k} e^{x \cdot \theta} d\mu(x)\right).$$
   The function $\psi$ is know as the log-partition function. For each $\theta \in \Theta$, define the non-negative function
   $$p_\theta(x) = \exp\left(x \cdot \theta - \log \psi(\theta)\right), \quad \forall x \in \mathbb{R}^k. \tag{1}$$
   Notice that, for each $\theta \in \Theta$, $\int_{\mathbb{R}^k} p_\theta(x) d\mu(x) = 1$ (this is because the exponential of the log-partition function serves as a normalizing constant), so that we can define the family $\mathcal{P} = \{P_\theta, \theta \in \Theta\}$ of probability measures on $(\mathbb{R}^k, \mathcal{B}^k)$, each of the form
   $$P_\theta(A) = \int_A p_\theta(x) d\mu(x), \quad \forall A \in \mathcal{B}^k.$$

[2] Ale's rant: in many theoretical papers you will see the following mis-use of the expression. In proving that a certain property holds, a general strategy is to define a high-probability good event and to show that the desired property always holds in that event. Way too often the authors will then say that *"...conditionally on this good event, the claimed result follows."* In fact, there is no conditioning at all! The argument is instead as follows: let $R$ the event that the result holds and $G$ the good event. Then if $G \subseteq R$ and $P(G)$ is large, we must have that the probability $P(R^c)$ that the result fails is small, smaller than $P(G^c)$. As you can see, we have not conditioned on any event.

In particular, since by construction $P_\theta << \mu$ for all $\theta$, we have that $p_\theta = \frac{dP_\theta}{d\mu}$.

The family $\mathcal{P}$ is known as a *$k$-dimensional standard exponential family* of probability distributions. These are the well-behaved type of distributions, with many interesting properties. Below you will derive some of them.

(a) Prove that all the probability measures in $\mathcal{P}$ are equivalent and have the same support (the support of a probability distribution $P$ on $(\mathbb{R}^k, \mathcal{B}^k)$ is the smallest closed set $S$ such that $P(S) = 1$; if $P$ has a density $p$ with respect to some $\sigma$-finite measure, then $S$ is $\mathrm{cl}(\{x : p(x) > 0\})$, the closure of all points of positive density).

(b) Prove that $\psi$ is a convex function on $\Theta$ and that $\Theta$ is a convex set. *Hint: use Hölder inequality.*

(c) Prove that $P_{\theta_1} = P_{\theta_2}$ if and only if, for some $\alpha \in (0,1)$,

$$\psi(\alpha\theta_1 + (1-\alpha)\theta_2) = \alpha\psi(\theta_1) + (1-\alpha)\psi(\theta_2).$$

Notice that if $\psi$ is strictly convex this cannot happen.
Prove that this is equivalent to $(\theta_1 - \theta_2) \cdot x = K$, a.e. $[\mu]$, for some $K \in \mathbb{R}$. In turn this is equivalent to $\mu(H^c) = 0$ for some affine subspace of dimension $k-1$ .

(d) **Sufficiency.** A more common form of the exponential family is obtained by assuming that the parameter space $\Theta$ is a subset (typically open) of $\mathbb{R}^d$, where $d < k$. In this case, the density (w.r.t. $\mu$) of a point $x \in \mathbb{R}^k$ is usually expressed, for a given value of the parameter vector $\theta \in \mathbb{R}^d$, as

$$p_\theta(x) = \exp\left(\tau(x) \cdot \theta - \log\psi(\theta)\right), \tag{2}$$

where $\tau : \mathbb{R}^k \to \mathbb{R}^d$ is a given function. Notice that in this representation, we can parametrize distributions on $\mathbb{R}^k$ with very few parameters $d < k$.

Let $X$ be a random vector in $\mathbb{R}^k$ with density (2), for some $\theta \in \Theta \subset \mathbb{R}^k$. Let $T = \tau(X)$, a $d$-dimensional vector. Show that the distribution of $T$ is an exponential family on $(\mathbb{R}^k, \mathcal{B}^k)$ with the same natural parameter space $\Theta$ as the distribution of $X$ and densities of the form (1) with respect to a new $\sigma$-finite measure $\nu$ on $(\mathbb{R}^k, \mathcal{B}^k)$. (Find that measure, too!).
Assuming that the common support is finite and that the dominating measure is the counting measure, show that the conditional distribution of $X$ given $T = t$ is uniform over the set $\{x \in \mathbb{R}^k : \tau(x) = t\}$. Conclude that $\tau(X)$ is a sufficient statistic for $\theta$.

(e) **Conditionals and Marginals of Exponential Families.** For any $x$ in the domain of $\tau$, writw $\tau(x) = (t_1, t_2)$, where $t_1 \in \mathbb{R}^l$ and $t_2 = \mathbb{R}^{k-l}$, for some $l = 1, \ldots, k-1$. Similarly, for any $\theta \in \Theta \subset \mathbb{R}^k$, write $\theta = (\theta_1, \theta_2)$ with $\theta_1 \in \mathbb{R}^l$ and $\theta_2 = \mathbb{R}^{k-l}$. Then

$$\tau(x) \cdot \theta = t_1 \cdot \theta_1 + t_2 \cdot \theta_2.$$

i. Show that, for a given $\theta = (\theta_1, \theta_2)$ the conditional distribution of $T_1$ given $T_2 = t_2$ has a density of the exponential form (1) with respect to a $\sigma$-finite measure $\nu_{t_2}$ (which depends on $t_2$) and natural parameter $\theta_1$. Thus, conditioning on $T_2$ eliminates the dependence on $\theta_2$. Conclude that the conditional distribution of $T_1$ given $T_2 = t_2$ is an exponential family of dimension $l$ and with natural parameter space given by $\{\theta_1 \colon (\theta_1, \theta_2) \in \Theta\}$.

ii. On the other hand, show that the marginal distribution of $T_1$ has a density of the exponential form (1) with respect to a $\sigma$-finite measure $\nu_{\theta_2}$, which depends on $\theta_2$. Notice that the marginal distribution of $T_1$ still depends on $\theta_2$ (the fact that the dominating measure depends on $\theta_2$ further implies that the log-partition function depends on $\theta_2$). Conclude that (unless $\theta_2$ is fixed and known) the marginal distribution of $T_2$ is not an exponential family.

iii. The Erdös-Rényi model is a statistical model for networks (i.e. random graphs). According to this model, the $\binom{n}{2}$ edges in a network with $n$ nodes are independent Bernoulli's with common parameter $p \in (0, 1)$. Show that this model is a one-dimensional (i.e. $d = 1$) exponential family of probability distributions over the set $\mathcal{G}_n$ of simple undirected graphs. *Hint: the one dimensional sufficient statistic is the number of edges...*