36-752, Spring 2018
Homework 3 Solution

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

Points: 100 pts total for the assignment.

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 $x \geq 0$.

Points: 10 pts.

Solution.
For $r > 1$, note that $f(x) = x^r$ for $x \geq 0$ is a convex function, and hence

$$\left(\frac{|X| + |Y|}{2} \right)^r \leq \frac{1}{2} \left(|X|^r + |Y|^r \right).$$

Taking expectation yields

$$\mathbb{E}\left[(|X| + |Y|)^r \right] \leq 2^{r-1} \left(\mathbb{E}|X|^r + \mathbb{E}|Y|^r \right).$$

For $r \in (0, 1]$, note that when $X \neq 0$,

$$(|X| + |Y|)^r = |X|^r \left(1 + \frac{|Y|}{|X|} \right)^r \leq |X|^r \left(1 + \frac{|Y|^r}{|X|^r} \right) = |X|^r + |Y|^r,$$

and such inequality holds when $X = 0$ as well. Hence taking expectation yields

$$\mathbb{E}\left[(|X| + |Y|)^r \right] \leq \mathbb{E}|X|^r + \mathbb{E}|Y|^r.$$

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.

Points: 10 pts.

Solution.

We apply mathematical induction. First, $k \le 2$ comes from Hölder inequality. Now, suppose the induction inequality holds for $k = m$. When $k = m + 1$, define $Y_1, \ldots, Y_m$ and $q_1, \ldots, q_m$ as

$$Y_i = X_i,\ i \le m-1,\ Y_m = X_m X_{m+1},\ q_i = p_i,\ i \le m-1,\ q_m = \frac{p_m p_{m+1}}{p_m + p_{m+1}}.$$

Then $\sum_{i=1}^{m} \frac{1}{q_i} = \sum_{i=1}^{m+1} \frac{1}{p_i} = 1$ holds. Hence applying the induction inequality on Y_i and q_i yields

$$\mathbb{E}\left[\left|\prod_{i=1}^{m} Y_i\right|\right] \le \prod_{i=1}^{m} \|Y_i\|_{p_i}.$$

Then applying the relation of X_i, Y_i, p_i, q_i gives

$$\mathbb{E}\left[\left|\prod_{i=1}^{m+1} X_i\right|\right] \le \left(\prod_{i=1}^{m-1} \|X_i\|_{p_i}\right) \|X_m X_{m+1}\|_{q_m}.$$

Then from $\frac{q_m}{p_m} + \frac{q_m}{p_{m+1}} = 1$, applying Hölder inequality on $\|X_m X_{m+1}\|_{q_m}$ gives

$$\begin{aligned}
\|X_m X_{m+1}\|_{q_m} &= \left(\mathbb{E}\left[|X_m X_{m+1}|^{q_m}\right]\right)^{\frac{1}{q_m}} \\
&\le \left(\left(\mathbb{E}\left[|X_m|^{q_m \times \frac{p_m}{q_m}}\right]\right)^{\frac{q_m}{p_m}} \left(\mathbb{E}\left[|X_{m+1}|^{q_m \times \frac{p_{m+1}}{q_m}}\right]\right)^{\frac{q_m}{p_{m+1}}}\right)^{\frac{1}{q_m}} \\
&= \|X_m\|_{p_m} \|X_{m+1}\|_{p_{m+1}}.
\end{aligned}$$

Hence applying this gives

$$\mathbb{E}\left[\left|\prod_{i=1}^{m+1} X_i\right|\right] \le \prod_{i=1}^{m+1} \|X_i\|_{p_i}.$$

3. Prove Paley-Zygmund's inequality: let X be a non-negative random variable with finite variance. Then, for ay $\lambda > 0$,

$$\mathbb{P}(X \ge \lambda) \ge \frac{[(\mathbb{E}[X] - \lambda)+]^2}{\mathbb{E}[X^2]}.$$

If X is non-negative and bounded – that is, $0 \le X \le b$ almost surely for some $b > 0$ – prove that, for all $\lambda \in (0, \mathbb{E}[X])$,

$$\mathbb{P}(X \ge \lambda) \ge \frac{\mathbb{E}[X] - \lambda}{b - \lambda}.$$

Points: 10 pts.

Solution.

Note first that $f : \mathbb{R} \to \mathbb{R}$ by $f(x) = x_+ := \max\{x, 0\}$ is convex function. And hence

$$(\mathbb{E}[X] - \lambda)_+ = (\mathbb{E}[X - \lambda])_+ \leq \mathbb{E}\left[(X-\lambda)_+\right] = \mathbb{E}\left[(X-\lambda)_+ I(X \geq \lambda)\right].$$

Then applying Cauchy-Schwarz inequality gives a further bound as

$$\begin{aligned}(\mathbb{E}[X] - \lambda)_+ &\leq \mathbb{E}\left[(X-\lambda)_+ I(X \geq \lambda)\right] \\ &\leq \sqrt{\mathbb{E}\left[(X-\lambda)_+^2\right] \mathbb{E}\left[I^2(X \geq \lambda)\right]} \text{ (Cauchy-Schwarz)} \\ &= \sqrt{\mathbb{E}\left[(X-\lambda)_+^2\right] \mathbb{P}(X \geq \lambda)}.\end{aligned}$$

Hence by using $(x-\lambda)_+^2 \leq x^2$ for $\lambda \geq 0$, $\mathbb{P}(X \geq \lambda)$ can be lower bounded as

$$\mathbb{P}(X \geq \lambda) \geq \frac{(\mathbb{E}[X] - \lambda)_+^2}{\mathbb{E}\left[(X-\lambda)_+^2\right]} \geq \frac{(\mathbb{E}[X] - \lambda)_+^2}{\mathbb{E}[X^2]}.$$

Also, $\lambda \in (0, \mathbb{E}[X])$ implies $(\mathbb{E}[X] - \lambda)_+ = \mathbb{E}[X] - \lambda$ and $\lambda \leq \mathbb{E}[X]$. Hence $0 \leq X \leq b$ a.s. implies $\lambda \leq \mathbb{E}[X] \leq b$ and $0 \leq (X-\lambda)_+ \leq b - \lambda$ a.s.. Then $\mathbb{E}[X] - \lambda$ can be bounded as

$$\begin{aligned}\mathbb{E}[X] - \lambda = (\mathbb{E}[X] - \lambda)_+ &\leq \mathbb{E}\left[(X-\lambda)_+ I(X \geq \lambda)\right] \\ &\leq \mathbb{E}\left[(b-\lambda) I(X \geq \lambda)\right] = (b-\lambda)\mathbb{P}(X \geq \lambda).\end{aligned}$$

Hence $\mathbb{P}(X \geq \lambda)$ can be lower bounded as

$$\mathbb{P}(X \geq \lambda) \geq \frac{(\mathbb{E}[X] - \lambda)_+}{b - \lambda}.$$

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 θ by proving that the conditional distribution of the X_i's given T is independent of θ. In this case $\sigma(T)$ is referred to as the sufficient σ-field.[1]

   Points: 10 pts.

Solution.

Let $\mathbb{R}_+ := (0, \infty)$ and fix a set $B \in \mathcal{B}_{\mathbb{R}_+^k}$. Consider a Lebesgue measure λ_B on $(B, \mathcal{B}_B)$ and a map $m : B \to \mathbb{R}_+$ by $m(x_1, \ldots, x_k) = \max_{1 \leq i \leq k} x_i$. Now, cosider

[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.

an induced measure μ_B on $(\mathbb{R}_+, \mathcal{B}_{\mathbb{R}_+})$ as $\mu_B(A) = \lambda_B(m^{-1}(A))$ for all $A \in \mathcal{B}_{\mathbb{R}_+}$. Let λ_1 be the Lebesgue measure on $(\mathbb{R}_+, \mathcal{B}_{\mathbb{R}_+})$, then $\lambda_1(A) = 0$ implies

$$\begin{aligned}\mu_B(A) = \lambda_B(m^{-1}(A)) &\leq \lambda_{\mathbb{R}_+^k}((A \times \mathbb{R}_+ \times \cdots \times \mathbb{R}_+) \cup \cdots \cup (\mathbb{R}_+ \times \cdots \times \mathbb{R}_+ \times A)) \\ &\leq 0 \times \infty \times \cdots \times \infty + \cdots + \infty \times \cdots \times \infty \times 0 = 0,\end{aligned}$$

and hence $\mu_B \ll \lambda_1$. Also, note that

$$\mu_B((0,n)) = \lambda_B(m^{-1}(0,n)) = \lambda_B((0,n)^k) \leq \lambda_{\mathbb{R}_+^k}((0,n)^k) = n^k < \infty$$

and $\mathbb{R}_+ \subset \bigcup_{n \in \mathbb{N}}(0,n)$, and hence μ_B is σ-finite. Since λ_1 is σ-finite as well, there exist a Radon-Nikodym derivative $\frac{d\mu_B}{d\lambda_1}$.

Note that conditional distribution $\mu_{X|\sigma(T)}(\cdot)(\cdot) : \mathcal{B}_{\mathbb{R}_+^k} \times \Omega \to [0,1]$ is characterized by that for all $B \in \mathcal{B}_{\mathbb{R}_+^k}$, $\mu_{X|\sigma(T)}(B)$ is a version of $\mathbb{E}\left[1_{X \in B} | \sigma(T)\right]$, i.e. $\mu_{X|\sigma(T)}(B)(\cdot)$ is $\sigma(T)$-measurable and for all $A \in \sigma(T)$,

$$\int_A \mu_{X|\sigma(T)}(B)(\omega) dP = \int_A 1_{X \in B}(\omega) dP.$$

We will argue that

$$\mu_{X|\sigma(T)}(B)(\omega) = \frac{1}{kT(\omega)^{k-1}} \frac{d\mu_B}{d\lambda_1}(T(\omega)).$$

Then since $\mu_{X|\sigma(T)}(B)(\cdot)$ is a function of T, it is $\sigma(T)$-measurable. Also note that from $X_1, \ldots, X_k \overset{i.i.d}{\sim} \text{Uniform}(0,\theta)$, $0 \leq T \leq \theta$ a.s., and hence $\sigma(T)$ is generated by $\{T^{-1}(0,t) : 0 < t < \theta\}$. Hence it suffices to show that for all $t \in (0,\theta)$,

$$\int_{T^{-1}(0,t)} \frac{1}{kT(\omega)^{k-1}} \frac{d\mu_B}{d\lambda_1}(T(\omega)) dP = \int_{T^{-1}(0,t)} 1_{X \in B} dP.$$

Note that the induced measure μ_T on $(\mathbb{R}_+, \mathcal{B}_{\mathbb{R}_+})$ by $\mu_T(A) = P(T^{-1}(A))$ has Radon-Nikodym derivative with respect to λ_1 as $\frac{d\mu_T}{d\lambda_1}(x) = \frac{kx^{k-1}}{\theta^k} I_{(0,\theta)}(x)$. Then by using change of variable, LHS can be expanded as

$$\begin{aligned}\int_{T^{-1}(0,t)} \frac{1}{kT(\omega)^{k-1}} \frac{d\mu_B}{d\lambda_1}(T(\omega)) dP(\omega) &= \int_{(0,t)} \frac{1}{kx^{k-1}} \frac{d\mu_B}{d\lambda_1}(x) d\mu_T(x) \\ &= \int_0^t \frac{1}{kx^{k-1}} \frac{d\mu_B}{d\lambda_1}(x) \frac{d\mu_T}{d\lambda_1}(x) I_{(0,\theta)}(x) d\lambda_1(x) \\ &= \frac{1}{\theta^k} \int_{(0,t)} \frac{d\mu_B}{d\lambda_1}(x) d\lambda_1(x) \\ &= \frac{\mu_B((0,t))}{\theta^k} = \frac{\lambda_B(m^{-1}(0,t))}{\theta^k} \\ &= \frac{\lambda_B((0,t)^k)}{\theta^k} = \frac{\lambda_{\mathbb{R}_+^k}(B \cap (0,t)^k)}{\theta^k}.\end{aligned}$$

And RHS can be expanded as

$$\int_{T^{-1}(0,t)} 1_{X\in B} dP = P(X \in B, T \in (0,t)) = P(X \in B \cap (0,t)^k)$$
$$= \frac{\lambda_{\mathbb{R}^k_+}(B \cap (0,t)^k)}{\theta^k}.$$

Hence $\int_{T^{-1}(0,t)} \frac{1}{kT(\omega)^{k-1}} \frac{d\mu_B}{d\lambda_1}(T(\omega))dP = \int_{T^{-1}(0,t)} 1_{X\in B} dP$, i.e. $\mu_{X|\sigma(T)}(B)(\omega) = \frac{1}{kT(\omega)^{k-1}} \frac{d\mu_B}{d\lambda_1}(T(\omega))$. Since $\mu_{X|\sigma(T)}$ doesn't depend on θ, T is a sufficient statistic for θ.

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

$$y \mapsto \frac{1}{P\left(Y^{-1}(\{y\})\right)} \int_{Y^{-1}(\{y\})} X dP.$$

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...).

Points: 10 pts.

Solution.

Let ν be a measure on $(\mathcal{Y}, 2^{\mathcal{Y}})$ induced by P and Y, so that for any $A \subset \mathcal{Y}$, $\nu(A) = P(Y^{-1}(A)) = \sum_{y\in A} P\left(Y^{-1}(\{y\})\right)$. Since $\mathcal{Y}$ is countable, $\sigma(Y) = \{Y^{-1}(A) : A \subset \mathcal{Y}\}$. Hence for any $B \in \sigma(Y)$, there exists $A \subset \mathcal{Y}$ with $B =$

[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.

$Y^{-1}(A)$, and hence applying change of variable gives

$$\begin{aligned}\int_B g(Y)dP(\omega) &= \int_{Y^{-1}(A)} g(Y(\omega))dP(\omega) \\ &= \int_A g(y)d\nu(y) \\ &= \sum_{y\in A} g(y)\nu(\{y\}) \\ &= \sum_{y\in A} \int_{Y^{-1}(\{y\})} XdP \\ &= \int_{Y^{-1}(A)} XdP.\end{aligned}$$

And hence $g(Y) = \mathbb{E}[X|Y]$.

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 σ-fields contained in $\mathcal{F}$, then $\mathbb{E}[X|\mathcal{C}] = \mathbb{E}[X]$, a.e. $[P]$.

Points: 10 pts.

Solution.
For any $B \in \mathcal{C}$, note that X and I_B is independent, and hence

$$\mathbb{E}[XI_B] = \mathbb{E}[X]\mathbb{E}[I_B] = \mathbb{E}[X]P(B).$$

And hence

$$\int_B XdP = \mathbb{E}[XI_B] = \mathbb{E}[X]P(B) = \int_B \mathbb{E}[X]dP.$$

Hence $\mathbb{E}[X]$ is a version of $\mathbb{E}[X|\mathcal{C}]$.

7. Let X be a random variable on $(\Omega, \mathcal{F}, P)$ and $\mathcal{C} \subset \mathcal{F}$ a σ-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.

Points: 10 pts.

Solution.

For $p \geq 1$, note that $f(x) = x^p$ for $x \geq 0$ is a convex funciton. Hence by applying conditional Jensen's inequality,

$$|\mathbb{E}[X|\mathcal{C}]|^p \leq \mathbb{E}[|X|^p|\mathcal{G}].$$

Then, taking expectation on both side and applying tower property on the right yields

$$\begin{aligned}\mathbb{E}[|\mathbb{E}[X|\mathcal{C}]|^p] &\leq \mathbb{E}[\mathbb{E}[|X|^p|\mathcal{G}]] \\ &= \mathbb{E}[|X|^p].\end{aligned}$$

And correspondingly,

$$\begin{aligned}Var[\mathbb{E}[X|\mathcal{C}]] &= \mathbb{E}\left[\mathbb{E}[X|\mathcal{C}]^2\right] - (\mathbb{E}[\mathbb{E}[X|C]])^2 \\ &\leq \mathbb{E}\left[X^2\right] - (\mathbb{E}[X])^2 = Var[X].\end{aligned}$$

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 μ be a σ-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 ψ 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

$$P_\theta(A) = \int_A p_\theta(x) d\mu(x), \quad \forall A \in \mathcal{B}^k.$$

In particular, since by construction $P_\theta << \mu$ for all θ, 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 σ-finite measure, then S is $\mathrm{cl}(\{x \colon p(x) > 0\})$, the closure of all points of positive density).

(b) Prove that ψ is a convex function on Θ and that Θ 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 ψ 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 Θ is a subset (typically open) of $\mathbb{R}^d$, where $d < k$. In this case, the density (w.r.t. μ) 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 $\tau \colon \mathbb{R}^k \to \mathbb{R}^d$ 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 Θ as the distribution of X and densities of the form (1) with respect to a new σ-finite measure ν on $(\mathbb{R}^k, \mathcal{B}^k)$. (Find that measure, too!).
More impoertantly, Show that the conditional distribution of X given $T = t$ is uniform over the set $\{x \in \mathbb{R}^k \colon \tau(x) = t\}$. Conclude that $\tau(X)$ is a sufficient statistic for θ.

(e) **Conditionals and Marginals of Exponential Families.** For any x in the domain of τ, 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 σ-finite measure ν_{t_2} (which depends on t_2) and natural parameter θ_1. Thus, conditioning on T_2 eliminates the dependence on θ_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 σ-finite measure ν_{θ_2}, which depends on θ_2. Notice that the marginal distribution of T_1 still depends on θ_2 (the fact that the dominating measure depends on θ_2 further implies that the log-partition function depends on θ_2). Conclude that (unless θ_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...*

Points: 30 pts = 4 + 4 + 4 + 6 + 12.

Solution.

(a)

Note that $p_\theta(x) = \exp(x \cdot \theta - \log \psi(\theta)) > 0$ for all $x \in \mathbb{R}^k$, and hence $\mu(A) > 0$ implies $P_\theta(A) = \int_A p_\theta(x) d\mu(x) > 0$, i.e. $\mu \ll P_\theta$. And hence

$$\mu \ll P_\theta \ll \mu.$$

Hence $\forall \theta_1, \theta_2 \in \Theta$, $P_{\theta_1} \ll \mu \ll P_{\theta_2} \ll \mu \ll P_{\theta_1}$ holds. Also, for all $A \in \mathcal{B}^k$, $P_\theta(A) > 0 \iff \mu(A) > 0$, and hence

$$\mathrm{supp}(\mu) = \mathrm{supp}(P_\theta).$$

Hence $\forall \theta_1, \theta_2 \in \Theta$, $\mathrm{supp}(P_{\theta_1}) = \mathrm{supp}(\mu) = \mathrm{supp}(P_{\theta_2})$.

(b)

For all $\theta_1, \theta_2 \in \mathbb{R}^k$ and $\lambda \in [0, 1]$,

$$\begin{aligned}
\psi(\lambda\theta_1 + (1-\lambda)\theta_2) &= \log\left(\int_{\mathbb{R}^k} \exp\left(x \cdot (\lambda\theta_1 + (1-\lambda)\theta_1)\right) d\mu(x)\right) \\
&= \log\left(\int_{\mathbb{R}^k} \left(\exp(x \cdot \theta_1)\right)^{\lambda} \left(\exp(x \cdot \theta_2)\right)^{1-\lambda} d\mu(x)\right) \\
&\leq \log\left(\left(\int_{\mathbb{R}^k} \exp(x \cdot \theta_1) d\mu(x)\right)^{\lambda} \left(\int_{\mathbb{R}^k} \exp(x \cdot \theta_2) d\mu(x)\right)^{1-\lambda}\right) \\
&= \lambda \log\left(\int_{\mathbb{R}^k} \exp(x \cdot \theta_1) d\mu(x)\right) + (1-\lambda) \log\left(\int_{\mathbb{R}^k} \exp(x \cdot \theta_2) d\mu(x)\right) \\
&= \lambda\psi(\theta_1) + (1-\lambda)\psi(\theta_2).
\end{aligned}$$

(c)

Note that $P_{\theta_1} = P_{\theta_2}$ if and only if $p_{\theta_1} = p_{\theta_2}$ a.e. $[\mu]$. And note that

$$\begin{aligned} p_{\theta_1}(x) = p_{\theta_2}(x) &\iff \exp\left(x\cdot\theta_1 - \psi(\theta_1)\right) = \exp\left(x\cdot\theta_2 - \psi(\theta_2)\right) \\ &\iff x\cdot\theta_1 - \psi(\theta_1) = x\cdot\theta_2 - \psi(\theta_2) \\ &\iff (\theta_1 - \theta_2)\cdot x = \psi(\theta_1) - \psi(\theta_2). \end{aligned}$$

(d)
Define ν on $(\mathbb{R}^k, \mathcal{B}^k)$ as $\nu(A) = \mu(\tau^{-1}(A))$, i.e. induced measure. Give partial order on $\mathbb{R}^k$ as $x \leq y \iff x_i \leq y_i$ for all $1 \leq i \leq k$. Then for all $t \in \mathbb{R}^k$ and $\Delta t \in \mathbb{R}^k$,

$$\begin{aligned} P(T\in A) &= P(\tau(X)\in A) \\ &= \int_{\{x:\,\tau(x)\in A\}} \exp(\tau(x)\cdot\theta - \psi(\theta))d\mu(x) \\ &= \int_{\tau^{-1}(A)} \exp(\tau(x)\cdot\theta - \psi(\theta))d\mu(x) \\ &= \int_A \exp(t\cdot\theta - \psi(\theta))d\nu(t), \end{aligned}$$

and hence

$$\frac{dP_T}{d\nu} = \exp(t\cdot\theta - \psi(\theta)),$$

and hence the distribution of T is of exponential family.
Note that conditional distribution $\mu_{X|\sigma(T)}(\cdot)(\cdot) : \mathcal{B}_{\mathbb{R}^k}\times\Omega \to [0,1]$ is characterized by that for all $B\in\mathcal{B}_{\mathbb{R}^k}$, $\mu_{X|\sigma(T)}(B)$ is a version of $\mathbb{E}\left[1_{X\in B}|\sigma(T)\right]$, i.e. $\mu_{X|\sigma(T)}(B)(\cdot)$ is $\sigma(T)$-measurable and for all $A\in\sigma(T)$,

$$\int_A \mu_{X|\sigma(T)}(B)(\omega)dP = \int_A 1_{X\in B}(\omega)dP.$$

Given that ν is a counting measure, we will argue that

$$\mu_{X|\sigma(T)}(B)(\omega) = \frac{\mu(B\cap\tau^{-1}(\{T(\omega)\}))}{\nu(\{T(\omega)\})}.$$

Since ν is counting measure, any $A\in\sigma(T)$ can be expressed as $A = T^{-1}(C)$ with C being countable and for all $t\in C$, $\nu(\{t\}) > 0$. Then LHS can be expanded as

$$\begin{aligned} \int_A \mu_{X|\sigma(T)}(B)(\omega)dP = \int_{T^{-1}(C)} \frac{\mu(B\cap\tau^{-1}(\{T(\omega)\}))}{\nu(\{T(\omega)\})}dP &= \int_C \frac{\mu(B\cap\tau^{-1}(\{t\}))}{\nu(\{t\})}dP_T(t) \\ &= \int_C \frac{\mu(B\cap\tau^{-1}(\{t\}))}{\nu(\{t\})}\exp(t\cdot\theta - \psi(\theta))d\nu(t) \\ &= \sum_{t\in C}\mu(B\cap\tau^{-1}(\{t\}))\exp(t\cdot\theta - \psi(\theta)). \end{aligned}$$

And RHS can be expanded as

$$\begin{aligned}
\int_A 1_{X\in B}(\omega)dP &= P(T\in C, X\in B) = P(X\in B\cap\tau^{-1}(C)) \\
&= \int_{B\cap\tau^{-1}(C)} \exp(\tau(x)\cdot\theta-\psi(\theta))d\mu(x) \\
&= \sum_{t\in C}\int_{B\cap\tau^{-1}(\{t\})} \exp(t\cdot\theta-\psi(\theta))d\mu(x) \\
&= \sum_{t\in C}\mu(B\cap\tau^{-1}(\{t\}))\exp(t\cdot\theta-\psi(\theta)).
\end{aligned}$$

Hence $\int_A \mu_{X|\sigma(T)}(B)(\omega)dP = \int_A 1_{X\in B}(\omega)dP$, i.e. $\mu_{X|\sigma(T)}(B)(\omega) = \frac{\mu(B\cap\tau^{-1}(\{T(\omega)\}))}{\nu(\{T(\omega)\})}$. In particular,. $\mu_{X|T=t}(B) = \frac{\mu(B\cap\tau^{-1}(\{t\})}{\nu(\{t\})}$ is uniform over the set $\{x\in\mathbb{R}^k : \tau(x) = t\}$.

(d)

Define ν on $(\mathbb{R}^k, \mathcal{B}^k)$ as $\nu(A) = \mu(\tau^{-1}(A))$, i.e. induced measure. Give partial order on $\mathbb{R}^k$ as $x\le y \iff x_i\le y_i$ for all $1\le i\le k$. Then for all $t\in\mathbb{R}^k$ and $\Delta t\in\mathbb{R}^k$,

$$\begin{aligned}
P(T\in A) &= P(\tau(X)\in A) \\
&= \int_{\{x:\,\tau(x)\in A\}} \exp(\tau(x)\cdot\theta-\psi(\theta))d\mu(x) \\
&= \int_{\tau^{-1}(A)} \exp(\tau(x)\cdot\theta-\psi(\theta))d\mu(x) \\
&= \int_A \exp(t\cdot\theta-\psi(\theta))d\nu(t),
\end{aligned}$$

and hence

$$\frac{dP_T}{d\nu} = \exp(t\cdot\theta-\psi(\theta)).$$

(e)

i.

Note that conditional distribution $\mu_{T_1|\sigma(T_2)}(\cdot)(\cdot) : \mathcal{B}_{\mathbb{R}^l}\times\Omega\to[0,1]$ is characterized by that for all $B\in\mathcal{B}_{\mathbb{R}^l}$, $\mu_{T_1|\sigma(T_2)}(B)$ is a version of $\mathbb{E}\left[1_{T_1\in B}|\sigma(T_2)\right]$, i.e. $\mu_{T_1|\sigma(T_2)}(B)(\cdot)$ is $\sigma(T_2)$-measurable and for all $A\in\sigma(T_2)$,

$$\int_A \mu_{T_1|\sigma(T_2)}(B)(\omega)dP = \int_A 1_{T_1\in B}(\omega)dP.$$

Let $\nu_{1|2}(\cdot)(\cdot) : \mathcal{B}_{\mathbb{R}^l}\times\mathbb{R}^k\to[0,1]$ be the regular conditional probability where $\nu_{1|2}(B)(t) = \mathbb{E}_\nu\left[1_{B\times\mathbb{R}^{k-l}}|\Pi_2^{-1}(\mathcal{B}^k)\right](t_2)$. Then $\int\left(\int f(t_1)d\nu_{1|2}(t_1)\right)(t_2)g(t_2)d\nu(t) =$

$\int f(t_1)g(t_2)d\nu(t)$ by standard machinary. Also, let $\psi_{t_2}(\theta_1) = \log\left(\int_{\mathbb{R}^l} \exp(t_1 \cdot \theta_1) d\nu_{1|2}(t_1)\right)(t_2)$. We will argue that

$$\mu_{T_1|\sigma(T_2)}(B)(\omega) = \left(\int_B \exp(t_1 \cdot \theta_1 - \psi_{T_2(\omega)}(\theta_1)) d\nu_{1|2}(t_1)\right)(T_2(\omega)).$$

Then $\mu_{T_1|\sigma(T_2)}(B)$ is $\sigma(T_2)$-measurable. Also, all $A \in \sigma(T_2)$ can be expressed as $A = T_2^{-1}(C)$. Hence LHS can be computed as

$$\begin{aligned}
&\int_A \mu_{T_1|\sigma(T_2)}(B)(\omega) dP \\
&= \int_{T_2^{-1}(C)} \mu_{T_1|\sigma(T_2)}(B)(\omega) dP(\omega) \\
&= \int_{\mathbb{R}^l \times C} (\int_B \exp(t_1 \cdot \theta_1 - \psi_{t_2}(\theta_1)) d\nu_{1|2}(t_1))(t_2) dP_T(t) \\
&= \int_{\mathbb{R}^l \times C} (\int_B \exp(t_1 \cdot \theta_1 - \psi_{t_2}(\theta_1)) d\nu_{1|2}(t_1))(t_2) \exp(t_1 \cdot \theta_1 + t_2 \cdot \theta_2 - \psi(\theta)) d\nu(t) \\
&= \int_{\mathbb{R}^l \times C} \left(\int_{\mathbb{R}^k} \exp(t_1 \cdot \theta_1) d\nu_{1|2}(t_1)\right)(t_2) \exp(\psi_{t_2}(\theta_1)) \\
&\qquad \times (\int_B \exp(t_1 \cdot \theta_1) d\nu_{1|2}(t_1))(t_2) \exp(t_2 \cdot \theta_2 - \psi(\theta)) d\nu(t) \\
&= \int_{\mathbb{R}^l \times C} (\int_B \exp(t_1 \cdot \theta_1) d\nu_{1|2}(t_1))(t_2) \exp(t_2 \cdot \theta_2 - \psi(\theta)) d\nu(t) \\
&= \int_{B \times C} \exp(t_1 \cdot \theta_1 + t_2 \cdot \theta_2 - \psi(\theta)) d\nu(t).
\end{aligned}$$

And RHS can be computed as

$$\begin{aligned}
\int_A 1_{T_1 \in B}(\omega) dP = P(T_1 \in B, T_2 \in C) &= \int_{B \times C} dP_T \\
&= \int_{B \times C} \exp(t_1 \cdot \theta_1 + t_2 \cdot \theta_2 - \psi(\theta)) d\nu(t).
\end{aligned}$$

Hence $\int_A \mu_{T_1|\sigma(T_2)}(B)(\omega) dP = \int_A 1_{T_1 \in B}(\omega) dP$, i.e.

$$\mu_{T_1|\sigma(T_2)}(B)(\omega) = \left(\int_B \exp(t_1 \cdot \theta_1 - \psi_{T_2(\omega)}(\theta_1)) d\nu_{1|2}(t_1)\right)(T_2(\omega)).$$

In particular, $\mu_{T_1|T_2=t_2}(B) = \int_B \exp(t_1 \cdot \theta_1 - \psi_{t_2}(\theta_1)) d\nu_{1|2}(t_1)$ has a density $\exp(t_1 \cdot \theta_1 - \psi_{t_2}(\theta_1))$ with respect to a σ-finite measure $\nu_{1|2}$. Hence it is an exponential family of dimension l. Also,

$$\begin{aligned}
\psi_{t_2}(\theta_1) < \infty &\iff \left(\int_{\mathbb{R}^l} \exp(t_1 \cdot \theta_1) d\nu_{1|2}(t_1)\right)(t_2) < \infty \\
&\iff \exists \theta_2 \text{ with } \int_{\mathbb{R}^{k-l}} \left(\int_{\mathbb{R}^l} \exp(t_1 \cdot \theta_1) d\nu_{1|2}(t_1)\right)(t_2) \exp(t_2 \cdot \theta_2) d\nu(t) < \infty \\
&\iff \exists \theta_2 \text{ with } \int_{\mathbb{R}^k} \exp(t_1 \cdot \theta_1 + t_2 \cdot \theta_2) d\nu(t) < \infty \\
&\iff \exists \theta_2 \text{ with } (\theta_1, \theta_2) \in \Theta,
\end{aligned}$$

and hence its natural parameter is $\{\theta_1 : (\theta_1, \theta_2) \in \Theta\}$.
ii.
Let ν_{θ_2} be the measure on $(\mathbb{R}^{k-l}, \mathcal{B}^{k-l})$ defined as for any $A_1 \in \mathcal{B}^{k-l}$, $\nu_{\theta_2}(A_1) = \int_{\mathbb{R}^l \times A_1} \exp(t_2 \cdot \theta_2) d\nu(t)$, and let $\psi_{\theta_2} : \mathbb{R}^l \to \mathbb{R}$ as $\psi_{\theta_2}(\theta_1) = \int_{\mathbb{R}^l} \exp(t_1 \cdot \theta_1) d\nu_{\theta_2}(t_1)$, then

$$\psi_{\theta_2}(\theta_1) = \int_{\mathbb{R}^l} \exp(t_1 \cdot \theta_1) d\nu_{\theta_2}(t_1) = \int_{\mathbb{R}^k} \exp(t_1 \cdot \theta_1) \exp(t_2 \cdot \theta_2) d\nu(t) = \psi(\theta).$$

Then $P(T_1 \in A_1)$ can be expanded as

$$\begin{aligned}
P(T_1 \in A_1) &= \int_{\Pi_1^{-1}(A_1)} \exp(t_1 \cdot \theta_1 + t_2 \cdot \theta_2 - \psi(\theta)) d\nu(t) \\
&= \int_{\Pi_1^{-1}(A_1)} \exp(t_1 \cdot \theta_1 - \psi_{\theta_2}(\theta_1)) \exp(t_2 \cdot \theta_2) d\nu(t) \\
&= \int_{A_1} \exp(t_1 \cdot \theta_1 - \psi_{\theta_2}(\theta_1)) d\nu_{\theta_2}(t_1),
\end{aligned}$$

Hence the marginal distribution of T_1 has a density $p_{T_1}(t_1) = \exp(t_1 \cdot \theta_1 - \psi_{\theta_1}(\theta_2))$ with respect to σ-finite measure ν_{θ_2}. Since ν_{θ_2} still depends on θ_2, the marginal distribution T_2 is not in general an exponential family.
iii.
Let $v_1, \ldots, v_n$ be n vertices, and for $1 \le i < j \le n$, let

$$X_{ij} = \begin{cases} 1, & \text{if there exists edges between } v_i \text{ and } v_j, \\ 0, & \text{otherwise.} \end{cases}$$

Then since X_{ij}'s are i.i.d. *Bernoulli*(p), $P(X_{ij} = x_{ij}) = p^{x_{ij}}(1-p)^{1-x_{ij}}$ and

$$\begin{aligned}
P(X = x) &= p^{\sum x_{ij}}(1-p)^{\frac{n(n-1)}{2} - \sum x_{ij}} I_{\{0,1\}^{n(n-1)/2}}(x) \\
&= \exp\left(\left(\sum x_{ij}\right) \log\left(\frac{p}{1-p}\right) + \frac{n(n-1)}{2} \log(1-p)\right) I_{\{0,1\}^{n(n-1)/2}}(x).
\end{aligned}$$

Hence it is one-dimensional exponential family with sufficient statistics $\sum X_{ij}$.