Neyman Pearson Causal Inference -Supplemental Notes

Joni

Theorem 1. *Let U be a uniform random variable on $[0,1]$. Then, the optimal detector $\hat{\boldsymbol{\chi}}^*$ for the Neyman-Pearson causal discovery problem is given as follows,*

$$\hat{\boldsymbol{\chi}}^*_{i,j} = \begin{cases} 0, & \frac{dP_{i,j}}{dQ_{i,j}} > \frac{w^-_{i,j}}{w^+_{i,j}}\gamma \\ 0, & \frac{dP_{i,j}}{dQ_{i,j}} = \frac{w^-_{i,j}}{w^+_{i,j}}\gamma \text{ and } U \le \eta \\ 1, & \frac{dP_{i,j}}{dQ_{i,j}} < \frac{w^-_{i,j}}{w^+_{i,j}}\gamma \\ 1, & \frac{dP_{i,j}}{dQ_{i,j}} = \frac{w^-_{i,j}}{w^+_{i,j}}\gamma \text{ and } U > \eta \end{cases} \tag{1}$$

where γ and $\eta \in [0,1]$ are chosen so that $\epsilon^+ = \epsilon$. ■

Proof. We proceed first by showing existence. Let $\gamma \ge 0$ be the largest γ such that

$$\sum_{i,j} w^+_{i,j} P_{i,j}\left(\frac{dP_{i,j}}{dQ_{i,j}} < \gamma\right) \le \epsilon. \tag{2}$$

Then, if the inequality is strict, select η to be

$$\eta = \frac{\epsilon - \sum_{i,j} w^+_{i,j} P_{i,j}\left(\frac{dP_{i,j}}{dQ_{i,j}} < \frac{w^-_{i,j}}{w^+_{i,j}}\gamma\right)}{\sum_{i,j} w^+_{i,j} P_{i,j}\left(\frac{dP_{i,j}}{dQ_{i,j}} = \frac{w^-_{i,j}}{w^+_{i,j}}\gamma\right)} \tag{3}$$

otherwise choose η arbitrarily. Hence, we have that

$$\epsilon^+ = \sum_{i,j} w^+_{i,j} P^+_{i,j} = \sum_{i,j} w^+_{i,j}\left(P_{i,j}\left(\frac{dP_{i,j}}{dQ_{i,j}} < \frac{w^-_{i,j}}{w^+_{i,j}}\gamma\right) + \eta P_{i,j}\left(\frac{dP_{i,j}}{dQ_{i,j}} = \frac{w^-_{i,j}}{w^+_{i,j}}\gamma\right)\right) \tag{4}$$

$$= \sum_{i,j} w^+_{i,j} P_{i,j}\left(\frac{dP_{i,j}}{dQ_{i,j}} < \frac{w^-_{i,j}}{w^+_{i,j}}\gamma\right) + \eta \sum_{i,j} w^+_{i,j} P_{i,j}\left(\frac{dP_{i,j}}{dQ_{i,j}} = \frac{w^-_{i,j}}{w^+_{i,j}}\gamma\right) = \epsilon. \tag{5}$$

We next show that threshold rules are optimal. We use the Lagrange multiplier $\lambda_0 > 0$ (if the optimal estimator is such that $\lambda_0 = 0$ then the constraint is inactive, and so the optimal estimator should always declare $\hat{\boldsymbol{\chi}}_{i,j} = 1$) and seek to minimize $\epsilon^- + \lambda_0\epsilon^+$. Then, we have

$$\epsilon^- + \lambda_0\epsilon^+ = \sum_{i,j} w^-_{i,j}\mathbb{P}(\hat{\boldsymbol{\chi}}_{i,j} = 0|\boldsymbol{\chi}_{i,j} = 1) + \lambda_0 \sum_{i,j} w^+_{i,j}\mathbb{P}(\hat{\boldsymbol{\chi}}_{i,j} = 1|\boldsymbol{\chi}_{i,j} = 0) \tag{6}$$

$$= \sum_{i,j}\int_{X_1,\dots,X_n} w^-_{i,j}\mathbb{P}(\hat{\boldsymbol{\chi}}_{i,j} = 0|X_1,\dots,X_n)Q_{i,j}(X_1,..,X_n) + \lambda_0 w^+_{i,j}\mathbb{P}(\hat{\boldsymbol{\chi}}_{i,j} = 1|X_1,\dots,X_n)P_{i,j}(X_1,..,X_n). \tag{7}$$

Then, to minimize $\epsilon^- + \lambda_0 \epsilon^+$, the estimator $\hat{\boldsymbol{\chi}}$ should be such that $\hat{\boldsymbol{\chi}}_{i,j} = 0$ if $w^+_{i,j} Q_{i,j} \le \lambda_0 w^-_{i,j} P_{i,j}(X_1, ..., X_n)$ and $\hat{\boldsymbol{\chi}}_{i,j} = 1$ otherwise. Since threshold rules are optimal, it remains to be seen how to select the threshold γ. Observe that the probabilities $P_{i,j}\left(\frac{dP_{i,j}}{dQ_{i,j}} < \frac{w^-_{i,j}}{w^+_{i,j}}\gamma\right)$ are increasing in γ for all pairs (i,j), and that the probabilities $Q_{i,j}\left(\frac{dP_{i,j}}{dQ_{i,j}} > \frac{w^-_{i,j}}{w^+_{i,j}}\gamma\right)$ are decreasing in γ. Hence, ϵ^+ is increasing in γ, and ϵ^- is decreasing in γ. Then, to minimize ϵ^-, γ should be chosen to make ϵ^+ as large as possible. □

We repeat some definitions here.

Definition 1. The Rényi divergence of order λ between two probability measures P and Q is given as

$$D_\lambda(P||Q) \doteq \frac{1}{\lambda - 1} \log \int_{\mathcal{X}} \left(\frac{dP}{dQ}\right)^\lambda dQ, \tag{8}$$

where $\frac{dP}{dQ}$ is the Radon-Nikodym derivative of P with respect to Q.

Definition 2. For any two probability measures P and Q, let

$$D'_\lambda(P||Q) \doteq \int_{\mathcal{X}} F_\lambda(x; P, Q) \log \frac{dP}{dQ} \tag{9}$$

$$D''_\lambda(P||Q) \doteq \int_{\mathcal{X}} F_\lambda(x; P, Q) \left(\log \frac{dP}{dQ}\right)^2 - \left(D'_\lambda(P||Q)\right)^2 \tag{10}$$

$$\text{where} \quad F_\lambda(x; P, Q) \doteq \frac{f(x)^\lambda g(x)^{1-\lambda}}{\int_{\mathcal{X}} f(x)^\lambda g(x)^{1-\lambda} d\mu} \quad \text{for} \ \lambda \in [0,1] \tag{11}$$

Theorem 2. *For any detector $\hat{\boldsymbol{\chi}}$ that satisfies $\epsilon^+ \le \epsilon$, we have that for any $\lambda \in [0,1]$.*

$$\begin{aligned} \epsilon^- \ge \frac{1}{2} \sum_{i,j} & w^-_{i,j} e^{-(1-\lambda) D_\lambda(P_{i,j}||Q_{i,j}) - \lambda D'_\lambda(P_{i,j}||Q_{i,j}) - \lambda\sqrt{2D''_\lambda(P_{i,j}||Q_{i,j})}} \\ & - \epsilon \max_{i,j} \left\{ \frac{w^-_{i,j}}{w^+_{i,j}} e^{-D'_\lambda(P_{i,j}||Q_{i,j}) + (1-2)\lambda\sqrt{2D''_\lambda(P_{i,j}||Q_{i,j})}} \right\} \end{aligned} \tag{12}$$

Proof. It is sufficient to show that for any pair (i,j),

$$Q_{i,j}(\hat{\boldsymbol{\chi}}_{i,j} = 0) \ge \frac{1}{2} e^{-(1-\lambda) D_\lambda(P_{i,j}||Q_{i,j}) - \lambda D'_\lambda(P_{i,j}||Q_{i,j}) - \lambda\sqrt{2D''_\lambda(P_{i,j}||Q_{i,j})}} - \epsilon e^{-D'_\lambda(P_{i,j}||Q_{i,j}) + (1-2\lambda)\sqrt{2D''_\lambda(P_{i,j}||Q_{i,j})}}, \tag{13}$$

It is relatively easy to show that for any pair (i,j),

$$Q_{i,j}(X_1, .., X_n) = e^{-(1-\lambda) D_\lambda(P_{i,j}||Q_{i,j}) - \lambda \log \frac{dP_{i,j}}{dQ_{i,j}}} F_\lambda(X_1, ..., X_n; P_{i,j}, Q_{i,j}), \tag{14}$$

$$P_{i,j}(X_1, .., X_n) = e^{-(1-\lambda) D_\lambda(P_{i,j}||Q_{i,j}) + (1-\lambda) \log \frac{dP_{i,j}}{dQ_{i,j}}} F_\lambda(X_1, ..., X_n; P_{i,j}, Q_{i,j}). \tag{15}$$

Then, consider the following sets

$$\mathcal{X}_{i,j,\lambda} = \left\{ X_1, ..., X_n : |\log \frac{dP_{i,j}}{dQ_{i,j}} - D'_\lambda(P_{i,j}||Q_{i,j})| \le \sqrt{2D''_\lambda(P_{i,j}||Q_{i,j})} \right\}, \tag{16}$$

$$\mathcal{X}^0_{i,j} = \left\{ X_1, ..., X_n : \hat{\boldsymbol{\chi}}_{i,j} = 0 \right\}, \tag{17}$$

$$\mathcal{X}^1_{i,j} = \left\{ X_1, ..., X_n : \hat{\boldsymbol{\chi}}_{i,j} = 1 \right\}. \tag{18}$$

Hence, for any $X_1, ..., X_n \in \mathcal{X}_{i,j,\lambda}$, we have that

$$Q_{i,j}(X_1, .., X_n) \geq e^{-(1-\lambda)D_\lambda(P_{i,j}||Q_{i,j}) - \lambda D'_\lambda(P_{i,j}||Q_{i,j}) - \lambda\sqrt{2D''_\lambda(P_{i,j}||Q_{i,j})}} F_\lambda(X_1, ..., X_n; P_{i,j}, Q_{i,j}), \tag{19}$$

$$P_{i,j}(X_1, .., X_n) \geq e^{-(1-\lambda)D_\lambda(P_{i,j}||Q_{i,j}) + (1-\lambda)D'_\lambda(P_{i,j}||Q_{i,j}) - (1-\lambda)\sqrt{2D''_\lambda(P_{i,j}||Q_{i,j})}} F_\lambda(X_1, ..., X_n; P_{i,j}, Q_{i,j}). \tag{20}$$

Notice that $D'_\lambda(P_{i,j}||Q_{i,j})$ is actually the mean of $\log \frac{dP_{i,j}}{dQ_{i,j}}$ with respect to the distribution F_λ, and $D''_\lambda(P_{i,j}||Q_{i,j})$ is its variance. Then, $\mathcal{X}_{i,j,\lambda}$ is the event that the log-likelihood ratio is within $\sqrt{2}$ standard deviations of its mean, and so from Chebychev's inequality,

$$\int_{\mathcal{X}_{i,j,\lambda}} F_\lambda(X_1, ..., X_n; P_{i,j}, Q_{i,j}) \geq \frac{1}{2}. \tag{21}$$

Then, from the union bound, we have that

$$\int_{\mathcal{X}_{i,j,\lambda} \cap \mathcal{X}^0_{i,j}} F_\lambda(X_1, ..., X_n; P_{i,j}, Q_{i,j}) + \int_{\mathcal{X}_{i,j,\lambda} \cap \mathcal{X}^1_{i,j}} F_\lambda(X_1, ..., X_n; P_{i,j}, Q_{i,j}) \geq \frac{1}{2}. \tag{22}$$

From (20), we have that

$$\begin{aligned}
&\int_{\mathcal{X}_{i,j,\lambda} \cap \mathcal{X}^1_{i,j}} F_\lambda(X_1, ..., X_n; P_{i,j}, Q_{i,j}) \\
&\leq e^{(1-\lambda)D_\lambda(P_{i,j}||Q_{i,j}) - (1-\lambda)D'_\lambda(P_{i,j}||Q_{i,j}) + (1-\lambda)\sqrt{2D''_\lambda(P_{i,j}||Q_{i,j})}} \int_{\mathcal{X}_{i,j,\lambda} \cap \mathcal{X}^1_{i,j}} P_{i,j}(X_1, ..., X_n) \\
&\leq e^{(1-\lambda)D_\lambda(P_{i,j}||Q_{i,j}) - (1-\lambda)D'_\lambda(P_{i,j}||Q_{i,j}) + (1-\lambda)\sqrt{2D''_\lambda(P_{i,j}||Q_{i,j})}} P^+_{i,j}
\end{aligned} \tag{23}$$

and so

$$\int_{\mathcal{X}_{i,j,\lambda} \cap \mathcal{X}^0_{i,j}} F_\lambda(X_1, ..., X_n; P_{i,j}, Q_{i,j}) \geq \frac{1}{2} - e^{(1-\lambda)D_\lambda(P_{i,j}||Q_{i,j}) - (1-\lambda)D'_\lambda(P_{i,j}||Q_{i,j}) + (1-\lambda)\sqrt{2D''_\lambda(P_{i,j}||Q_{i,j})}} P^+_{i,j}. \tag{24}$$

Then, (19) gives

$$Q^-_{i,j} \geq e^{-(1-\lambda)D_\lambda(P_{i,j}||Q_{i,j}) - \lambda D'_\lambda(P_{i,j}||Q_{i,j}) - \lambda\sqrt{2D''_\lambda(P_{i,j}||Q_{i,j})}} \int_{\mathcal{X}_{i,j,\lambda} \cap \mathcal{X}^0_{i,j}} F_\lambda(X_1, ..., X_n; P_{i,j}, Q_{i,j}), \tag{25}$$

together with (24) we have

$$\begin{aligned}
Q^-_{i,j} \geq \frac{1}{2} &e^{-(1-\lambda)D_\lambda(P_{i,j}||Q_{i,j}) - \lambda D'_\lambda(P_{i,j}||Q_{i,j}) - \lambda\sqrt{2D''_\lambda(P_{i,j}||Q_{i,j})}} \\
&- e^{-D'_\lambda(P_{i,j}||Q_{i,j}) + (1-2)\lambda\sqrt{2D''_\lambda(P_{i,j}||Q_{i,j})}} P^+_{i,j}.
\end{aligned} \tag{26}$$

Multiplying both sides by $w^-_{i,j}$ and summing over all (i,j) pairs gives us

$$\begin{aligned}
\sum_{i,j} w^-_{i,j} Q^-_{i,j} \geq \frac{1}{2} \sum_{i,j} &w^-_{i,j} e^{-(1-\lambda)D_\lambda(P_{i,j}||Q_{i,j}) - \lambda D'_\lambda(P_{i,j}||Q_{i,j}) - \lambda\sqrt{2D''_\lambda(P_{i,j}||Q_{i,j})}} \\
&- \sum_{i,j} w^+_{i,j} \frac{w^-_{i,j}}{w^+_{i,j}} e^{-D'_\lambda(P_{i,j}||Q_{i,j}) + (1-2)\lambda\sqrt{2D''_\lambda(P_{i,j}||Q_{i,j})}} P^+_{i,j},
\end{aligned} \tag{27}$$

which yields

$$\begin{aligned}\sum_{i,j} w_{i,j}^- Q_{i,j}^- \geq \frac{1}{2} \sum_{i,j} w_{i,j}^- e^{-(1-\lambda) D_\lambda (P_{i,j}||Q_{i,j}) - \lambda D'_\lambda (P_{i,j}||Q_{i,j}) - \lambda \sqrt{2 D''_\lambda (P_{i,j}||Q_{i,j})}} \\ - \max_{i,j} \left\{ \frac{w_{i,j}^-}{w_{i,j}^+} e^{-D'_\lambda (P_{i,j}||Q_{i,j}) + (1-2)\lambda \sqrt{2 D''_\lambda (P_{i,j}||Q_{i,j})}} \right\} \sum_{i,j} w_{i,j}^+ P_{i,j}^+ .\end{aligned} \tag{28}$$

Noticing that $\sum_{i,j} w_{i,j}^+ P_{i,j}^+ = \epsilon^+ \leq \epsilon$ completes the proof. $\square$