E2 209 – Topics in Information Theory and Statistical Learning, Jan-Apr 2018
Homework #1

---

(1) *Properties of total variation distance*
Consider two distributions P and Q on $\mathcal{X}$. Show the following properties of $d(P,Q) = \sup_A P(A) - Q(A)$.

(a) $d(P,Q) = d(Q,P)$.

(b) $d(P,Q) = \sup\{\frac{1}{2}\sum_{i=1}^{k} |P(A_i) - Q(A_i)| : \{A_1, ..., A_k\} \text{ is a partition of } \mathcal{X}\}$.

(c) If P and Q have densities f and g w.r.t. μ,

$$d(P,Q) = \frac{1}{2}\int |f(x) - g(x)|\mu(dx),$$

and

$$d(P,Q) = P(\{x : f(x) \geq g(x)\}) - Q(\{x : f(x) \geq g(x)\}).$$

(2) *Bounds among distances and divergences*
Consider two distributions P and Q such that $P \ll Q$. Denote by f the Raydon-Nikodym derivative of P w.r.t. Q (you can think of the discrete case where $f(x) = P(x)/Q(x)$).
The *chi-squared divergence* $\chi^2(P,Q)$ between P and Q is given by

$$\chi^2(P,Q) = \mathbb{E}_Q\left\{(f(X) - 1)^2\right\}.$$

The *squared Hellinger distance* $\mathcal{H}(P,Q)$ between P and Q is given by

$$\mathcal{H}(P,Q) = \frac{1}{2}\mathbb{E}_Q\left\{(\sqrt{f(X)} - 1)^2\right\}.$$

Establish the following bounds relating these distances to the total variation distance and the KL divergence

(a) $D(P\|Q) \leq \chi^2(P,Q)$.

(b) $\mathcal{H}^2(P,Q) \leq d(P,Q)^2 \leq \mathcal{H}(P,Q)(2 - \mathcal{H}(P,Q))$.

(3) *Pinsker's inequality*
Show that for $p, q \in [0,1]$

$$|p-q|^2 \leq c \cdot \left(p \ln\frac{p}{q} + (1-p)\ln\frac{1-p}{1-q}\right)$$

if and only if $c \geq 1/2$.

(4) *Estimating k-ary distribution*
Let $\mathcal{P}_k$ denote the $(k-1)$-dimensional probability simplex. Consider the problem of estimating $P \in \mathcal{P}_k$ by observing n independent samples from P. Denote by $\mathcal{F}$ the family of estimators $\hat{P} : \mathbf{x}^n \mapsto \hat{P}_{\mathbf{x}^n} \in \mathcal{P}_k$. Define the minimax risk $R(k,n)$ as

$$R(k,n) = \min_{\hat{P}\in\mathcal{F}} \max_{P\in\mathcal{P}_k} \mathbb{E}_P\left\{d(P, \hat{P}_{X^n})\right\}.$$

Find upper and lower bounds for $R(k,n)$.

(5) *Bias of Estimators*
For $P \in \mathcal{P}_k$, let $X_1, ..., X_n$ denote n independent samples from P.

(a) *(Estimating moments of a distribution)* For $l \in \mathbb{N}, l \geq n$, find an unbiased estimator of $\sum_{i=1}^{k} P(i)^l$ from n independent samples from P, namely $e : [k]^n \to \mathbb{R}_+$ such that $\mathbb{E}_P \{e(X^n)\} = \sum_{i=1}^{k} P(i)^l$.

(b) *(Missing mass estimation)* Denote by N_x the number of times a symbol x appears in X^n. Find an estimator e for the probability of missing mass $M_n = \sum_{x:N_x=0} P(x)$ such that

$$\mathbb{E}_P \{M_{n-1}\} \leq \mathbb{E}_P \{e(X^n)\}.$$

(c) *(Linear estimators)* Denote by n_l the number of symbols that appear l times, $0 \leq l \leq n$. A linear estimator of a parameter has the form $\sum_l a_l n_l$. For a given function $f : [0,1] \to [0,1]$, consider the estimation of $F(P) = \sum_{i=1}^{n} f(P(i))$. Find the bias of a linear estimator for $F(P)$.

(6) *Sheffé estimators*
Consider the following modification of the standard parametric estimation problem: Given a parametric family $\mathcal{P} = \{P_\theta, \theta \in \Theta\}$, we seek to estimate P_θ by observing n independent samples $X_1, ..., X_n$ from it. For a minimax-risk formulation with $d(P_\theta, P_{\hat{\theta}})$ as the loss function, use Scheffé selectors to give estimators for the following problems and analyse their performances:

(a) $\Theta = [0, 1]$, $P_\theta = Ber(\theta)$, $\theta \in \Theta$.

(b) $\Theta = \mathbb{R}_+$, $P_\lambda = Poi(\lambda)$, $\lambda \in \Theta$.