Stats 210A, Fall 2021
Homework 1

Due date: Wednesday, Sep. 8

You may disregard measure-theoretic niceties about conditioning on measure-zero sets, almost-sure equality vs. actual equality, “all functions” vs. “all measurable functions,” etc. (unless the problem is explicitly asking about such issues).

1. Bias-Variance Tradeoff

Consider a generic estimation setting where we observe $X \sim P_\theta$, for a model $P = \{P_\theta : \theta \in \Theta \subseteq \Theta\}$, and we want to estimate $\theta$ using some estimator $\delta(X) \in \mathbb{R}^d$. The bias of $\delta$ (under sampling from $P_\theta$) is defined as

$$
\text{Bias}_{\theta}(\delta) = \mathbb{E}_{\theta}[\delta(X)] - \theta.
$$

For $d = 1$, it is well-known that the mean squared error $\text{MSE}(\theta; \delta)$ can be decomposed as the sum of the squared bias of $\delta$ and its variance:

$$
\text{MSE}(\theta; \delta) = \text{Bias}_{\theta}(\delta)^2 + \text{Var}_{\theta}(\delta).
$$

(a) Derive the correct generalization of (1) for general $d \geq 1$, where the MSE is defined as

$$
\text{MSE}(\theta; \delta) = \mathbb{E}_{\theta}\|\delta(X) - \theta\|^2.
$$

It might help to start with $d = 1$.

(b) Suppose that we are estimating the false positive rate of a new diagnostic test for some disease, using a sample of $n$ specimens taken from a population known not to have the disease we are testing for. If $X$ is the number of false positives and $\theta \in (0, 1)$ is the false positive rate, assume $X \sim \text{Binom}(n, \theta)$. The “obvious” estimator is $\delta_0(X) = X/n$.

However, biological samples are expensive to obtain and the new test is a slightly modified version of an old test whose false positive rate is known to be $\theta_0 \in (0, 1)$, so we might want to “shrink” the estimator toward $\theta_0$ as follows:

$$
\delta_\gamma(X) = \gamma \theta_0 + (1 - \gamma) \frac{X}{n}, \quad \text{for } \gamma \in [0, 1],
$$

where taking $\gamma = 0$ reduces to the “obvious” estimator $\delta_0(X) = X/n$.

Find the MSE of $\delta_\gamma(X)$ as an explicit expression in $\theta_0, \theta, n$, and $\gamma$.

(c) Find the parameter $\gamma^*$ for which the MSE is minimized, as an expression in $n, \theta$, and $\theta_0$. What happens to $\gamma^*$ if we send $\theta \to \theta_0$ holding $\theta_0$ and $n$ fixed? What if we send $n \to \infty$ holding $\theta$ and $\theta_0$ fixed instead? Explain why these limits make sense.

(d) In our calculation above, $\gamma^*$ is never exactly zero. That is, a smidgeon of shrinkage always beats no shrinkage. Does this prove that $\delta_0$ is inadmissible? Prove or disprove whether $\delta_0$ is dominated by any $\delta_\gamma$.

2. Convexity of $A(\eta)$ and $\Xi_1$

Let $P = \{p_\eta : \eta \in \Xi_1\}$ denote an $s$-parameter exponential family in canonical form

$$
p_\eta(x) = e^{\eta^T(x) - A(\eta)} h(x), \quad A(\eta) = \log \int_X e^{\eta^T(x)} h(x) \, d\mu(x),
$$

where $\eta \in \mathbb{R}^s$. The function $A(\eta)$ is convex on $\Xi_1$, and

$$
\frac{d^2}{d\eta^2} A(\eta) = \mathbb{E}_{\theta}[\eta^2] > 0.
$$

(a) Show that $A(\eta)$ is strictly convex on $\Xi_1$.

(b) Show that $A(\eta)$ is concave on $\Xi_1$.

(c) Let $\eta \in \mathbb{R}^s$. Show that $A(\eta)$ is convex on $\mathbb{R}^s$.

(d) Let $\eta \in \mathbb{R}^s$. Show that $A(\eta)$ is concave on $\mathbb{R}^s$.
4. Exponential families maximize entropy

Recall Hölder’s inequality: if $q_1, q_2 \geq 1$ with $q_1^{-1} + q_2^{-1} = 1$, and $f_1$ and $f_2$ are ($\mu$-measurable) functions from $\mathcal{X}$ to $\mathbb{R}$, then

$$\|f_1 f_2\|_{L^q(\mu)} \leq \|f_1\|_{L^{q_1}(\mu)} \|f_2\|_{L^{q_2}(\mu)} , \quad \text{where } \|f\|_{L^q(\mu)} = \left( \int_{\mathcal{X}} |f(x)|^q \, d\mu(x) \right)^{1/q} .$$

(Note that $q_1 = q_2 = 2$ reduces to Cauchy-Schwarz).

(a) Show that $A(\eta) : \mathbb{R}^s \to [0, \infty]$ is a convex function: that is, for any $\eta_1, \eta_2 \in \mathbb{R}^s$ (not just in $\Xi_1$), and $c \in [0, 1]$ then

$$A(c\eta_1 + (1-c)\eta_2) \leq cA(\eta_1) + (1-c)A(\eta_2)$$

(Hint: try $q_1 = c^{-1}, f_1(x)^{1/c} = e^{\eta_1 T(x)} h(x)$.)

(b) Conclude that $\Xi_1 \subseteq \mathbb{R}^s$ is convex.

3. Expectation of an increasing function

(a) Assume $X \sim P$ is a real-valued random variable. Show that if $f(x)$ and $g(x)$ are non-decreasing functions of $x$, then

$$\text{Cov}(f(X), g(X)) \geq 0$$

(Hint: derive the identity $E[(f(X_1) - f(X_2))(g(X_1) - g(X_2))] = 2\text{Cov}(f(X_1), g(X_1))$, where $X_1, X_2 \overset{i.i.d.}{\sim} P$).

(b) Let $p_\eta(x)$ be a one-parameter canonical exponential family with non-decreasing sufficient statistic $T(x)$, where $x \in \mathcal{X} \subseteq \mathbb{R}$:

$$p_\eta(x) = e^{\eta T(x) - A(\eta)} h(x).$$

Let $\psi(x)$ be any non-decreasing bounded function. Show that, for $\eta \in \Xi_1, \frac{d\eta}{d\eta} E_\eta[\psi(X)] \geq 0$.

(Hint: find an expression for $\frac{d\eta}{d\eta} E_\eta[\psi(X)]$ by using methods akin to the ones we used in class to derive the differential identities. You may appeal to Keener Theorem 2.4 to justify differentiating under the integral sign.)

(c) Conclude that $X$ is stochastically increasing in $\eta$; that is, show $P_\eta(X \leq c)$ is non-increasing in $\eta$, for every $c \in \mathbb{R}$.

4. Exponential families maximize entropy

The entropy (with respect to $\mu$) of a random variable $X$ with density $p$, is defined by

$$h(p) = \mathbb{E}_p(-\log p(X)) = -\int_{\{x : p(x) > 0\}} \log(p(x)) p(x) \, d\mu(x).$$

This quantity arises naturally in information theory as a minimal expected code length.

Let $T : \mathcal{X} \to \mathbb{R}^s$ denote a generic function, and let $\alpha$ be some vector in the interior of the convex hull of $T(\mathcal{X}) = \{T(x) : x \in \mathcal{X}\}$. Consider the problem of maximizing $h(p)$ over all probability densities subject to the constraint that $E_p[T(X)] = \alpha$. That is, we want to solve

$$\text{maximize} \quad -\int_{\{x : p(x) > 0\}} \log(p(x)) p(x) \, d\mu(x)$$

s.t. $p(x) \geq 0$, $\int_{\mathcal{X}} p(x) \, d\mu(x) = 1$, and $\int_{\mathcal{X}} p(x) T(x) \, d\mu(x) = \alpha \in \mathbb{R}^s$. 

2
(a) If $X$ is a finite set with $\mu(\{x\}) > 0$ for all $x \in X$, show that the optimal $p^*$ is a member the $s$-parameter exponential family

$$p_\eta(x) = e^{\eta^T(x) - A(\eta)},$$

with parameter $\eta^* \in \mathbb{R}^s$ chosen so that $p_{\eta^*}$ satisfies the constraints.

(Hint: use Lagrange multipliers).

(b) Blithely\(^1\) applying the result of (a) to $X = \mathbb{R}$, find the distribution that maximizes entropy with respect to the Lebesgue measure, subject to the constraint that $\mathbb{E}(X) = \mu, \text{Var}(X) = \sigma^2$.

5. Gamma family

The gamma family is a two-parameter family of distributions on $\mathbb{R}_+ = [0, \infty)$, with density

$$p_{k,\theta}(x) = \frac{x^{k-1}e^{-x/\theta}}{\Gamma(k)\theta^k}$$

with respect to the Lebesgue measure on $\mathbb{R}_+, k > 0$ and $\theta > 0$ are respectively called the shape and scale parameters, and $\Gamma(k)$ is the gamma function, defined as

$$\Gamma(k) = \int_0^\infty x^{k-1}e^{-x} \, dx.$$

The gamma distribution generalizes the exponential distribution ($\text{Exp}(\theta) = \theta^{-1}e^{-x/\theta} = \text{Gamma}(1, \theta)$) and the chi-squared distribution ($\chi^2_d = \frac{x^{d/2-1}e^{-x/2}}{\Gamma(d/2)2^{d/2}} = \text{Gamma}(d/2, 2)$).

(a) Show that the Gamma is a 2-parameter exponential family by putting it into its canonical form. Find the natural parameter, sufficient statistic, carrier density, and log-partition function (Note: there are multiple valid ways of doing this).

(b) Find the mean and variance of $X \sim \Gamma(k, \theta)$.

(c) Find the moment generating function of $X \sim \Gamma(k, \theta)$:

$$M_X(u) = \mathbb{E}_{k,\theta}[e^{uX}],$$

and use it to find the distribution of $X_+ = \sum_{i=1}^n X_i$ where $X_1, \ldots, X_n$ are mutually independent with $X_i \sim \text{Gamma}(k_i, \theta)$.

You may use without proof the following uniqueness result about MGFs: If $Y$ and $Z$ are two random variables whose MGFs coincide in a neighborhood of 0 ($\exists \delta > 0$ for which $M_Y(u) = M_Z(u) < \infty$ for all $u \in [-\delta, \delta]$), then $Y$ and $Z$ have the same distribution.

---

\(^1\)Meaning naively, without any concern that anything new might go wrong in a continuous space