STAT3006/STAT7305 Assignment 1—Probability and Optimization Theory

Due Date: 26th August 2022
Weighting: 20%

Instructions

- The assignment consists of **four (4) problems, each problem is worth 25 marks, and each mark is equally weighted.**
- The mathematical elements of the assignment can be completed by hand, **in LaTeX (preferably)**, or in Word (or other typesetting software). The mathematical derivations and manipulations should be accompanied by clear explanations in English regarding necessary information required to interpret the mathematical exposition.
- Computation problems should be answered using programs in the R language.
- Computer generated plots and hand drawn graphs should be included together with the text where problems are answered.
- Submission files should include the following (which ever applies to you):
  - Scans of handwritten mathematical exposition.
  - Typeset mathematical exposition, outputted as a `pdf` file.
  - Typeset answers to computational problems, outputted as a `pdf` file.
  - Program code/scripts that you wish to submit, outputted as a `txt` file.
- Mathematical problems should be answered with reference to results presented in the Main Text (refer, page numbers), Remarks, Exercises, and Corollaries/Lemma/Propositions/Theorems from the Lecture Notes, if required. If a mathematical result is used that is not presented in the Lecture Notes, then its common name (e.g., "Bayes' Theorem", "Intermediate Value Theorem", "Borel–Cantelli Lemma", etc.) should be cited, or else a reference to a text containing the result should be provided (preferably a textbook).

- **All submission files should be labeled with your name and student number and archived together in a `zip` file and submitted at the TurnItIn link on Blackboard.** We suggest naming using the convention:

  `[LastName_FirstName/StudentNumber]_STAT3006A1_[AnythingElse].[FileExtension]`.

- As per `my.uq.edu.au/information-and-services/manage-my-program/student-integrityand-conduct/academic-integrity-and-student-conduct`, what you submit should be your own work. Even where working from sources, you should endeavour to write in your own words. You should use consistent notation throughout your assignment and define whatever is required.

Problem 1 [25 Marks]

Let $(\Omega, \mathcal{F}, \mathrm{P})$ be a probability space, and let $(X_i(\omega))_{i\in[n]}$ be a sequence of independent and identically distribution (IID) random variables from $(\Omega, \mathcal{F})$ to $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$, each with the same image measure P_X, as the random variable $X(\omega)$.

(a) Assuming that X has expectation $\mu_X = \mathrm{E}(X)$ and finite variance

$$\sigma_X^2 = \mathrm{E}\left[(X - \mu_X)^2\right] < \infty,$$

prove that the sample mean,

$$\bar{X}_n(\omega) = \frac{1}{n}\sum_{i=1}^{n} X_i(\omega),$$

is a *consistent* estimator of μ_X in the sense that $\bar{X}_n$ converges in probability to μ_X, as $n \to \infty$. Such a result is usually referred to as a *weak law of large numbers*.

[10 Marks]

Recall that the cumulative distribution function (CDF) of X is defined as:

$$F(x) = \mathrm{P}(\omega : X(\omega) \leq x).$$

Using the sequence $(X_i)_{i\in[n]}$ estimate the CDF $F(x)$ using the *empirical CDF*:

$$\bar{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} \chi_{(-\infty,x)}(X_i).$$

(b) For each fixed $x \in \mathbb{R}$, show that

$$\mathbb{I}_{\alpha,n}(x) = \left\{ y \in \mathbb{R} : \bar{F}_n(x) - \frac{1}{2\sqrt{n\alpha}} \leq y \leq \bar{F}_n(x) + \frac{1}{2\sqrt{n\alpha}} \right\}$$

is a $(1-\alpha) \times 100\%$ confidence interval for $F(x)$ in the sense that

$$\mathrm{P}\left(\omega : F(x) \in \mathbb{I}_{\alpha,n}(x)\right) \geq 1 - \alpha.$$

[10 Marks]

The *Kolmogorov's strong law of large numbers* improves upon the result from (a) stating that if the expectation of X is finite (i.e. $|\mu_X| < \infty$), then $\bar{X}_n$ converges almost surely to μ_X, as $n \to \infty$.

(c) Assuming that $\mathrm{P} \ll \mathrm{Leb}$, use the strong law of large numbers to prove the *Glivenko–Cantelli Theorem*; i.e., show that

$$\sup_{x \in \mathbb{R}} \left| \bar{F}_n(x) - F(x) \right|$$

converges to 0, almost surely, as $n \to \infty$.

[5 Marks]

This result, along with the laws of large numbers, is generally considered to be the fundamental law(s) of statistics in the sense that they provide a mechanism for probability distributions to be realized in reality; i.e., repeated observation of IID random variables and averaging the outcomes yields insight regarding the properties of the underlying probability measure P.

Problem 2

Let $(\Omega, \mathcal{F}, \mathrm{P})$ be a probability space, and let $(X_i(\omega))_{i \in [n]}$ be a sequence of IID random variables from $(\Omega, \mathcal{F})$ to $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$, each with the same image measure P_X, as the random variable $X(\omega)$.

Suppose that $\mathrm{P}_X \ll \mathrm{Leb}$, and has Radon–Nykodym derivative (probability density function; PDF) $p_{\theta_0} : \mathbb{R} \to \mathbb{R}_{\geq 0}$, where $\theta_0 \in \mathbb{R}$, the so-called data generating parameter corresponding to P_X, is unknown to us. That is, we only know that P_X has a PDF of the form p_θ, for some value of $\theta \in \mathbb{R}$, but not which particular value of θ (i.e., θ_0).

We wish to test the simple hypotheses that either the *null hypothesis*

$$\mathrm{H}_0\text{: } \theta_0 = \theta_1^*$$

is true, or the *alternative hypothesis*

$$\mathrm{H}_1\text{: } \theta_0 = \theta_2^*$$

is true, for a pair of predetermined values $\theta_1^*, \theta_2^* \in \mathbb{R}$, where $\theta_1^* \neq \theta_2^*$. To construct a test between the two hypotheses, we construct the so-called *likelihood ratio statistic*:

$$\mathscr{L}_n(\omega) = \frac{\prod_{i=1}^n p_{\theta_2^*}(X_i(\omega))}{\prod_{i=1}^n p_{\theta_1^*}(X_i(\omega))}.$$

(a) Under the null hypothesis (i.e., we assume that $\theta_0 = \theta_1^*$ is true, and subsequently $p_{\theta_0} = p_{\theta_1^*}$), show that

$$\mathrm{E}[\mathscr{L}_n(\omega)] = 1.$$

You may use the fact that if $(Y_i(\omega))_{i\in[n]}$ is a sequence of independent random variables, then so $(f_i \circ Y_i)_{i\in[n]}$ is also a sequence of independent random variables, where $f_i : \mathbb{R} \to \mathbb{R}$ is a function, for each $i \in [n]$.

[5 Marks]

We say that a random variable $E : \Omega \to \mathbb{R}_{\geq 0}$ is an e-value if

$$\mathrm{E}[E] \leq 1,$$

and we say that a random variable $P(\omega)$ is a p-value if for any $\alpha \in [0,1]$:

$$\mathrm{P}(\omega : P(\omega) \leq \alpha) \leq \alpha.$$

(b) Prove that $P = \min\{1, 1/E\}$ is a p-value.

[5 Marks]

We say that $\mathscr{R}_n$, $(X_i)_{i\in[n]} \mapsto \{0,1\}$, is a *rejection rule* (or *test*) for the hypotheses H_0 and H_1, where we say that we *reject* H_0 if $\mathscr{R}_n = 1$ and we say that we *accept* H_0 if $\mathscr{R}_n = 0$. We further say that $\mathscr{R}_n$ has *size* (or *significance level*) $\alpha \in (0,1)$ if

$$\mathrm{P}(\omega : \mathscr{R}_n = 1) \leq \alpha,$$

under the assumption that H_0 is true.

(c) Using the conclusions from (a) and (b), for any $\alpha \in (0,1)$, construct a rejection rule with size α for the test of H_0: $\theta_0 = \theta_1^*$ versus H_1: $\theta_0 = \theta_2^*$.

[5 Marks]

In the R programing language, we can generate an sequence $(Y_i)_{i\in[n]}$ of IID random variables, each with the same image probability measures P_Y, and PDF

$$p_\theta(y) = \phi_\theta(y) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(y-\theta)^2\right),$$

using the command `rnorm` (n, θ). We can also compute $\phi_\theta(y)$ using the command `dnorm` (y, θ).

(d) Assuming that X has image measure P_X has a PDF of the form $p_{\theta_0}(x) = \phi_{\theta_0}(x)$, use the commands above to program an R script that implements the rejection rule from (c) to test the hypotheses H_0: $\theta_0 = 0$ versus H_1: $\theta_0 = \theta_2^*$, and assess how often the test rejects H_0, as θ_2^* increases. We say that a test is powerful if H_0 is rejected with high probability when H_1 is true. Via computational experiments or otherwise, comment on the power of the test as θ_2^* increases away from 0.

[10 Marks]

Problem 3

(a) For non-negative numbers $u, v \in \mathbb{R}_{\geq 0}$ and positive coefficients $p, q \in \mathbb{R}_{>0}$, such that $1/p + 1/q = 1$, show that

$$uv \leq \frac{u^p}{p} + \frac{v^q}{q}.$$

You may use the fact that $x \mapsto \log(x)$ is concave on $\mathbb{R}_{>0}$.

[5 Marks]

Let $X(\omega)$ and $Y(\omega)$ be functions from measure space $(\Omega, \mathcal{F}, \mathrm{M})$ to $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$.

(b) Use the result from (a) to prove *Hölder's inequality*. That is, show that if

$$\int_\Omega |X|^p \,\mathrm{dM} < \infty \text{ and } \int_\Omega |Y|^q \,\mathrm{dM} < \infty,$$

for $1/p + 1/q = 1$, then

$$\int_\Omega |XY| \,\mathrm{dM} \leq \left(\int_\Omega |X|^p \,\mathrm{dM} \right)^{1/p} \left(\int_\Omega |Y|^q \,\mathrm{dM} \right)^{1/q}.$$

[10 Marks]

(c) Use Hölder's inequality (or otherwise) to prove *Minkowski's inequality*. That is, show that if

$$\int_\Omega |X|^p \,\mathrm{dM} < \infty \text{ and } \int_\Omega |Y|^p \,\mathrm{dM} < \infty,$$

for $p \geq 1$, then

$$\left(\int_\Omega |X + Y|^p \,\mathrm{dM} \right)^{1/p} \leq \left(\int_\Omega |X|^p \,\mathrm{dM} \right)^{1/p} + \left(\int_\Omega |Y|^p \,\mathrm{dM} \right)^{1/p}.$$

[5 Marks]

Let $(X_i)_{i\in[n]}$ be a sequence of functions from $(\Omega, \mathcal{F}, \mathrm{M})$ to $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$.

(d) Consider a generalization of Hölder's inequality of the form:

$$\int_\Omega \left| \prod_{i=1}^{n} X_i \right| \mathrm{dM} \leq \prod_{i=1}^{n} \left(\int_\Omega |X|^{q_i} \, \mathrm{dM} \right)^{1/q_i}, \tag{1}$$

where $(q_i)_{i\in[n]}$ is a sequence of constants, such that $q_i \in \mathbb{R}$ for each $i \in [n]$. Propose conditions on the values that $(q_i)_{i\in[n]}$ can take so that (1) is true, and argue (formally or informally) why you believe that your suggested conditions are correct.

[5 Marks]

Problem 4

Let $f : \mathbb{T} \to \mathbb{R}$ be a function on the domain $\mathbb{T} \subset \mathbb{R}^d$, where $\mathbb{T}$ is said to be convex. We say that f is μ-strongly convex on $\mathbb{T}$ (with respect to the Euclidean norm $\|\cdot\|$) if there exists a $\mu > 0$ such that for any $\theta, \tau \in \mathbb{T}$ and $\lambda \in [0, 1]$, we have

$$f(\lambda\theta + (1-\lambda)\tau) \leq \lambda f(\theta) + (1-\lambda) f(\tau) + \mu\lambda(1-\lambda)\|\theta - \tau\|^2. \tag{2}$$

This more structured notion of convexity allows for proofs of convergence and rates for many modern optimization algorithms, and problems relating to strongly convex functions are mathematically "easier" to solve that others.

(a) Characterization (2) is one of many characterizations of μ-strong convexity. Another useful characterization is that

$$g(\theta) = f(\theta) - \mu\|\theta\|^2 \tag{3}$$

is convex on $\mathbb{T}$. That is, f is so convex that even when we remove a quadratic function from f (i.e., g) we still have a convex function. Prove that these two characterizations ((2) and (3)) of μ-strong convexity are equivalent.

[5 Marks]

(b) Using either (2) or (3), prove that if f is μ-strongly convex on $\mathbb{T}$ then it is also strictly convex on $\mathbb{T}$, and propose a counterexample of a strictly convex function that is not μ-strongly convex for any $\mu > 0$.

[5 Marks]

We now assume that f is differentiable on a an open set $\bar{\mathbb{T}} \supset \mathbb{T}$, and thus is differentiable on $\mathbb{T}$. Then, we can further characterize μ-strong convexity on $\mathbb{T}$ by the inequality

$$f(\tau) \geq f(\theta) + (\tau - \theta)^\top \nabla f(\theta) + \frac{1}{2}\mu \|\tau - \theta\|. \tag{4}$$

(c) Using (4), show that if f is μ-strongly convex on $\mathbb{T}$ and $\theta^* \in \mathbb{T}$ is a critical point of f in the sense that

$$\nabla f(\theta^*) = 0,$$

then for all $\theta \in \mathbb{T}$ we have

$$f(\theta) \geq f(\theta^*) + \frac{1}{2}\mu \|\theta - \theta^*\|^2$$

and

$$f(\theta^*) \geq f(\theta) - \frac{1}{2\mu} \|\nabla f(\theta)\|^2.$$

[10 Marks]

For a differentiable function f, we say that it is β-smooth on $\mathbb{T}$ if

$$\|\nabla f(\theta) - \nabla f(\tau)\| \leq \beta \|\theta - \tau\|,$$

for every $\theta, \tau \in \mathbb{T}$.

(d) A typical algorithm for computing

$$\theta^* = \arg\min_{\theta \in \mathbb{T}} f(\theta)$$

is the *gradient descent method*, where we generate a sequence $(\theta_t)_{t \in \mathbb{N}}$ such that

$$\theta_{t+1} = \theta_t - \eta \nabla f(\theta_t),$$

for some positive constant $\eta > 0$, starting from some initial value $\theta_1 \in \mathbb{T}$. When f is convex and β-smooth on $\mathbb{T}$, it is provable that after $T \in \mathbb{N}$ steps of the gradient descent algorithm, we have the inequality:

$$f(\theta_T) - f(\theta^*) \leq \frac{1}{T-1} 2\beta \|\theta_1 - \theta^*\|^2, \tag{5}$$

by taking $\eta = 1/\beta$. On the other hand, if f is also μ-strongly convex on $\mathbb{T}$, then after

$T+1$ steps of gradient descent, we have

$$f\left(\theta_T\right) - f\left(\theta^*\right) \leq \exp\left(-\frac{4\left(T-1\right)}{\beta/\mu}\right)\frac{\beta}{2}\left\|\theta_1 - \theta^*\right\|^2, \tag{6}$$

when taking $\eta = 2/\left(\mu+\beta\right)$. Discuss the meanings of inequalities (5) and (6), and argue whether or not gradient descent performs better when f is strongly convex.

[5 Marks]