Stats 210A, Fall 2024
Homework 1

Due on: Wednesday, Sep. 11

Instructions: 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).

Problem 1 (Non-measurable sets)**.** This problem goes through a construction of a non-measurable set, meant to motivate measure theory from a real analysis perspective. It concerns the impossibility of defining "volume" for every subset of the unit interval $U = [0, 1)$.

For $x, y \in \mathbb{R}$ define the "wraparound addition" (modulo 1) as the fractional part of their sum:

$$x \oplus y = x + y - \lfloor x + y \rfloor.$$

Recall that for $x \in \mathbb{R}$ and $A \subseteq \mathbb{R}$ we define the set $x + A = \{x + a : \ a \in A\}$. Analogously, we can define

$$x \oplus A = \{x \oplus a : \ a \in A\} \subseteq U$$

Any reasonable definition of "volume" on the interval should have several properties:

(i) Additivity: $\lambda(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} \lambda(A_i)$ if all $A_i \subseteq U$ and $A_i \cap A_j = \emptyset$ for all $i \neq j$.

(ii) Translation invariance: $\lambda(x \oplus A) = \lambda(A), \ \forall x \in U, A \subseteq U$.

(iii) Interval length: $\lambda([x, y]) = y - x, \ \forall 0 \leq x \leq y < 1$.

Assume that some measure λ exists which satisfies (i)–(iii) and which is defined for all subsets of U. We will go through several steps to derive a contradiction.

(a) Define the function $A(x)$ mapping elements of U to subsets of U, via $A(x) = x \oplus \mathbb{Q}$, where $\mathbb{Q}$ is the set of rational numbers. Show that $\lambda(A(x)) = 0$ for any x.

(b) Consider the range $\mathcal{R}_A = \{A(x) : \ x \in U\}$. Show that $\mathcal{R}_A$ is a collection of uncountably many subsets of U, all of which are disjoint from each other. That is, show that for any $x, y \in U$, we have either $A(x) = A(y)$ or $A(x) \cap A(y) = \emptyset$.

(c) Now, let $B \subseteq U$ denote a new set, which we construct by selecting a *single element* from each set $R \in \mathcal{R}_A$ (it doesn't matter which element; note this step uses the axiom of choice.)

Define a new function $C(x) = x \oplus B$ and define $\mathcal{R}_C = \{C(x) : \ x \in \mathbb{Q}\}$. Show that $\mathcal{R}_C$ is a collection of *countably* many subsets of U, all of which are disjoint from each other, and whose union is U.

(d) Show that no matter what value $\lambda(B)$ takes, λ will have to violate one of the properties (i)–(iii)

Hint: what does the value of $\lambda(B)$ imply about $\lambda(U)$?

Because the Lebesgue measure satisfies properties (i)–(iii), it follows that λ must not be defined for every subset of U.

Moral: One motivation for measure theory is to find a way to exclude counterexamples like this. The pathological sets we're constructing here do not make it into the Borel σ-field.

Problem 2 (A conditional probability paradox)**.** Let $X, Y \overset{\text{i.i.d.}}{\sim} N(0,1)$. This problem is meant to show that by carelessly conditioning on probability-zero events we can get ourselves into trouble. It is directly inspired by a calculation I personally flubbed in graduate school.

(a) Defining $S = X + Y$ and $D = X - Y$, show S and D are independent and conclude that

$$\mathbb{E}[X^2 + Y^2 \mid D] = D^2/2 + 1$$

(b) Now define the polar parameterization (R, Θ) with $R = \sqrt{X^2 + Y^2}$ and $\Theta \in [0, 2\pi)$ such that $X = R\cos\Theta$ and $Y = R\sin\Theta$. Show that R is independent of Θ and conclude that

$$\mathbb{E}[X^2 + Y^2 \mid \Theta] = 2$$

(c) Use (a) and then (b) to find the expectation of $X^2 + Y^2$ conditional on the event $X = Y$. Can you come up with an intuitive explanation for how we could have arrived at two different answers?

Moral: Intuition may fail us when we condition on a measure-zero event, and in cases like this the meaning can be ambiguous and give different answers. Conditioning on a random variable, on the other hand, tends to give less ambiguous answers (there are still some ambiguities, similar to those we encounter in defining densities, but they don't really matter).

Problem 3 (Non-uniqueness of densities)**.** It can be good to keep in mind that, in general, densities are not unique. For example, depending on what textbook you look in you might find the standard exponential distribution $\text{Exp}(1)$ defined as the distribution with probability density function $e^{-x} \cdot 1\{x > 0\}$ or as the distribution with probability density function $e^{-x} \cdot 1\{x \geq 0\}$. Which one is the real exponential distribution?

The answer is that both are: the two densities are different functions but they result in exactly the same probability distribution because changing the integrand at a single point never affects a (Lebesgue or Riemann) integral. If we wanted to be perverse for some reason we could even define the density as something like $e^{-x} \cdot 1\{x > 0\} \cdot 1\{x \notin \mathbb{Q}\}$, and we'd still end up with exactly the same probability measure. So, while there is only one $\text{Exp}(1)$ distribution, there are many densities that equivalently describe its relationship to the Lebesgue measure.

However, you will show in this problem that any two density functions do have to be equal *almost everywhere*, meaning the set of points where they differ has to have measure 0.

(a) Consider two densities p_1 and p_2 with respect to some common measure μ on a sample space $\mathcal{X}$. Suppose p_1 and p_2 both result in the same probability measure P defined by $P(A) = \int 1_A(x) p_i(x)\, d\mu(x)$.

Define the set $A = \{x : \ p_1(x) \neq p_2(x)\}$, and show that $\mu(A) = 0$.

Hint: consider sets like

$$A_n = \left\{ x : \ p_1(x) - p_2(x) \in \left[\frac{1}{n+1}, \ \frac{1}{n} \right) \right\}$$

for $n = 1, 2, \ldots$. Don't worry about whether the sets A_n are measurable (they are).

(b) If $\mathcal{X}$ is countable, show that any probability measure P on the sample space $\mathcal{X}$ has a unique density with respect to the counting measure $\#$ on $\mathcal{X}$.

(c) Let P be a probability measure on $\mathbb{R}$, which has a density with respect to the Lebesgue measure. Show that P has at most one continuous density p.

Problem 4 (Densities for continuous-discrete mixtures)**.** Suppose μ_1 and μ_2 are both measures on $\mathcal{X}$, and $a_1, a_2 \geq 0$. You may use without proof that the sum $\nu = a_1\mu_1 + a_2\mu_2$ is also a measure, and that we have

$$\int f(x)\, d\nu(x) = a_1 \int f(x)\, d\mu_1(x) + a_2 \int f(x)\, d\mu_2(x),$$

provided the two integrals on the right are well-defined and finite.

(a) For $\mathcal{X} = \mathbb{R}$, define the measure $\mu(A) = \lambda(A) + \#(A)$, where λ represents the Lebesgue measure and $\#$ represents the counting measure on the set of integers $\mathbb{Z}$. For fixed $\theta \in \mathbb{R}$, define the random variable

$$X = \max(0, Z) \text{ where } Z \sim N(\theta, 1),$$

Let P_θ represent the probability distribution of X. Show that P_θ has no density with respect to λ or $\#$. Find a density of P_θ with respect to μ.

(b) Find a density for P_θ with respect to P_0, or show that none exists.

Moral: Our more general definition of densities extends to situations where there is no probability mass function or probability density function.

Problem 5 (Bias-variance tradeoff)**.** Consider a generic estimation setting where we observe $X \sim P_\theta$, for a model $\mathcal{P} = \{P_\theta : \theta \in \Theta \subseteq \mathbb{R}^d\}$, and we want to estimate θ using some estimator $\delta(X) \in \mathbb{R}^d$. The *bias* of δ (under sampling from P_θ) is defined as

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

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 δ and its variance:

$$\text{MSE}(\theta; \delta) = \text{Bias}_\theta(\delta)^2 + \text{Var}_\theta(\delta). \tag{1}$$

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

$$\text{MSE}(\theta; \delta) = \mathbb{E}_\theta \|\delta(X) - \theta\|_2^2.$$

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 θ_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 θ_0, θ, n, and γ.

(c) Find the parameter γ^* for which the MSE is minimized, as an expression in n, θ, and θ_0. What happens to γ^* if we send $\theta \to \theta_0$ holding θ_0 and n fixed? What if we send $n \to \infty$ holding θ and θ_0 fixed instead? Explain why these limits make sense.

(d) In our calculation above, γ^* is never exactly zero. That is, a smidgeon of shrinkage always beats no shrinkage. Does this prove that δ_0 is inadmissible? Prove or disprove whether δ_0 is dominated by any δ_γ.

Moral: Shading our estimate toward some "hunch" value can be an effective technique to improve an estimator's performance. This is a central idea in statistics and machine learning that goes by many names: regularization, shrinkage, and inductive bias, to name a few. The optimal amount of bias in an estimator depends on the sample size, and the accuracy of our hunch, but is rarely zero. This may give us pause about insisting that estimators should be unbiased, a theme to which we will return later.