EE 126 — Probability and Random Processes

April 10, 2018 — LLSE

1 Lecture

Neyman-Pearson Hypothesis Testing

There is an observation Y and two hypotheses – one from one setting $H_0 : Y \sim f_Y(y \mid 0)$ and one from another setting $H_1 : Y \sim f_Y(y \mid 1)$. We can invent a decision rule r, which takes us from an observation in $\mathbb{R}$ to a binary target $\{0, 1\}$. The goal is to minimize the false negative error $P(r(Y) = 0 \mid X = 1)$ subject to the constraint that the false positive error $P(r(Y) = 1 \mid X = 0)$ is less than or equal to β.

In other words, we're most concerned about the false negative errors (which are perhaps a matter of life and death, if it's a medical diagnosis), but we have to constrain false positives too (or else the system will just say "positive" every time, in order to avoid false negatives).

We define the **likelihood ratio** to be $L(y) = \frac{f_Y(y|1)}{f_Y(y|0)}$. The Neyman-Pearson theorem states that

$$r^*(Y) = \begin{cases} 1 & \text{if } L(Y) > \lambda \\ 0 & \text{if } L(Y) < \lambda \\ 1 \text{ w.p. } \gamma & \text{if } L(Y) = \lambda \end{cases}$$

The λ threshold we choose depends on our problem.

Some acronyms: PCD is the *probability of correct detection* (1 minus the false negative probability). PFA is the *probability of false alarm* (the false positive probability).

Example: bias of a coin. Under H_0, the coin is fair, with $P(H) = 0.5$. Under H_1, the coin is biased, with $P(H) = 0.6$. Accordingly, X is either 0 (fair) or 1 (biased). The goal is for the PFA to be less than or equal to 5%, i.e. our tolerance for false positives is 5% of the time. *What is the optimal decision rule for minimizing the probability of false negatives (i.e. for maximizing PCD)?*

We flip the coin n times and observe what comes out. We are interested in the probability we get a particular sequence, conditioned on each hypothesis:

$$P(Y_1 = y_1, ..., Y_n = y_n \mid X = 0) = 0.5^n \quad \text{(fair coin)}$$
$$P(Y_1 = y_1, ..., Y_n = y_n \mid X = 0) = 0.6^H 0.4^{n-H} \quad \text{(biased coin)}$$

where H is the number of heads.

The Neyman-Pearson theorem tells us to take the likelihood ratio

$$L(y_1, ..., y_n) = \frac{0.6^H 0.4^{n-H}}{0.5^n} = \left(\frac{0.4}{0.5}\right)^n \left(\frac{0.6}{0.4}\right)^H$$

Note: we don't need to know the ordering of heads; we only need the number. H is called a sufficient statistic for our decision rule.

As H goes up, $L(y)$ goes up exponentially. The Neyman-Pearson theorem says "see when $L(y)$ is above or below some λ." But since $L(y)$ increases monotonically as a function of H, the threshold λ on $L(y)$ is equivalent to a threshold n_0 on H.

We should thus calculate n_0 for a PFA of 5%. For $X = 0$, H is distributed as $\text{Bin}\big(n, \frac{1}{2}\big)$ (so $\mathbb{E}[H] = \frac{n}{2}$ & $var(H) = \frac{n}{4}$). We can use the CLT to determine an n_0 for which $P(H \geq n_0 \mid X = 0) = P\left((H - \mathbb{E}[H])/\sqrt{var(H)} \geq \frac{n_0 - n/2}{\sqrt{n}/2}\right) = 0.05$.

Estimation

LLSE (linear least squares estimation) and MMSE (minimum MSE) estimation have many applications, e.g. in sensor networks, radar, and ML. In them, we observe the random variable Y and turn it into an estimate $\hat{X}$ of the desired random variable X. The goal is to estimate X from Y as accurately as possible, i.e. minimize the error Δ. Specifically, we would like to minimize $\mathbb{E}[\Delta^2]$.

We are assumed to know the joint distribution of X and Y (Bayesian setting).

When we do linear estimation, $\hat{X}$ is constrained to be of the form $\hat{X} = a + bY$ for $a, b \in \mathbb{R}$. MMSE estimation explicitly minimizes an error and gives us the best of all estimators (linear, quadratic, etc.) but is harder to do.

LLSE

LLSE finds $\hat{X} = a + bY$ as $\min_{a,b} \mathbb{E}[(X - (a + bY))^2] = \min_{a,b} \mathbb{E}[\Delta^2]$. This can be done via calculus: just set the partial derivatives of $\mathbb{E}[\Delta^2]$ (which we'll alternatively call ξ) to 0 and solve for a and b.

$$\begin{aligned}\xi(a, b) &= \mathbb{E}[X^2 - 2(a + bY)X + (a + bY)^2] \\ &= \mathbb{E}[X^2] - 2a\mathbb{E}[X] - 2b\mathbb{E}[XY] + a^2 + 2ab\mathbb{E}[Y] + b^2\mathbb{E}[Y^2]\end{aligned}$$

Then

$$\begin{aligned}\frac{\partial \xi}{\partial a} &= 0 \Rightarrow -2\mathbb{E}[X] + 2a + 2b\mathbb{E}[Y] = 0 \\ \frac{\partial \xi}{\partial b} &= 0 \Rightarrow -2\mathbb{E}[XY] + 2a\mathbb{E}[Y] + 2b\mathbb{E}[Y^2] = 0\end{aligned}$$

Solving, we find that

$$\begin{aligned}b &= \frac{cov(X, Y)}{var(Y)} \\ a &= \mathbb{E}[X] - b\mathbb{E}[Y] \\ cov(X, Y) &= \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]\end{aligned}$$

and therefore $\hat{X} = L[X \mid Y] = \mathbb{E}[X] + \frac{cov(X,Y)}{var(Y)}(Y - \mathbb{E}[Y])$.

There are two important properties of our estimate $L[X \mid Y]$.

1. **It is unbiased.** $\mathbb{E}[\hat{X}] = \mathbb{E}[X]$, i.e. $\mathbb{E}[\Delta] = 0$.
2. **The error and the observation are uncorrelated.** $cov(\Delta, Y) = 0$, i.e. $\mathbb{E}[\Delta Y] = 0$ (projection property).

Note: we can greatly simplify our lives by working with zero-mean versions of X and Y. Let

$$\begin{aligned}X &= \bar{X} + \mathbb{E}[X] \\ Y &= \bar{Y} + \mathbb{E}[Y]\end{aligned}$$

where $\bar{X}$ and $\bar{Y}$ have zero mean. Then we can form $L[\bar{X} \mid \bar{Y}]$ and "add back" the means later. There are two main benefits: (1) it simplifies calculations (a becomes 0, while b stays the same), and (2) it permits a geometric treatment.

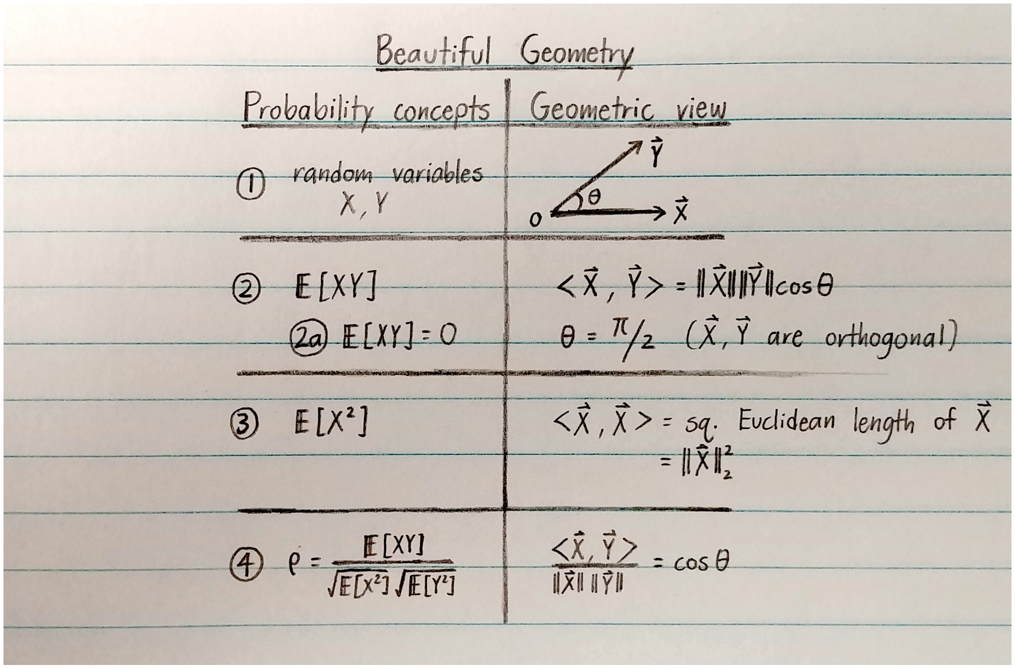
To be concrete,

$$L[\bar{X} \mid \bar{Y}] = \frac{cov(\bar{X}, \bar{Y})}{var(\bar{Y})}\bar{Y}$$

But this is the same as $\frac{cov(X,Y)}{var(Y)} = b$ in the general formula.

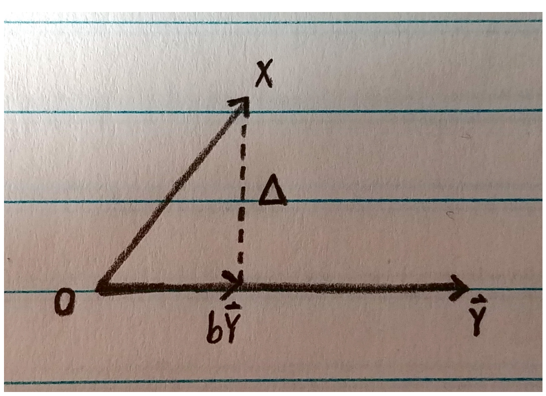
Vector Space Representation of Random Variables

There is a notion of a **Hilbert space**, which is called a "complete inner product vector space" and which allows us to think about random variables geometrically. From this point on, we will assume that X, Y are zero-mean random variables with finite variance.



Note: horizontal alignment signifies equivalence.

- (2) is a key point, as it describes how we define inner products.
- By (3), the norm (length) of a vector $\vec{X}$ is $\|\vec{X}\| = \texttt{sqrt}(\langle \vec{X}, \vec{X} \rangle)$ and is equivalent to $\sqrt{\mathbb{E}[X^2]}$.
- ρ in (4) is the correlation coefficient, meaning angles tell us how correlated the random variables are.



Error is minimized when Δ is orthogonal to Y.

$$\begin{aligned}\text{Therefore, } \hat{X} = bY &= \text{Projection}_{\vec{Y}} \vec{X} \\ &= \left\langle X, \frac{Y}{\|Y\|} \right\rangle \frac{Y}{\|Y\|} \\ &= \frac{\langle X, Y \rangle}{\|Y\|^2} \cdot Y.\end{aligned}$$