Section 8: Tail Bounds, MLE

Review of Main Concepts

- **Markov's Inequality**: Let X be a non-negative random variable, and $\alpha > 0$. Then,

$$\mathbb{P}(X \geq \alpha) \leq \frac{\mathbb{E}[X]}{\alpha}$$

- **Chebyshev's Inequality**: Suppose Y is a random variable with $\mathbb{E}[Y] = \mu$ and $\text{Var}(Y) = \sigma^2$. Then, for any $\alpha > 0$,

$$\mathbb{P}(|Y - \mu| \geq \alpha) \leq \frac{\sigma^2}{\alpha^2}$$

- **(Multiplicative) Chernoff Bound**: Let $X_1, X_2, ..., X_n$ be *independent* Bernoulli random variables.

  Let $X = \Sigma_{i=1}^{n} X_i$, and $\mu = \mathbb{E}[X]$. Then, for any $0 \leq \delta \leq 1$,

  - $\mathbb{P}(X \geq (1+\delta)\mu) \leq e^{-\frac{\delta^2 \mu}{3}}$
  - $\mathbb{P}(X \leq (1-\delta)\mu) \leq e^{-\frac{\delta^2 \mu}{2}}$

- **Realization/Sample**: A realization/sample x of a random variable X is the value that is actually observed.

- **Likelihood**: Let $x_1, \ldots x_n$ be iid realizations from probability mass function $p_X(\mathrm{x}\,; \theta)$ (if X discrete) or density $f_X(\mathrm{x}\,; \theta)$ (if X continuous), where θ is a parameter (or a vector of parameters). We define the likelihood function to be the probability of seeing the data.

  If X is discrete:

$$L(x_1, \ldots, x_n\,; \theta) = \prod_{i=1}^{n} p_X(x_i\,; \theta)$$

  If X is continuous:

$$L(x_1, \ldots, x_n\,; \theta) = \prod_{i=1}^{n} f_X(x_i\,; \theta)$$

- **Maximum Likelihood Estimator (MLE)**: We denote the MLE of θ as $\hat{\theta}_{\text{MLE}}$ or simply $\hat{\theta}$, the parameter (or vector of parameters) that maximizes the likelihood function (probability of seeing the data).

$$\hat{\theta}_{\text{MLE}} = \arg\max_{\theta} L(x_1, \ldots, x_n\,; \theta) = \arg\max_{\theta} \ln L(x_1, \ldots, x_n\,; \theta)$$

- **Log-Likelihood**: We define the log-likelihood as the natural logarithm of the likelihood function. Since the logarithm is a strictly increasing function, the value of θ that maximizes the likelihood will be exactly the same as the value that maximizes the log-likelihood.

  If X is discrete:

$$\ln L(x_1, \ldots, x_n\,; \theta) = \sum_{i=1}^{n} \ln p_X(x_i\,; \theta)$$

  If X is continuous:

$$\ln L(x_1, \ldots, x_n\,; \theta) = \sum_{i=1}^{n} \ln f_X(x_i\,; \theta)$$

- **Bias**: The bias of an estimator $\hat{\theta}$ for a true parameter θ is defined as $\text{Bias}(\hat{\theta}, \theta) = \mathbb{E}[\hat{\theta}] - \theta$. An estimator $\hat{\theta}$ of θ is unbiased iff $\text{Bias}(\hat{\theta}, \theta) = 0$, or equivalently $\mathbb{E}[\hat{\theta}] = \theta$.

- **Steps to find the maximum likelihood estimator, $\hat{\theta}$:**

(a) Find the likelihood and log-likelihood of the data.

(b) Take the derivative of the log-likelihood

(c) Set it to 0 to find a candidate for the MLE, $\hat{\theta}$. (note: at this step, we change from the θ to the $\hat{\theta}$ because in this step we are solving for the *maximum* likelihood estimator for θ)

(d) Take the second derivative and show that $\hat{\theta}$ indeed is a maximizer, that $\frac{\partial^2 L}{\partial \theta^2} < 0$ at $\hat{\theta}$. Also ensure that it is the global maximizer: check points of non-differentiability and boundary values.

1. Content Review

(a) True or false: the Union Bound always gives a result in $[0, 1]$.

(b) True or false: Markov's Inequality always gives a non-negative result.

(c) True or false: Chebyshev's Inequality can best be described as giving an upper bound on the distribution's right tail.

(d) **True or False**: The Log-Likelihood gives a slightly different estimate, but because it is close enough and easier to compute we use it for our estimate of θ.

(e) **True or False**: $\hat{\theta}$ is the true parameter and θ is our estimate.

(f) **True or False**: An estimator is unbiased if $\text{Bias}(\hat{\theta}, \theta) = \mathbb{E}[\hat{\theta}] - \theta = 0$ or equivalently $\mathbb{E}[\hat{\theta}] = \theta$

(g) You flip a coin 10 times and observe HHHTHHTHHH (8 heads, 2 tails). What is the MLE of θ, where θ is the true probability of seeing tails?

- ◯ $\hat{\theta} = .2$
- ◯ $\hat{\theta} = .25$
- ◯ $\hat{\theta} = .8$
- ◯ $\hat{\theta} = .3$

2. Tail bounds

Suppose $X \sim \text{Binomial}(6, 0.4)$. We will bound $\mathbb{P}(X \geq 4)$ using the tail bounds we've learned, and compare this to the true result.

(a) Give an upper bound for this probability using Markov's inequality. Why can we use Markov's inequality?

(b) Give an upper bound for this probability using Chebyshev's inequality. You may have to rearrange algebraically and it may result in a weaker bound.

(c) Give an upper bound for this probability using the Chernoff bound.

(d) Give the exact probability.

3. Exponential Tail Bounds

Let $X \sim \text{Exp}(\lambda)$ and $k > 1/\lambda$.

(a) Use Markov's inequality to bound P($X \geq k$).

(b) Use Markov's inequality to bound P($X < k$).

(c) Use Chebyshev's inequality to bound P($X \geq k$).

(d) What is the exact formula for P($X \geq k$)?

(e) For $\lambda k \geq 3$, how do the bounds given in parts (a), (c), and (d) compare?

4. Robbie's Late!

Suppose the probability Robbie is late to teaching lecture on a given day is at most 0.01. Do not make any independence assumptions.

(a) Use a Union Bound to bound the probability that Robbie is late at least once over a 30-lecture quarter.

(b) Use a Union Bound to bound the probability that Robbie is **never** late over a 30-lecture quarter.

(c) Use a Union Bound to bound the probability that Robbie is late at least once over a 120-lecture quarter.

5. Mystery Dish!

A fancy new restaurant has opened up which features only 4 dishes. The unique feature of dining here is that they will serve you any of the four dishes randomly according to the following probability distribution: give dish A with probability 0.5, dish B with probability θ, dish C with probability 2θ, and dish D with probability $0.5 - 3\theta$

Each diner is served a dish independently. Let x_A be the number of people who received dish A, x_B the number of people who received dish B, etc, where $x_A + x_B + x_C + x_D = n$. Find the $\hat{\theta}$, the maximum likelihood estimator for θ.

6. A Red Poisson

Suppose that $x_1, \ldots, x_n$ are i.i.d. samples from a Poisson(θ) random variable, where θ is unknown. Find the MLE for θ.

7. Independent Shreds, You Say?

You are given 100 independent samples $x_1, x_2, \ldots, x_{100}$ from Bernoulli(θ), where θ is unknown. (Each sample is either a 0 or a 1). These 100 samples sum to 30. You would like to estimate the distribution's parameter θ. Give all answers to 3 significant digits.

(a) What is the maximum likelihood estimator $\hat{\theta}$ of θ?

(b) Is $\hat{\theta}$ an unbiased estimator of θ?

8. Y Me?

Let $y_1, y_2, ...y_n$ be i.i.d. samples of a random variable with density function

$$f_Y(y\,;\,\theta) = \frac{1}{2\theta} \exp\left(-\frac{|y|}{\theta}\right) .$$

Find the MLE for θ in terms of $|y_i|$ and n.

9. Bird Watching

You are an ornithologist studying a rare species of birds in a nature reserve. Over a period of 50 days, you record the number of sightings of this bird (you see $x_1, x_2, ..., x_{50}$ birds on each day). Your research has shown that the number of sightings on this species depends on the average number of monkeys in the reserve, θ_1, and the average number of eagles in the reserve, θ_2. After years of studying this rare species in other environments, you've found the number of birds observed on a particular day follows the following distribution:

$$p_X(k) = \frac{1}{k!}(\theta_1^k \cdot e^{-\theta_1} \cdot \theta_2^k \cdot e^{-3\theta_2})$$

Find the MLE for θ_1 and θ_2 (i.e., find $\hat{\theta_1}$ and $\hat{\theta_2}$).

(a) What is the likelihood function?

(b) What is the log-likelihood function?

(c) We want to find values of θ_1 and θ_2 that maximize the likelihood function. To do this, we will take the partial derivative with respective to each of these parameters and solve for the values that make them both zero. First, take the partial derivative of the likelihood function with respect to θ_1.

(d) Now, take the partial derivative with respect to θ_2.

(e) Set both these partial derivatives to 0, and solve for $\hat{\theta_1}$ and $\hat{\theta_2}$.

10. A biased estimator

In class, we showed that the maximum likelihood estimate of the variance θ_2 of a normal distribution (when both the true mean μ and true variance σ^2 are unknown) is what's called the *population variance*. That is

$$\hat{\theta}_2 = \left(\frac{1}{n} \sum_{i=1}^{n} (x_i - \hat{\theta}_1)^2) \right)$$

where $\hat{\theta}_1 = \frac{1}{n} \sum_{i=1}^{n} x_i$ is the MLE of the mean. Is $\hat{\theta}_2$ unbiased?

11. It Means Nothing

Suppose $x_1, x_2, \ldots, x_n$ are samples from a normal distribution whose mean is known to be μ but the variance is unknown. How does the maximum likelihood estimator for the variance differ from the maximum likelihood estimator when both mean and variance are unknown? Which if any is unbiased?

12. Covariance Connection

Let X be the network connection status, where $X = 0$ represents a stable connection and $X = 1$ represents an unstable connection. Let Y be the number of successes in data transmission, taking values in the set $\{0, 1, 2\}$. If $X = 0$, Y follows a Binomial distribution Bin$(2, 0.8)$, and if $X = 1$, Y follows a Binomial distribution Bin$(2, 0.3)$. The probabilities for X are given by $P(X = 0) = 0.8$ and $P(X = 1) = 0.2$. Find Cov(X, Y). (note that we don't know that X and Y are independent here!)