CS 109 — Final Exam Review

Adithya Ganesh

December 20, 2018

1 Key Topics

1. Balls and urns
   (a) k distinguishable objects to n distinguishable buckets:
   $$n^k.$$
   (b) k indistinguishable objects to n distinguishable buckets. If each bucket gets a positive number of objects:
   $$\binom{k-1}{n-1}.$$
   (c) If each bucket gets a nonnegative number of objects:
   $$\binom{n-1+k}{n-1}.$$
2. Balls and urns: Ordered vs. unordered set
   (a) Unordered interpretation: k people each get a set of objects
   (b) Ordered interpretation: 1 person gets a series of sets of objects
3. Bayes Theorem
   $$P(B|A) = \frac{P(A|B)P(B)}{P(A)} = \frac{P(A|B)P(B)}{P(A|B)P(B) + P(A|\neg B)P(\neg B)}.$$
   Typically use the second version for computation.
4. Principle of inclusion - exclusion
   $$\left|\bigcup_{i=1}^{n} A_i\right| = \sum_{k=1}^{n} (-1)^{k+1} \left(\sum_{1 \le i_1 < \cdots < i_k \le n} |A_{i_1} \cap \cdots \cap A_{i_k}| \right).$$
5. Computing CDF in terms of Φ:
   $$P(X \le x) = P(\frac{X-\mu}{\sigma} \le \frac{x-\mu}{\sigma}) = P(Z \le \frac{x-\mu}{\sigma}) = \Phi(\frac{x-\mu}{\sigma}).$$
6. Expectation properties
   (a) Definition
   $$\mathbb{E}[X] = \sum_x x p_X(x).$$
   $$\mathbb{E}[X] = \int_x x p(x) dx.$$

More generally, you can compute

$$\mathbb{E}[g(x)] = \int_{-\infty}^{\infty} g(x)p(x)dx.$$

(b) Linearity

$$\mathbb{E}[f(X) + g(X)] = \mathbb{E}[f(X)] + \mathbb{E}[g(X)].$$

7. Variance properties

(a) Definition

$$\text{Var}(X) = \mathbb{E}[(X - \mu)^2]$$

(b) Key identity

$$\text{Var}(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$$

(c) Linear combinations

$$\text{Var}(aX + b) = a^2 \,\text{Var}(X)$$

(d) Sums

$$\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y) + 2\,\text{Cov}(X, Y).$$

(e) Standard deviation

$$\text{SD}(X) = \sqrt{\text{Var}(X)}.$$

8. Covariance properties

(a) Definition

$$\text{Cov}(X, Y) = \mathbb{E}[(X - \mu_X)(Y - \mu_Y)] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y].$$

(b) Sum of variance

$$\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y) + 2\,\text{Cov}(X, Y).$$

(c) If X, Y independent, then $\text{Cov}(X, Y) = 0$.

(d) If X, Y independent, then

$$\mathbb{E}[f(X)g(Y)] = \mathbb{E}[f(X)]\mathbb{E}[g(Y)].$$

9. Correlation of X and Y:

$$\rho(X, Y) = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)\,\text{Var}(Y)}}$$

10. Key distributions

Discrete:

(a) $X \sim \mathit{Bernoulli}(p)$, $0 \leq p \leq 1$. 1 if coin with heads probability p comes up heads, zero otherwise.

$$p(x) = \begin{cases} p; & x = 1; \\ 1 - p; & x = 0. \end{cases}$$

$$\mathbb{E}[X] = p; \qquad \text{Var}(X) = p(1 - p).$$

(b) $X \sim \mathit{Binomial}(n, p)$, $0 \leq p \leq 1$. The number of heads in n independent flips of a coinw ith heads probability p.

$$p(x) = \binom{n}{x} p^x (1 - p)^{n-x}$$

$$\mathbb{E}[X] = np; \qquad \text{Var}(X) = np(1 - p).$$

(c) $X \sim Geometric(p)$, $p > 0$. The number of flips of a coin with heads probability p until the first heads.

$$p(x) = p(1-p)^{x-1}.$$

$$\mathbb{E}[X] = \frac{1}{p}; \qquad \text{Var}(X) = \frac{1-p}{p^2}.$$

(d) $X \sim Poisson(\lambda)$, $\lambda > 0$. A probability distribution over the nonnegative integers used for the modeling the frequency of rare events.

$$p(x) = e^{-\lambda}\frac{\lambda^x}{x!}.$$

$$\mathbb{E}[X] = \lambda; \qquad \text{Var}(X) = \lambda.$$

Intuition: let $n \to \infty, p \to 0$, and let $np = \lambda$ stay constant. Binomial distribution will converge to this density function.

The binomial in the limit, with $\lambda = np$, when n is large, p is small, and λ is "moderate"

Let X be binomial. Then if $p = \lambda/n$, we obtain

$$P(X = i) = \frac{n!}{i!(n-i)!}p^i(1-p)^{n-i} = \frac{n!}{i!(n-i)!}\left(\frac{\lambda}{n}\right)^i\left(1-\frac{\lambda}{n}\right)^{n-i}$$

$$= \frac{n(n-1)\dots(n-i+1)}{n^i}\frac{\lambda^i}{i!}\frac{(1-\lambda/n)^n}{(1-\lambda/n)^i}.$$

When n is large, p is small, and λ is moderate, we obtain

$$\frac{n(n-1)\dots(n-i+1)}{n^i} \approx 1; \qquad (1-\lambda/n)^n \approx e^{-\lambda}; \qquad (1-\lambda/n)^i \approx 1.$$

Recall that the definition of e is

$$e = \lim_{n\to\infty}(1+1/n)^n.$$

Understand how this derivation works with the exponential term

It follows that

$$P(X = i) \approx \frac{\lambda^i}{i!}e^{-\lambda}.$$

Continuous:

(a) $X \sim Uniform(a, b)$, $a < b$. Equal probability density to every value between a and b on the real line.

$$f(x) = \begin{cases} \frac{1}{b-a}; & a \le x \le b \\ 0; & \text{else.} \end{cases}$$

$$\mathbb{E}[X] = \frac{a+b}{2}; \qquad \text{Var}(X) = \frac{(b-a)^2}{12}.$$

(b) $X \sim Exponential(\lambda)$, $\lambda > 0$. Decaying probability density over the nonnegative reals.

$$f(x) = \begin{cases} \lambda e^{-\lambda x}; & x \ge 0 \\ 0; & \text{else.} \end{cases}$$

$$\mathbb{E}[X] = \frac{1}{\lambda}; \qquad \text{Var}(X) = \frac{1}{\lambda^2}.$$

(c) $X \sim Normal(\mu, \sigma^2)$. Gaussian distribution.

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right).$$

$$\mathbb{E}[X] = \mu; \qquad \mathrm{Var}(X) = \sigma^2.$$

11. Moment generating function (MGF) of X:

$$M(t) = \mathbb{E}[e^{tX}].$$

Intuition: uniquely determines the distribution. Can differentiate to compute useful quantities

12. Joint MGF of $X_1, X_2, \dots, X_n$:

$$M(t_1, t_2, \dots, t_n) = \mathbb{E}[e^{t_1 X_1 + t_2 X_2 + \cdots + t_n X_n}]$$

13. Markov's inequality. Let X be non-negative RV:

$$P(X \geq a) \leq \frac{\mathbb{E}[X]}{a}; \qquad \text{for all } a > 0.$$

Proof - indicator random variables.

14. Chebyshev's Inequality. Let X be an RV with $\mathbb{E}[X] = \mu, \mathrm{Var}(X) = \sigma^2$. Then

$$P(|X-\mu| \geq k) \leq \frac{\sigma^2}{k^2}; \qquad \text{for all } k > 0.$$

Proof, apply Markov's Inequality with $a = k^2$.

15. One-sided Chebyshev's Inequality. Let X be an RV with $\mathbb{E}[X] = 0, \mathrm{Var}(X) = \sigma^2$. Then

$$P(X \geq a) \leq \frac{\sigma^2}{\sigma^2 + a^2}.$$

Proof. Note that $P(X \geq a) = P(X + b \geq a + b)$, and apply Markov's inequality. Minimize the resulting quadratic as a function of offset b.

Or, if $\mathbb{E}[Y] = \mu$, and $\mathrm{Var}(Y) = \sigma^2$, we obtain

$$P(Y \geq \mathbb{E}[Y] + a) \leq \frac{\sigma^2}{\sigma^2 + a^2}; \qquad \text{for any } a > 0$$

$$P(Y \leq \mathbb{E}[Y] - a) \leq \frac{\sigma^2}{\sigma^2 + a^2} \qquad \text{for any } a > 0.$$

16. Chernoff bound. Let $M(t)$ be an MGF of RV X. Then

$$P(X \geq a) \leq e^{-ta} M(t); \qquad \text{for all } t > 0.$$

$$P(X \leq a) \leq e^{-ta} M(t); \qquad \text{for all } t < 0.$$

Bounds hold for $t \neq 0$, so use t that minimizes $e^{-ta}M(t)$ (i.e. makes bound strictest).

Proof: $P(X \geq a) = P(e^{tX} \geq e^{ta})$, and then apply Markov's inequality.

17. Jensen's Inequality. If $f(x)$ is convex, then

$$\mathbb{E}[f(X)] \geq f(\mathbb{E}[X]).$$

Equality when $f''(x) = 0$. Proof: Taylor series of $f(x)$ about μ.

18. Law of Large Numbers. Consider I.I.D. random variables $X_1, X_2, \dots$. Suppose $\mathbb{E}[X_i] = \mu$ and $\text{Var}(X_i) = \sigma^2$. Let $\overline{X} = \frac{1}{n}\sum_{i=1}^n X_i$. For any $\epsilon > 0$:

$$P(|\overline{X} - \mu| \geq \epsilon) \to 0.$$

Proof: Apply Chebyshev's inequality on $\overline{X}$. $\mathbb{E}[\overline{X}] = \mu, \text{Var}(\overline{X}) = \frac{\sigma^2}{n}$.

$$P(|\overline{X} - \mu| \geq \epsilon) \leq \frac{\sigma^2}{n\epsilon^2} \to 0.$$

19. Strong Law of Large Numbers. Consider I.I.D. random variables $X_1, X_2, \dots$. Suppose X_i has distribution F with $\mathbb{E}[X_i] = \mu$.

Then
$$P\left(\lim_{n\to\infty}\left[\frac{X_1 + X_2 + \cdots + X_n}{n} = \mu\right]\right) = 1.$$

20. Central Limit Theorem (CLT). Consider I.I.D. random variables $X_1, X_2, \dots$. Suppose $\mathbb{E}[X_i] = \mu$, and $\text{Var}(X_i) = \sigma^2$. Then

$$\frac{X_1 + X_2 + \cdots + X_n - n\mu}{\sigma\sqrt{n}} \to \mathcal{N}(0,1); \qquad \text{as } n \to \infty.$$

Intuition – the $n\mu$ is for mean normalization, the $\sigma\sqrt{n}$ is for variance normalization. This is why many real world distributions look normally distributed.

21. Method of moments. Let $\hat{m}_k = \frac{1}{n}\sum_{i=1}^n X_i^k$ (sample moments). Set each of these sample moments equal to the "true" moments.

22. Estimator Bias. Defined as

$$\mathbb{E}[\hat{\theta}] - \theta.$$

When bias = 0, estimator is unbiased.

23. Estimator Consistency. Defined as

$$\lim_{n\to\infty} P(|\hat{\theta} - \theta| < \epsilon) = 1; \qquad \text{for } \epsilon > 0.$$

24. Maximum Likelihood Estimation. Define the likelihood function as

$$L(\theta) = \prod_{i=1}^n f(X_i|\theta),$$

where this is a product since the X_i are IID. Then

$$\theta_{MLE} = \arg\max_\theta L(\theta).$$

25. Log-likelihood

$$LL(\theta) = \log L(\theta) = \sum_{i=1}^n \log f(X_i|\theta).$$

26. Bayesian Estimation. Let θ = model parameters, D = data. Then

$$P(\theta|D) = \frac{P(D|\theta)P(\theta)}{P(D)}.$$

We have prior $P(\theta)$ and can compute likelihood $P(D|\theta)$. Posterior $P(\theta|D)$ is assumed to have same parameter form as prior. The term $P(D)$ is a constant that can be ignored (just for integration).

Example: Let $\theta \sim \text{Beta}(a,b)$, $D = \{n \text{ heads}, m \text{ tails}\}$. Then maximum a posteriori will give you $\text{Beta}(a+n, b+m)$.

27. Maximum A Posteriori (MAP) estimator of θ:

$$\theta_{MAP} = \arg\max_\theta f(\theta|X_1, X_2, \ldots, X_n) = \arg\max_\theta \frac{f(X_1, X_2, \ldots, X_n|\theta)g(\theta)}{h(X_1, X_2, \ldots, X_n)}$$

$$= \arg\max_\theta \frac{\left(\prod_{i=1}^n f(X_i|\theta)\right) g(\theta)}{h(X_1, X_2, \ldots, X_n)} = \arg\max_\theta g(\theta) \prod_{i=1}^n f(X_i|\theta).$$

28. Log a posteriori

$$\theta_{MAP} = \arg\max_\theta \left(\log(g(\theta)) + \sum_{i=1}^n \log(f(X_i|\theta)) \right).$$

29. Naive Bayes. Estimate probabilities $P(Y)$ and each $P(X_i|Y)$ for all i. Classify as spam or not using $\hat{Y} = \arg\max_y \hat{P}(\mathbf{X}|Y)\hat{P}(Y)$.

    Employ conditional independence assumption:

$$\hat{P}(\mathbf{X}|Y) = \prod_{i=1}^m \hat{P}(X_i|Y).$$

30. Laplace estimate, Naive Bayes.

$$P(X_i = 1|Y = \text{spam}) = \frac{(\text{ spam emails with word } i) + 1}{\text{total spam emails} + 2}.$$

31. Logistic regression. Learn weights β_i to estimate

$$P(Y = 1|\mathbf{X}) = \frac{1}{1 + e^{-z}}; \qquad z = \beta^T x.$$

    Learn weights β_i from gradient descent.

32. Linear congruential generator. Start with seed number X_0. Next random number is given by

$$X_{n+1} = (aX_n + c) \pmod m.$$

33. Bayesian network. Graphical representation of joint probability distribution. Each node X has a conditional probability $P(X|parents(X))$. Graph has no cycles (directed acylic graph).

34. Showing two distributions are independent. If

$$P(x, y) = P(x)P(y); \qquad \forall x, y.$$

    then the two random variables are independent.

2 Theory

1.

3 Problems to Review

3.1 Problem Set 1

1. Classical combinatorics.
2. Balls and urns, and variations.
3. 1.13 - Unordered vs. ordered ways of counting a set for probability.

3.2 Problem Set 2

1. Basic applications of Bayes' Theorem.
2. Principle of inclusion - exclusion.
3. Classical combinatorics.

3.3 Problem Set 3

1. Infinite summations to compute expectation (typically, arithmetico-geomtric series).
2. CDF of normal in terms of Φ.
3. Binary random variable + sum of expectations.

3.4 Problem Set 4

1. Multiple integrals of a density function
2. Independence of two distributions + joint density

3.5 Problem Set 5

1. Recursive expectation calculation
2. MGF calculation

3.6 Problem Set 6

1. 6.1 - Confidence intervals
2. 6.2 - Maximum likelihood estimation + Jensen for bias

4 Practice Problems 3-20-17

1. (See notebook), went through all problems from final review document.
2. PS5.1(a)
3. PS5.4 – the relationship between independence, correlation, and covariance
4. PS5.8, using MGFs to obtain the distribution
5. PS3.3(a)
6. Problem from midterm that uses infinite series

5 Practice Final 3-21-17

1. Remember to take $\sqrt{\sigma^2}$ when doing Φ transformation.
2. Review continuity correction