MATH 408 Final Review

Qilin Ye

December 3, 2021

Main Concepts

(1) **Probability space**: let $(\mathbb{P}, \mathcal{F}, \mathbb{P})$ be a probability space ($\mathbb{P}$ the probability law and Ω the sample space). Then $\mathbb{P}$ needs to satisfy three conditions:

(i) $\mathbb{P}(A) \geqslant 0$ for all $A \in \mathcal{F}$, i.e., events have a nonnegative probability of happening;

(ii) if $A_1, A_2, \ldots$ satisfy $A_i \cap A_j = \varnothing$ for $i \neq j$, then $\mathbb{P}(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} \mathbb{P}(A_i)$; and

(iii) $\mathbb{P}(\Omega) = 1$.

(2) **Conditional**: $\mathbb{E}(X \mid Y)$, a random variable, can be viewed as a function $g(Y)$ defined by $g(y) := \mathbb{E}(X \mid Y = y)$.

(3) **Modes of convergence**: from strongest to weakest: almost surely (a.e.) convergence, convergence in probability, and convergence in distribution. They appear in the Strong LLN, Weak LLN, and CLT, respectively. See the theorem section.

(4) **Sampling from the normal** (3.7): let $X_1, \ldots, X_n$ be i.i.d. Gaussians $\sim \mathcal{N}(\mu, \sigma^2)$. The **sample mean** and **sample variance** are

$$\overline{X} := \frac{1}{n}\sum_{i=1}^{n} X_i \qquad \text{and } S^2 := \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2.$$

(i) $\overline{X} \sim \mathcal{N}(\mu, \sigma^2/n)$;

(ii) $\overline{X}$ and S^2 are independent; and

(iii) $(n-1)S^2/\sigma^2 \sim \chi^2_{n-1}$.

(5) **Convolution**: let $f, g : \mathbb{R} \to \mathbb{R}$. The convolution product $f * g$ is defined by

$$(f * g)(x) = \int_{-\infty}^{\infty} f(t)g(x-t)\,\mathrm{d}t.$$

Interpretation in probability: if X and Y are independent, then the PDF/PMF of $X + Y$ is the convolution of those of X and Y:

$$\text{PMF: } \mathbb{P}(X + Y = k) = \sum_{j \in \mathbb{Z}} \mathbb{P}(X = j)(Y = k - j) \qquad \text{PDF: } f_{X+Y}(x) = \int_{-\infty}^{\infty} f_X(t) f_Y(x-t)\,\mathrm{d}t.$$

(6) **Method of Moments estimator** (4.4): define the j^{th} sample moment by $M_j := n^{-1} \sum_{i=1}^{n} X_i^j$ (which is unbiased and consistent) and let $\mu_j := \mathbb{E}X_1^j$. If $g(\theta)$ can be expressed as a function $h(\mu, ..., \mu_j)$ then $h(M_1, ..., M_j)$ is the method of moments estimator for $g(\theta)$.

- Consistent, asymptotically unbiased, not not necessarily unbiased. *Counterexample: MoM for variance is*

$$M_2 - M_1^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)^2.$$

Let $n = 1$ *and* X *the uniform on* $[0, 1]$*. Then the MoM gives* 0 *whereas the variance is actually* $1/12$*.*

(7) **Sufficient statistics** (4.9): Y is sufficient for θ if the conditional distribution of $X = (X_1, ..., X_n)$ given $Y = y$ does not depend on θ.

- The **factorization theorem** (4.12) characterizes sufficiency: $Y = t(X)$ is sufficient if and only if

$$f_\theta(x) = g_\theta(y)h(x) \qquad \text{for all } \theta.$$

(8) **Maximum likelihood estimator** (MLE, 4.32): the estimator that maximizes the **likelihood function** defined by $\ell(\theta) = \prod_{i=1}^{n} f_\theta(x_i)$, if the maximum exists.

- MLE needs not be unique: let $f_\theta(x_i) = 1_{[\theta,\theta+1]}(x_i)$ so

$$\ell(\theta) = \prod_{i=1}^{n} 1_{[\theta,\theta+1]}(x_i).$$

If $x_1 = ... = x_n = 0$ then $\ell(\theta) = 1_{\theta\in[-1,0]}$, so any $\theta \in [-1, 0]$ works as an MLE.

- MLE is consistent and has asymptotically minimal variance by Cramér-Rao. It can be biased. *For example, consider* $X_1, ..., X_n$ *i.i.d. from uniform on* $[0, \theta]$ *and we try to estimate* θ*. The likelihood function is*

$$\ell(\theta) = \theta^{-n} 1_{0\leqslant X_{(1)}\leqslant X_{(n)}\leqslant\theta}$$

so the MLE is attained at $X_{(n)}$*.* $\mathbb{E}X_{(n)} = \theta \cdot n/(n + 1) \neq \theta$*. For a derivation of the last step, see here.*

(9) **Fisher information** (4.24): $I_X(\theta) := \mathbb{E}_\theta \left(\frac{\mathrm{d}}{\mathrm{d}\theta} \log f_\theta(X)\right)^2$.

- Alternatively,

$$I_X(\theta) = \operatorname{var}_\theta \left(\frac{\mathrm{d}}{\mathrm{d}\theta} \log f_\theta(X)\right) = -\mathbb{E}_\theta \left(\frac{\mathrm{d}^2}{\mathrm{d}\theta^2} \log f_\theta(X)\right).$$

- (4.26) Fisher information is additive for independent distributions.

(10) **Significance level & UMP test** (5.8): the significance level is defined by

$$\alpha := \sup_{\theta\in\Theta_0} \beta(\theta) = \sup_{\theta\in\Theta_0} \mathbb{P}_\theta(X \in C).$$

A hypothesis test in $\mathcal{T}$, a collection of hypothesis tests, is UMP class $\mathcal{T}$ if its $\beta(\theta)$ is the largest on Θ_1 among all tests in $\mathcal{T}$.

(11) **Likelihood ratio test, generalized LR test** (5.9, 5.22): first fix $k \geqslant 0$. If $\Theta = \{\theta_0, \theta_1\}$, then the likelihood ratio test for $H_0 : \theta = \theta_0$ is a test with rejection region

$$C := \{x \in \mathbb{R}^n : f_{\theta_1}(x) > k f_{\theta_0}(x)\}.$$

Now let $\Theta, \Theta_0, \Theta_1$ be more general and let $k > 1$ (since $k \leqslant 1$ makes the claim trivial). The generalized likelihood ratio test of $H_0 : \theta \in \Theta_0$ is the test with rejection region

$$C := \{x \in \mathbb{R}^n : \sup_{\theta \in \Theta} f_\theta(x) \geqslant k \sup_{\theta \in \Theta_0} f_\theta(x)\}.$$

- If $\lambda(x) := \dfrac{\sup_{\theta \in \Theta} f_\theta(x)}{\sup_{\theta \in \Theta_0} f_\theta(x)}$, then $2 \log \lambda(x)$ converges in distribution to χ_1^2.

(12) ***p*-value** (5.17): define $p(x) := \sup_{\theta \in \Theta_0} (t(X) \geqslant t(x))$ and we say the statistic $p(X)$ is the p-value for the hypothesis test with rejection region $\{x \in \mathbb{R}^n : t(x) \geqslant c\}$ for some constant c.

- "The probability of our statistic being at least as extreme as what is observed."
- Small p-value corresponds to high confidence in rejecting the null hypothesis.

(13) **Mann-Whitney test and signed rank test** (sections 6.2 & 6.3). See notes for more detail.

(14) **ANOVA, general linear model** (section 7.1): $Y = A\beta + \epsilon$ where Y, A are given, ϵ a random Gaussian vector, and β to be estimated.

(15) **One-way ANOVA and *F*-test** (section 7.2): test to see if each "groups" have the same β_i. If

$$\begin{aligned} Y_i &= \beta_1 + \epsilon_i \qquad 1 \leqslant i \leqslant m_1 \\ Y_i &= \beta_2 + \epsilon_i \qquad m_1 \leqslant i \leqslant m_2 \\ &\dots \\ Y_i &= \beta_p + \epsilon_i \qquad m_{p-1} \leqslant i \leqslant m_p, \end{aligned}$$

we define $\overline{Y}_j := n_j^{-1} \sum_{i=n_{j-1}+1}^{m_j} Y_j$ and $\overline{\beta}$ to be the weighted average parameter $(m_p)^{-1} \sum_{j=1}^{p} n_p \beta_p$.

We also define $S_j^2 := \dfrac{\sum_{i=n_{j-1}+1}^{m_j} (Y_i - \overline{Y}_j)^2}{n_j - 1}$ and $S^2 := \dfrac{\sum_{j=1}^{p} (n_j - 1) S_j^2}{-p + \sum_{i=1}^{n} n_j}$. Then the following follows a Snedecor's F-distribution with $p - 1$ and $m_p - p$ degrees of freedom:

$$F := S^{-2} \sum_{j=1}^{p} n_j [(\overline{Y}_j - \overline{Y}) - (\beta_i - \overline{\beta})]^2$$

- Use a table to compute the p-value of null $\beta_1 = \dots = \beta_n$ where $(\beta_i - \overline{\beta})$ vanishes.

Main Theorems

(1) **Jensen's inequality** (1.91): if $\varphi : \mathbb{R} \to \mathbb{R}$ is convex, then $\varphi(\mathbb{E}X) \leqslant \mathbb{E}(\varphi(X))$.

(2) **Laws of Large Numbers** (2.10 weak & 2.11 strong):

(i) (Weak): if $X_1, \dots, X_n$ are i.i.d. with $\mathbb{E}|X_1| < \infty$, then $n^{-1} \sum_{i=1}^{n} X_i$ converges in probability to $\mathbb{E}X_1$ as $n \to \infty$:

$$\lim_{n \to \infty} \mathbb{P}\left(\left| \frac{1}{n} \sum_{i=1}^{n} X_i - \mathbb{E}X_i \right| > t \right) = 0.$$

(ii) (Strong): same assumption, then $n^{-1}\sum_{i=1}^{n} X_i$ converges almost surely to $\mathbb{E}X_1$:

$$\mathbb{P}\left(\lim_{n\to\infty} \frac{X_1 + \ldots + X_n}{n} = \mu\right) = 1.$$

(3) **The Central Limit Theorem** (2.13): if $X_1, \ldots, X_n$ are i.i.d. with $0 < \sigma^2 := \mathrm{var}(X_1) < \infty$ and $\mu := \mathbb{E}X_1 \in \mathbb{R}$,

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{X_1 + \ldots + X_n - n\mu}{\sigma\sqrt{n}} \leqslant t\right) = \mathbb{P}(Z \leqslant t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{t} e^{-^2/2}\,\mathrm{d}s \qquad \text{for all } t \in \mathbb{R}.$$

- Error bound for CLT, *the Berry-Esseén theorem* (2.30): $\sqrt{n}(\overline{X} - \mu)$ converges in distribution to $\mathcal{N}(0, \sigma^2)$.

(4) **The Delta method** (3.13): let $\theta \in \mathbb{R}$ and $Y_1, Y_2, \ldots$ i.i.d. with $\sqrt{n}(Y_n - \theta)$ converging in distribution to $\mathcal{N}(0, \sigma^2)$ with $\sigma^2 > 0$ as $n \to \infty$. Let $f : \mathbb{R} \to \mathbb{R}$ be differentiable at θ with $f'(\theta) \neq 0$. Then $\sqrt{n}(f(Y_n) - f(\theta))$ converges in distribution to $\mathcal{N}(0, \sigma^2 f'(\theta)^2)$, or equivalently $\sigma f'(\theta) \cdot \mathcal{N}(0, 1)$.

More generally (3.16), if $f'(\theta) = 0$ but $f''(\theta)$ exists and is nonzero, then $n(f(Y_n) - f(\theta))$ converges in distribution to χ_1^2 multiplied by $\sigma^2 f''(\theta)/2$, or equivalently $\sigma^2 f''(\theta)/2 \cdot \mathcal{N}(0, 1)^2$.

Even more generally, if $f^{(k)}(\theta) = 0$ for all $k < m$ and $f^{(m)}(\theta)$ exists and is nonzero, then $n^{m/2}(f(Y_n) - f(\theta))$ converges in distribution to $\sigma^m f^{(m)}(\theta)/m! \cdot \mathcal{N}(0, 1)^m$.

(5) **The Rao-Blackwell Theorem** (4.17): *conditioning an unbiased estimator on a sufficient statistic will not increase the variance*. Formally, if Y is unbiased for $g(\theta)$ and Z sufficient, assuming $\mathrm{var}_\theta(Y) < \infty$, then

$$\mathrm{var}_\theta(\mathbb{E}_\theta(Y \mid Z)) \leqslant \mathrm{var}_\theta(Y).$$

(6) **Cramér-Rao inequality** (4.28): (with the usual assumptions) if Y is unbiased then $\mathrm{var}(Y) \geqslant 1/I_X(\theta)$.

More formally and generally, if $g(\theta) = \mathbb{E}_\theta Y$, then

$$\mathrm{var}_\theta(Y) \geqslant \frac{|g'(\theta)|^2}{I_X(\theta)} \qquad \text{for all } \theta \in \Theta.$$

(7) **The Neyman-Pearson Lemma** (5.9): if $\Theta = \{\theta_0, \theta_1\}$ with H_0 being $\theta = \theta_0$ and H_1 being $\theta = \theta_1$, then a likelihood ratio test with the appropriate ratio constant k (i.e., the k that makes significance level α) is UMP for all tests with significance level $\leqslant \alpha$.