Math 541A Final Cheat Sheet

Gregory Faletto

Contents

1 Midterm 1 **3**

1.1 Limit Theorems . . . 3

1.2 Exponential Families . . . 4

1.3 Random Samples . . . 5

1.3.1 The Delta Method . . . 5

2 Midterm 2 **5**

2.1 Data Reduction . . . 5

2.1.1 Sufficient Statistics . . . 5

2.1.2 Ancillary Statistics . . . 6

2.2 Point Estimation . . . 7

2.2.1 Evaluating Estimators . . . 7

2.2.2 Efficiency of an Estimator . . . 9

2.2.3 Bayes Estimation . . . 10

3 Post-Midterm 2 (HW 6 and 7) **11**

3.1 Results from Convexity . . . 11

3.2 Point Estimation (continued) . . . 12

3.2.1 Method of Moments . . . 12

3.2.2 Maximum Likelihood Estimator . . . 12

3.2.3 EM Algorithm . . . 14

3.3 Resampling and Bias Reduction . . . 14

3.4 Some Concentration of Measure . . . 15

1 Midterm 1

1.1 Limit Theorems

Theorem 1. [**Hölder (Grimmett and Stirzaker p. p. 143, 319; Theorem 1.99 in Math 541A lecture notes) Generalization of Cauchy-Schwarz**] Let $X, Y : \Omega \to \mathbb{R}$ be random variables. For $p, q \geq 1$ satisfying $1/p + 1/q = 1$ we have

$$\mathbb{E}(|XY|) \leq (\mathbb{E}|X^p|)^{1/p}(\mathbb{E}|X^q|)^{1/q} = \|X\|_p \|Y\|_q.$$

The equality case happens only if X is a constant multiple of Y with probability 1 (see Corollary 7.3.15 in Casella and Berger). Note that the case $p = q = 2$ recovers the Cauchy-Schwarz Inequality.

Definition 1.1. Convergence in probability. $\{X_n\}$ is said to **converge in probability** to X if

- Grimmett and Strizaker definition:

$$\lim_{n\to\infty} \Pr(|X_n - X| > \epsilon) = 0, \text{ for every } \epsilon > 0$$

- More formal (from Math 541A):

$$\forall \epsilon > 0, \lim_{n\to\infty} \Pr(\{\omega \in \Omega : |X_n(\omega) - X(\omega)| > \epsilon) = 0$$

Definition 1.2. Convergence with probability 1 or almost surely. The sequence of random variables $\{X_n\}$ is said to **converge with probability 1** (or **almost surely**) to X if

$$\Pr\left(\{\omega \in \Omega : \lim_{n\to\infty} X_n(\omega) = X(\omega)\}\right) = 1$$

Definition 1.3. Convergence in r-th mean or **convergence in ℓ_p.** $X_n \to X$ **in rth mean** (or **in ℓ_p**) where $r \geq 1$ (or $0 < p \leq \infty$) if $\mathbb{E}|X_n^r| < \infty$ for all n and

$$\lim_{n\to\infty} \mathbb{E}(|X_n - X|^r) = 0$$

or if $\|X\|_p < \infty$ and

$$\lim_{n\to\infty} \|X_n - X\|_p = 0$$

Remark. Recall that $\|X\|_p := (\mathbb{E}(X)^p)^{1/p}$ if $0 < p < \infty$ and $\|X\|_\infty := \inf\{c > 0 : \Pr(|X| \leq c) = 1\}$. Note that if $p < 1$, $\|\cdot\|_p$ is no longer a norm because it does not satisfy the Triangle Inequality, but this property still holds. Convergence in rth mean is often written $X_n \xrightarrow{r} X$.

Definition 1.4. Convergence in Distribution. Let $X_1, X_2, \ldots$ have distribution functions $F_1(\cdot), F_2(\cdot), \ldots$ respectively. Then X_n is said to **converge in distribution to** X if

$$\lim_{n\to\infty} \Pr(X_n \leq u) = \Pr(X \leq u)$$

for all u at which $F_X(x) = \Pr(X \leq x)$ is continuous.

Theorem 2. Weak Law of Large Numbers (Khinchine) (541A notes Theorem 2.10). Suppose that $\{X_k\}$ is a sequence of (i) independent (ii) identically distributed random variables with (iii) constant means, i.e., $\mathbb{E}(X_k) = \mu < \infty$. Then

$$\overline{X}_k = \frac{1}{n}\sum_{k=1}^{n} X_k \xrightarrow{p} \mu$$

Remark. Used for showing consistency of MLE estimators, etc.

Theorem 3. Strong Law of Large Numbers (541A notes Theorem 2.11). Let $\{X_k\}$ be a sequence of (i) independent (ii) identically distributed random variables. Then if and only if (iii) $\mathbb{E}|X_k| < \infty$,

$$\frac{1}{n}\sum_{k=1}^{n} X_k \xrightarrow{a.s.} \mu$$

1.2 Exponential Families

Definition 1.5. Let n, k be positive integers and let μ be a *measure* on $\mathbb{R}^n$ (that is, a probability law that does not necessarily sum to 1). Let $t_1, \dots, t_k : \mathbb{R}^n \to \mathbb{R}$. Let $h : \mathbb{R}^n \to [0, \infty]$, and assume h is not identically zero. For any $w = (w_1, \dots, w_k) \in \mathbb{R}^k$, define

$$a(w) := \log\left[\int_{\mathbb{R}^n} h(x) \exp\left(\sum_{i=1}^{k} w_i t_i(x)\right) d\mu(x)\right], \quad \forall x \in \mathbb{R}^n$$

The set $\{w \in \mathbb{R}^k\}$ is called the **natural parameter space**. On this set, the function

$$f_w(x) := h(x) \exp\left(\sum_{i=1}^{k} w_i t_i(x) - a(w)\right), \quad \forall x \in \mathbb{R}^n$$

satisfies $\int_{\mathbb{R}^n} f_w(x) d\mu(s) = 1$ (by the definition of $a(w)$). So, the set of functions (which can be interpreted as probability density functions, or as probability mass functions according to μ) $\{f_w : \theta \in \Theta : a(w(\theta)) < \infty\}$ is called a **k-parameter exponential family in canonical form.**

More generally, let $\Theta \in \mathbb{R}^k$ be any set an let $w : \Theta \to \mathbb{R}^k$. We define a **$k$-parameter exponential family** to be a set of functions $\{f_\theta : \theta \in \Theta\}$, where

$$f_\theta(x) := h(x) \exp\left(\sum_{i=1}^{k} w_i(\theta) t_i(x) - a(w(\theta))\right), \quad \forall x \in \mathbb{R}^n$$

Theorem 4 (Theorem 3.4.2 from Casella and Berger). If X is a random variable in an exponential family, then

$$\mathbb{E}\left(\sum_{i=1}^{k} \frac{\partial w_i(\theta)}{\partial \theta_j} t_i(X)\right) = \frac{\partial}{\partial \theta_j} a(w(\theta)). \tag{1}$$

1.3 Random Samples

1.3.1 The Delta Method

Theorem 5 (**Delta Method, Theorem 4.14 in 541A notes, 5.5.24 in Casella and Berger**). Let $\theta \in \mathbb{R}$. Let $Y_1, Y_2, \ldots$ be random variables such that $\sqrt{n}(Y_n - \theta)$ converges in distribution to a mean zero Gaussian random variable with variance $\sigma^2 > 0$. Let $f : \mathbb{R} \to \mathbb{R}$. Assume that f' exists and is continuous, and $f'(\theta) \neq 0$. Then

$$\sqrt{n}(f(Y_n) - f(\theta)) \xrightarrow{d} \mathcal{N}(0, \sigma^2(f'(\theta))^2).$$

2 Midterm 2

Remark. Unlike Exam 1, Exam 2 will have a reference sheet at the beginning of the exam, where the following definitions will be stated: sufficient statistic, minimal sufficient, ancillary, complete, and the definition of conditional expectation as a random variable.

Exercises to do (in relevant parts of lecture notes but not assigned in HW): 6.22 in lecture notes; problems 1 - 3 in HW6.

2.1 Data Reduction

2.1.1 Sufficient Statistics

Definition 2.1 (**Sufficient Statistic; definition 5.1 in Math 541A notes**). Suppose $X_1, \ldots, X_n$ is a sample of size n from a distribution f where $f \in \{f_\theta : \theta \in \Theta\}$ is a family of distributions (such as an exponential family). Let $g : \mathbb{R}^n \to \mathbb{R}^k$ so that $Y := g(X_1, \ldots, X_n)$ is a statistic. We say that Y is a **sufficient statistic** for θ if for every $y \in \mathbb{R}^k$ and for every $\theta \in \Theta$, the conditional distribution of $(X_1, \ldots, X_n)$ given $Y = y$ (with respect to probabilities given by f_θ) does not depend on θ. That is, Y provides sufficient information to determine θ from $X_1, \ldots, X_n$.

How to show sufficiency: either use the definition (as we did in some examples; show that conditional density does not depend on θ), or use the Factorization Theorem:

Theorem 6 (**Factorization Theorem, Theorem 5.4 in 541A notes**). Suppose $X = (X_1, \ldots, X_n)$ is a random sample of size n from a distribution f where $f \in \{f_\theta : \theta \in \Theta\}$ is a family of probability density functions or probability mass functions. Let $t : \mathbb{R}^n \to \mathbb{R}^k$, so $Y := t(X_1, \ldots, X_n)$ is a statistic. Then Y is sufficient for θ if and only if there exists a nonnegative $\{g_\theta : \theta \in \Theta\}$, $h : \mathbb{R}^n \to \mathbb{R}$, $g_\theta : \mathbb{R}^k \to \mathbb{R}$ such that

$$f_\theta(x) = g_\theta(t(x))h(x), \quad \forall\, x \in \mathbb{R}^n, \ \ \forall \theta \in \Theta. \tag{2}$$

Definition 2.2 (**Minimal sufficient statistic**). Suppose $X = (X_1, \dots, X_n)$ is a random sample of size n from a distribution f where $f \in \{f_\theta : \theta \in \Theta\}$ is a family of probability density functions or probability mass functions. Let $t : \mathbb{R}^n \to \mathbb{R}^k$. Let $Y := t(X_1, \dots, X_n)$. Assume Y is sufficient of θ. Then Y is a **minimal sufficient statistic** for θ if for every statistic $Z : \Omega \to \mathbb{R}^m$ that is sufficient for θ there exists a function $\mathbb{R}^m \to \mathbb{R}^k$ such that $Y = r(Z)$.

How to prove a minimal sufficient statistic exists: pretty much always exists by Proposition 5.12.

How to show minimal sufficiency: typically, use Theorem 5.8:

Theorem 7 (**Theorem 5.8 in 541A notes**). Let $\{f_\theta : \theta \in \Theta\}$ be a family of probability density functions or probability mass functions. Let $X_1, \dots, X_n$ be a random sample from a member of the family. Let $t : \mathbb{R}^n \to \mathbb{R}^m$ and define $Y := t(X_1, \dots, X_n)$. Y is minimal sufficient if and only if the following condition holds for every $x, y \in \mathbb{R}^n$:

There exists $c(x, y) \in \mathbb{R}$ that does not depend on θ such that $f_\theta(x) = c(x, y) f_\theta(y) \quad \forall\, \theta \in \Theta$

if and only if

$t(x) = t(y)$.

Remark. If a minimal sufficient exists, it is unique up to an invertible transformation. By this uniqueness, the converse of Theorem 5.8 also holds.

2.1.2 Ancillary Statistics

Definition 2.3 (**Ancillary Statistic**). Suppose $X_1, \dots, X_n$ is a random sample of size n from a distribution f where $f \in \{f_\theta : \theta \in \Theta\}$ is a family of distributions. A statistic $Y = t(X_1, \dots, X_n), t : \mathbb{R}^n \to \mathbb{R}^m$ is **ancillary** for θ if the distribution of Y does not depend on θ.

Definition 2.4 (**Complete statistic; definition 5.16 in 541A notes**). Suppose $X_1, \dots, X_n$ is a random sample of size n from a distribution f where $f \in \{f_\theta : \theta \in \Theta\}$ is a family of distributions. A statistic $Y = t(X_1, \dots, X_n), t : \mathbb{R}^n \to \mathbb{R}^k$ is **complete** for θ if the following holds:

For any $f : \mathbb{R}^k \to \mathbb{R}$ such that $\mathbb{E}_\theta f(Y) = 0 \ \forall\, \theta \in \Theta$, it holds that $f(Y) = 0$.

How to show a statistic is complete: Use the definition (difficult except for discrete random variables where you can express expectation without θ, as in binomial).

Why do we care if a statistic is complete?

(1) **Complete sufficiency implies minimal sufficiency:**

> **Theorem 8** (**Bahadur's Theorem; Theorem 5.25 in Math 541A notes**)**.** If Y is a complete sufficient statistic for a family $\{f_\theta : \theta \in \Theta\}$ of probability densities or probability mass functions, then Y is a minimal sufficient statistic for θ.

Remark. Because it is true that if a minimal sufficient exists, it is unique up to an invertible transformation, from Bahadur's Theorem we also know that a complete sufficient statistic is unique up to an invertible mapping. However, the converse of Bahadur's Theorem is false.

(2) **With a complete sufficient statistic (and any unbiased estimator), we can find a UMRU/UMVU estimator** (see Lehmann-Scheffe below).

(3) **Complete sufficient statistics are independent from ancillary statistics:**

> **Theorem 9** (**Basu's Theorem, Theorem 5.27 in Math 541A notes**)**.** Let $Y : \Omega \to \mathbb{R}^k$ and $Z : \Omega \to \mathbb{R}^m$ be statistics. If Y is a complete sufficient statistic for $\{f_\theta : \theta \in \Theta\}$ and Z is ancillary for θ, then for all $\theta \in \Theta$, Y and Z are independent with respect to f_θ.

2.2 Point Estimation

2.2.1 Evaluating Estimators

> **Definition 2.5** (**UMVU, sometimes called MVUE (minimum variance unbiased estimator); Definition 6.3 in Math 541A Notes**)**.** Let $X_1, \dots, X_n$ be a random sample of size n from a family of distributions $\{f_\theta : \theta \in \Theta\}$. Let $g : \Theta \to \mathbb{R}$. Let $t : \mathbb{R}^n \to \mathbb{R}$ and let $Y := t(X_1, \dots, X_n)$ be an unbiased estimator for $g(\theta)$. We say that Y is **uniformly minimum variance unbiased (UMVU)** if, for any other unbiased estimator Z for $g(\theta)$, we have $\text{Var}_\theta(Y) \leq \text{Var}_\theta(Z)$ for all $\theta \in \Theta$.

Remark. The "uniform" property has to do with the fact that this inequality must hold for every $\theta \in \Theta$ (as opposed to for a particular θ, or averaged over all $\theta \in \Theta$, or something like that).

More generally, given a family of distributions $\{f_{\tilde{\theta}} : \tilde{\theta} \in \Theta\}$, we could be given a **loss function** $\ell(\theta, y) : \Theta \times \mathbb{R}^k \to \mathbb{R}$ and be asked to minimize the **risk function** $r(\theta, Y) := \mathbb{E}_{\tilde{\theta}}\big(\ell(\theta, Y)\big)$ over all possible estimators Y. In the case of mean squared error loss, we have $\ell(\theta, y) := (y - g(\theta))^2$ for all $y, \theta \in \mathbb{R}$.

> **Definition 2.6** (**Definition 6.4 in 541A notes**)**.** Let $X_1, \dots, X_n$ be a random sample of size n from a family of distributions $\{f_\theta : \theta \in \Theta\}$. Let $g : \Theta \to \mathbb{R}$. Let $t : \mathbb{R}^n \to \mathbb{R}$ and let $Y := t(X_1, \dots, X_n)$ be an unbiased estimator for $g(\theta)$. We say Y is **uniformly minimum risk unbiased** (UMRU) if for any other unbiased estimator Z for $g(\theta)$,
>
> $$r(\theta, Y) \leq r(\theta, z), \qquad \forall\ \theta \in \Theta$$

The Rao-Blackwell Theorem says that we can lower the risk of an estimator Y by conditioning on a sufficient statistic Z.

Theorem 10 (**Rao-Blackwell; Theorem 6.4 in Math 541A notes**)**.** Let Z be a sufficient statistic for $\{f_\theta : \theta \in \Theta\}$ and let Y be an estimator for $g(\theta)$. Define $W := \mathbb{E}_\theta(Y \mid Z)$. Let $\theta \in \Theta$. Then

$$\text{Var}_\theta(W) \leq \text{Var}_\theta(Y).$$

Further, let $r(\theta, y) < \infty$ and such that $\ell(\theta, y)$ is convex in y. Then

$$r(\theta, W) \leq r(\theta, Y).$$

Remark. In the in-class review, Heilman gave the variance version of Rao-Blackwell, not loss.

In practice, if Z is not complete then we might not see an improvement.

Theorem 11 (**Lehmann-Scheffe, Theorem 6.13 in Math 541A notes**)**.** Let Z be a complete sufficient statistic for a family of distributions $\{f_\theta : \theta \in \Theta\}$. Let Y be an unbiased estimator for $g(\theta)$. Define $W := \mathbb{E}_\theta(Y \mid Z)$. (Since Z is sufficient, W does not depend on θ; see Definition 2.1.) Then W is UMRU for $g(\theta)$. Further, if $\ell(\theta, y)$ is strictly convex in y for all $\theta \in \Theta$, then W is unique. In particular, W is the unique UMVU for $g(\theta)$.

Remark (**Remark 6.14 in Math 541A notes**)**.** Let $Z : \Omega \to \mathbb{R}^k$ be a complete sufficient statistic for $\{f_\theta : \theta \in \Theta\}$ and let $h : \mathbb{R}^k \to \mathbb{R}^m$. Let $g(\theta) := \mathbb{E}_\theta h(Z)$ for all $\theta \in \Theta$. Then $h(Z)$ is unbiased for $g(\theta)$, since $\mathbb{E}_\theta h(Z) = g(\theta) = \mathbb{E}_\theta(g(\theta))$. Applying Theorem 11 (Lehmann-Scheffe, Theorem 6.13 in Math 541A notes), we have

$$W := \mathbb{E}_\theta(h(Z) \mid Z) = \mathbb{E}_\theta(\mathbb{E}_\theta[h(Z) \mid h(Z)] \mid Z) = \mathbb{E}_\theta[h(Z) \mid h(Z)] = h(Z).$$

Therefore by Theorem 11 (Lehmann-Scheffe, Theorem 6.13 in Math 541A notes), $h(Z)$ is UMVU for $g(\theta)$. That is, **any function of a complete sufficient statistic is UMVU for its expected value** (see also Theorem 7.3.23 in Casella and Berger). So one way to find a UMVU is to come up with a function of a complete sufficient statistic that is unbiased for a given function $g(\theta)$.

Summary of methods for finding a UMVU estimator for $g(\theta)$:

(1) If we have a complete sufficient statistic Z:

(a) **(Condition method/Rao-Blackwell):** Follow Theorem 11 (Lehmann-Scheffe, Theorem 6.13 in Math 541A notes): find an unbiased Y and let $W := \mathbb{E}_\theta(Y \mid Z)$; this is UMVU. (problem: can be hard to find an unbiased Y.)

(b) Solve for $h : \mathbb{R}^k \to \mathbb{R}$ satisfying

$$\mathbb{E}_\theta h(Z) = g(\theta) \tag{3}$$

by Remark 6.14 in the lecture notes, $h(Z)$ is UMVU for $g(\theta)$. By "solve", consider that we have g and Z and somehow solve for the h satisfying (3). Two examples of how this works:

- Problem 3 on Midterm 2
- (From lecture notes) If Z is binomial the left side of (3) will be the sum of a bunch of numbers. Find the h values that satisfy (3), if possible.

(2) If we don't have a complete sufficient statistic:

(a) For a one-parameter family of distributions, follow the equality case of Theorem 12 (Cramer-Rao/Information Inequality, Theorem 6.23 in Math 541A Notes). That is, an unbiased estimator Y that is a constant multiple of $\frac{\mathrm{d}}{\mathrm{d}\theta}\log f_\theta(X)$ is UMVU. (Note that this avoids any discussion of complete sufficient statistics, but it's generally not easy to do.)

(b) For a multiple-parameter family of distributions, apply Theorem 6.18 in Math 541A notes, but it will probably be difficult to apply. **Important to note: we skipped over Theorem 6.18 in the in-class review.**

2.2.2 Efficiency of an Estimator

Definition 2.7 (**Fisher Information, Definition 6.19 in Math 541A notes**). Let $f \in \{f_\theta : \theta \in \Theta\}$ be a family of multivariate probability densities or probability mass functions. Assume $\Theta \subseteq \mathbb{R}$ (this is a one-parameter situation). Let X be a random variable with distribution f_θ. Define the **Fisher information** of the family to be

$$I(\theta) = I_X(\theta) := \mathbb{E}_\theta\left(\frac{\mathrm{d}}{\mathrm{d}\theta}\log f_\theta(X)\right)^2, \qquad \forall\theta \in \Theta$$

if this quantity exists and is finite.

Remark. Note that if X is continuous,

$$\mathbb{E}_\theta\left(\frac{\mathrm{d}}{\mathrm{d}\theta}\log f_\theta(X)\right) = \int_{\mathbb{R}^n}\frac{1}{f_\theta(x)}\frac{\mathrm{d}}{\mathrm{d}\theta}f_\theta(x)\cdot f_\theta(x)dx = \int_{\mathbb{R}^n}\frac{\mathrm{d}}{\mathrm{d}\theta}f_\theta(x)dx = \frac{\mathrm{d}}{\mathrm{d}\theta}\int_{\mathbb{R}^n} f_\theta(x)dx = \frac{\mathrm{d}}{\mathrm{d}\theta}1 = 0.$$

So we could have equivalently defined the Fisher information as

$$I_X(\theta) = \mathrm{Var}_\theta\left(\frac{\mathrm{d}}{\mathrm{d}\theta}\log f_\theta(X)\right)$$

Theorem 12 (**Cramer-Rao/Information Inequality, Theorem 6.23 in Math 541A Notes**). Let $X : \Omega \to \mathbb{R}^n$ be a random variable with distribution from a family of multivariable probability densities or probability mass functions $\{f_\theta : \theta \in \Theta\}$ with $\Theta \subseteq \mathbb{R}$. Let $t : \mathbb{R}^n \to \mathbb{R}$ and let $Y := t(X)$ be a statistic. For any $\theta \in \Theta$ let $g(\theta) := \mathbb{E}_\theta Y$. Then

$$\text{Var}_\theta(Y) \geq \frac{|g'(\theta)|^2}{I_X(\theta)}, \qquad \forall \theta \in \Theta.$$

In particular, if Y is unbiased for θ, then $g(\theta) = \theta$, so,

$$\text{Var}_\theta(Y) \geq \frac{1}{I_X(\theta)}, \qquad \forall \theta \in \Theta.$$

Equality occurs for some $\theta \in \Theta$ only when $\frac{\mathrm{d}}{\mathrm{d}\theta} \log f_\theta(x)$ and $Y - \mathbb{E}_\theta Y$ are multiples of each other.

Remark. For homework 7 problem 5 (Fall 2011 Qual Exam Question 1): Fisher information was not well-defined because boundary depended on θ (had an indicator variable that depended on θ).

Definition 2.8 (**Efficiency, Defintion 6.25 in Math 541A notes**). Let $X : \Omega \to \mathbb{R}$ be a random variable with distribution from a family of multivariable probability densities or probability mass functions $\{f_\theta : \theta \in \Theta\}$ with $\Theta \in \mathbb{R}$. Let $t : \mathbb{R}^n \to \mathbb{R}$ and let $Y := t(X)$ be a statistic. Define the **efficiency** of Y to be

$$\frac{1}{I_X(\theta)\text{Var}_\theta(Y)}, \qquad \forall \theta \in \Theta$$

2.2.3 Bayes Estimation

In Bayes estimation, the parameter $\theta \in \Theta$ is regarded as a random variable Ψ. The distribution of Ψ reflects our prior knowledge about the probable values of Ψ. Then, given that $\Psi = \theta$, the conditional distribution of $X \mid \Psi = \theta$ is assumed to be $\{f_\theta : \theta \in \Theta\}$, where $f_\theta : \mathbb{R}^n \to [0, \infty)$. Suppose $t : \mathbb{R}^n \to \mathbb{R}^k$ and we have a statistic $Y := t(X)$ and a loss function $\ell : \mathbb{R}^k \times \mathbb{R}^k \to \mathbb{R}$. Let $g : \Theta \to \mathbb{R}^k$.

Definition 2.9 (**Bayes estimator, Defintion 6.26 in Math 541A notes**). A **Bayes estimator** Y for $g(\theta)$ with respect to Ψ is defined such that

$$\mathbb{E}\ell(g(\Psi), Y) \leq \mathbb{E}\ell(g(\Psi), Z)$$

for all estimators Z. Here the expectation is with respect to both Ψ and Y. Note that we have not made any assumptions about bias for Y or Z.

Remark. $t(X)$ can depend on Ψ.

In order to find a Bayes estimator, it is sufficient to minimize the conditional risk.

Proposition 13 (**Proposition 6.27 in Math 541A notes**)**.** Suppose there exists $t : \mathbb{R}^k \to \mathbb{R}$ such that for almost every $x \in \mathbb{R}^n$, $Y := t(X)$ minimizes

$$\mathbb{E}(\ell(g(\Psi), Z) \mid X = x)$$

over all estimators Z. Then $t(X)$ is a Bayes estimator for $g(\theta)$ with respect to Ψ.

3 Post-Midterm 2 (HW 6 and 7)

Exercises to do (in relevant parts of lecture notes but not assigned in HW): Exercise 6.43 (wasn't assigned on homework and is pretty simple), 6.44, 6.55, 8.3.

Proofs to be comfortable with (pretty simple, could show up on final): Proposition 6.40, Proposition 6.49, Lemma 6.50, Proposition 7.2, some of the convexity results.

3.1 Results from Convexity

Definition 3.1 (**Convex function in $\mathbb{R}^n$, Math 541A Definition**)**.** Let $\phi : \mathbb{R}^n \to \mathbb{R}$. We say that ϕ is **convex** if, for any $x, y \in \mathbb{R}^n$ and for any $t \in [0, 1]$, we have

$$\phi(tx + (1-t)y) \leq t\phi(x) + (1-t)\phi(y). \tag{4}$$

Lemma 14 (Result from Math 541A Homework 2)**.** The slope of a convex function is nondecreasing. More formally, let $\phi : \mathbb{R} \to \mathbb{R}$ be a convex function. For any $x \in \mathbb{R}$, let

$$M_R := \left\{ \frac{\phi(c) - \phi(x)}{c - x} : c > x \right\}, \quad M_L := \left\{ \frac{\phi(x) - \phi(b)}{x - b} : b > x \right\}$$

be the slopes of the secant lines through ϕ using points to the right and left of x, respectively. Then for any $m \in M_R$, $p \in M_L$ we have $m \geq p$.

Theorem 15 (**Result from 541A Homework 2; equivalent conditions for convexity**)**.** Let $\phi : \mathbb{R} \to \mathbb{R}$. Then ϕ is convex if and only if: for any $y \in \mathbb{R}$, there exists a constant a and there exists a function $L : \mathbb{R} \to \mathbb{R}$ defined by $L(x) = a(x - y) + \phi(y)$, $x \in \mathbb{R}$, such that $L(y) = \phi(y)$ and such that $L(x) \leq \phi(x)$ for all $x \in \mathbb{R}$. (In the case that ϕ is differentiable, the latter condition says that ϕ lies above all of its tangent lines.)

Theorem 16 (**Global minimum of convex functions; Math 541A Homework 6 problem**)**.** Let $f : \mathbb{R}^n \to \mathbb{R}$ be a convex function. Let $x \in \mathbb{R}^n$ be a local minimum of f. Then

(a) x is a global minimum of f.

(b) If f is strictly convex, then there is at most one global minimum of f.

(c) If f is a C^1 function (all derivatives of f exist and are continuous), and $x \in \mathbb{R}^n$ satisfies $\nabla f(x) = 0$, then x is a global minimum of f.

Theorem 17 (**Jensen's Inequality, from Math 541A**). Let $X : \Omega \to [-\infty, \infty]$ be a random variable. Let $\phi : \mathbb{R} \to \mathbb{R}$ be convex. Assume that $\mathbb{E}|X| < \infty$ and $\mathbb{E}|\phi(X)| < \infty$. Then

$$\phi(\mathbb{E}X) \leq \mathbb{E}\phi(X).$$

(**Mentioned in in-class 541A review; might have been on HW?** If ϕ is strictly convex and $\mathbb{E}(\phi(X)) = \phi(\mathbb{E}(X))$ then X is almost surely constant.)

Theorem 18 (**Conditional Jensen Inequality**). Let $X, Y : \Omega \to \mathbb{R}$ be random variables that are either both discrete or both continuous. Let $\phi : \mathbb{R} \to \mathbb{R}$ be convex. Then

$$\phi(\mathbb{E}(X|Y)) \leq \mathbb{E}(\phi(X)|Y).$$

If ϕ is strictly convex, then equality holds only if X is constant on any set where Y is constant. That is, (by an Exercise from the previous homework) equality holds only if X is a function of Y.

Proposition 19. Let A be a real $m \times n$ matrix. Let $x \in \mathbb{R}^n$ and let $b \in \mathbb{R}^m$. Then the function $f \colon \mathbb{R}^n \to \mathbb{R}$ defined by $f(x) = \frac{1}{2}\|Ax - b\|^2$ is convex.

3.2 Point Estimation (continued)

3.2.1 Method of Moments

Definition 3.2 (**Method of Moments; Definition 6.32 in Lecture Notes**). Let $g : \Theta \to \mathbb{R}^k$. Let

$$M_j := \frac{1}{n}\sum_{i=1}^{n} X_i^k$$

be the jth sample moment. Suppose we want to estimate $g(\theta)$ for any $\theta \in \Theta$. Suppose there exists $h : \mathbb{R}^j \to \mathbb{R}^k$ such that $g(\theta) = h(\mu_1, \ldots, \mu_j)$. Then the estimator

$$h(M_1, \ldots, M_j)$$

is a **method of moments** estimator for $g(\theta)$.

Remark. This is kind of the most obvious thing to do, but it turns out it doesn't work very well, which is why we spent much more time on maximum likelihood estimators.

3.2.2 Maximum Likelihood Estimator

Definition 3.3 (**Maximum Likelihood Estimator; Definition 6.36 in Lecture Notes**). The **maximum likelihood estimator** (MLE) Y is the estimator maximizing the likelihood function. That is, $Y := t(X)$, $t : \mathbb{R}^n \to \mathbb{R}$ and $t(x_1, \ldots, x_n)$ is defined to be any value of $\theta \in \Theta$ that maximizes the function

$$\prod_{i=1}^{n} f_\theta(x_1)$$

if this value of θ exists.

Remark (**Remark 6.37 in Math 541A notes**)**.** Maximizing the likelihood $\ell(\theta)$ is equivalent to maximizing $\log \ell(\theta)$ since log is monotone increasing.

- **Be able to prove Proposition 6.40? Seems a little tricky, because proof requires that a local max exists, and I don't think we ever showed that was the case.**
- Review exercise 6.41, maybe 6.42
- Understand examples 6.45 - 6.48

Proposition 20 (**Functional Equivariance of the MLE; Proposition 6.49 from Lecture Notes**)**.** Let $g : \Theta \to \Theta'$ be a bijection. Suppose Y is the MLE of θ. Then $g(Y)$ is the MLE of $g(\theta)$.

Remark. The proof is pretty simple, and this is an important result, so this may be a good proof to know how to do.

Be able to prove Lemma 6.50?

Theorem 21 (**Consistency of MLE; Theorem 6.52 from Lecture Notes**)**.** Let $X_1, X_2, \ldots : \Omega \to \mathbb{R}^n$ be i.i.d. random variable with common probability density $f_\theta : \mathbb{R}^n \to [0, \infty)$. Fix $\theta \in \Theta \subseteq \mathbb{R}^m$. Suppose Θ is compact and $f_\theta(x_1)$ is a continuous function of θ for almost every $x_1 \in \mathbb{R}$. (Then the maximum of $\ell(\theta)$ exists, since it is a continuous function on a compact set.) Assume that $\mathbb{E}_\theta \sup_{\theta' \in \Theta} |\log f_{\theta'}(X_1)| < \infty$, and $\mathbb{P}_\theta \neq \mathbb{P}_{\theta'}$ for all $\theta' \neq \theta$. Then as $n \to \infty$, the MLE of θ converges in probability to the constant function θ, with respect to $\mathbb{P}_\theta$.

Theorem 22 (**Theorem 6.53 from Lecture Notes**)**.** Let $\{f_\theta : \theta \in \Theta\}$ be a family of probability density functions, so that $f_\theta : \mathbb{R}^n \to [0, \infty)$ for all $\theta \in \Theta$. Let $X_1, X_2, \ldots$ be i.i.d. such that X_1 has density f_θ. Let $\Theta \in \mathbb{R}$. Then if the following conditions hold:

(i) The set $A := \{x \in \mathbb{R} : f_\theta(x) > 0\}$ does not depend on θ;

(ii) For every $x \in \mathbb{A}$, $\frac{\partial^2}{\partial \theta^2} f_\theta(x)$ exists and is continuous in θ;

(iii) The Fisher Information $I_{X_1}(\theta)$ exists and is finite, with $\mathbb{E}_\theta \frac{\mathrm{d}}{\mathrm{d}\theta} \log f_\theta(X_1) = 0$ and

$$I_{X_1}(\theta) = \mathbb{E}_\theta \left(\frac{\mathrm{d}}{\mathrm{d}\theta} \log f_\theta(X_1) \right)^2 = -\mathbb{E}_\theta \frac{\mathrm{d}^2}{\mathrm{d}\theta^2} \log f_\theta(X_1) > 0;$$

(iv) For every θ in the interior of Θ, there exists $\epsilon > 0$ such that

$$\mathbb{E}_\theta \sup_{\theta' \in \Theta} \left| \mathbf{1}_{\{\theta' \in [\theta - \epsilon, \theta + \epsilon]\}} \frac{\mathrm{d}^2}{\mathrm{d}[\theta']^2} \log f_{\theta'}(X_1) \right| < \infty; \textit{and}$$

(v) The MLE Y_N of θ is consistent;

then for any θ in the interior of Θ, as $n \to \infty$

$$\sqrt{n}(Y_n - \theta) \xrightarrow{d} \mathcal{N}(0, I_{X_1}(\theta)^{-1})$$

with respect to $\mathbb{P}_\theta$.

Remark. Heilman mentioned in in-class review that he probably won't make us verify all the assumptions/conditions for Theorem 6.53. Also see Theorem 6.61 in lecture notes for multidimensional version.

3.2.3 EM Algorithm

Let $X : \Omega \to \mathbb{R}^n$ be a discrete or continuous random variable with distribution from a family $\{f_\theta : \theta \in \Theta\}$ where $f_\theta : \mathbb{R}^n \to [0, \infty)$ for all $\theta \in \Theta$. Let $t : \mathbb{R}^n \to \mathbb{R}^m$ be a non-invertible function (typically with $m < n$), and let $Y := t(X)$. Suppose we would like to find the MLE of θ by maximizing $\log \ell(\theta) = \log f_\theta(X)$. But we cannot directly observe the full sample X; instead we observe the "incomplete" sample Y. The EM algorithm allows us to approximate the MLE of X by conditioning on Y.

Algorithm 6.56 (Expectation-Maximization Algorithm). Initalize $\theta_0 \in \Theta$. Fix $k \geq 1$. For all $1 \leq j \leq k$, repeat the following procedure:

(1) **(Expectation)** Given θ_{j-1}, let $\phi_j(\theta) := \mathbb{E}_{\theta_{j-1}}(\log f_\theta(X) \mid Y)$, for any $\theta \in \Theta$.

(2) **(Maximization)** Set $\theta_j \in \Theta$ to maximize ϕ_j (if such a θ_j exists).

Remark (**Correction to Remark 6.57**)**.** If Y is constant, the algorithm just outputs θ_0 in one step by the Likelihood Inequality (Lemma 6.50 in lecture notes):

$$\mathbb{E}_{\theta_0}\left[\log\left(\frac{f_\theta(X)}{f_\omega(X)}\right)\middle| Y\right] \geq 0 \iff \mathbb{E}_{\theta_0}\left[\log f_\theta(X) \mid Y\right] - \mathbb{E}_{\theta_0}\left[\log f_\omega(X) \mid Y\right] \geq 0$$

has equality only when $\omega = \theta$ (if $\mathbb{P}_\theta \neq \mathbb{P}_\omega \ \forall \theta \neq \omega$). So

$$\mathbb{E}_{\theta_0}\left[\log f_{\theta_0}(X) \mid Y\right] \geq \mathbb{E}_{\theta_0}\left[\log f_\omega(X) \mid Y\right], \qquad \forall \omega \in \Theta.$$

If $Y = X$, the algorithm just outputs the MLE of X in one step:

$$\arg\max_{\theta \in \Theta}\left\{\mathbb{E}_{\theta_0}\left[\log f_\theta(X) \mid Y\right]\right\} = \arg\max_{\theta \in \Theta}\left\{\mathbb{E}_{\theta_0}\left[\log f_\theta(X)\right]\right\} = \hat{\theta}$$

where $\hat{\theta}$ is the MLE of θ.

3.3 Resampling and Bias Reduction

Definition 3.4 (**Jackknife Estimator (Definition 7.1 in Lecture Notes)**)**.** Let $X_1, X_2, \ldots : \Omega \to \mathbb{R}^m$ be i.i.d. random variables so that X_1 has distribution $f_\theta : \mathbb{R}^m \to [0, \infty), \theta \in \Theta$. Let $Y_1, Y_2, \ldots$ be a sequence of estimators for θ so that for any $n \geq 1$, $Y_n := t_n(X_1, \ldots, X_n)$ for some $t_n : \mathbb{R}^{n \cdot m} \to \Theta$. For any $n \geq 1$, define the **jackknife estimator** of Y_n to be

$$Z_n := nY_n - \frac{n-1}{n}\sum_{i=1}^{n} t_{n-1}(X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n).$$

Be able to prove Proposition 7.2

Definition 3.5 (**Bootstrap Estimator; Definition 7.5 in Lecture Notes**)**.** Let $X_1, \ldots, X_n$ be a random sample of size n. Let $m \geq 1$. We define the **bootstrap sample** $Y_1, \ldots, Y_m$ as follows. Given $X_1, \ldots, X_n$, let $Y_1, \ldots, Y_m$ be a random sample of size m from the values $\{X_1, \ldots, X_n\}$ with replacement.

3.4 Some Concentration of Measure

Probably not important? Heilman mentioned during the review that this is "not qual material" and didn't go over it much. **Except maybe Exercise 8.3?**