Contents

Part I

Calculus

Chapter 1

Limit Theory

1.1 Function

Definition 1.1.1 (Mapping)

Let $X : \Omega_1 \to \Omega_2$ be a mapping.

1. For every subset $B \in \Omega_2$, the inverse image of B is

$$X^{-1}(B) = \{\omega : \omega \in \Omega_1, X(\omega) \in B\} := \{X \in B\}.$$

2. For every class

Definition 1.1.2 (Closed Function)

A function $f : \mathbb{R}^n \to \mathbb{R}$ is said to be closed, if for each $\alpha \in \mathbb{R}$, the sublevel set

$$\{\mathbf{x} \in \operatorname{dom} f \mid f(\mathbf{x}) \leq \alpha\}, \tag{1.1}$$

is closed.

Chapter 2

Differential Calculus

Chapter 3

Integral Calculus

Part II

Real Analysis

Chapter 4

Measure Theory

4.1 Semi-algebras, Algebras and Sigma-algebras

Definition 4.1.1 (Semi-algebra)

A nonempty class of $\mathcal{S}$ of subsets of Ω is an **semi-algebra** on Ω that satisfy

1. if $A, B \in \mathcal{S}$, then $A \cap B \in \mathcal{S}$.
2. if $A \in \mathcal{S}$, then A^C is a finite disjoint union of sets in $\mathcal{S}$, i.e.,

$$A^C = \sum_{i=1}^{n} A_i, \text{where} A_i \in \mathcal{S}, A_i \cap A_j = \emptyset, i \neq j.$$

Definition 4.1.2 (Algebra)

A nonempty class $\mathcal{A}$ of subsets of Ω is an **algebra** on Ω that satisfy

1. if $A \in \mathcal{A}$, then $A^C \in \mathcal{A}$.
2. if $A_1, A_2 \in \mathcal{A}$, then $A_1 \cup A_2 \in \mathcal{A}$.

Definition 4.1.3 (σ-algebra)

A nonempty class $\mathcal{F}$ of subsets of Ω is a **σ-algebra** on Ω that satisfy

1. if $A \in \mathcal{F}$, then $A^C \in \mathcal{F}$.
2. if $A_i \in \mathcal{F}$ is a countable sequence of sets, then $\cup_i A_i \in \mathcal{F}$.

Example (Special σ-algebra).

1. **Trival σ-algebra** $:= \{\emptyset, \Omega\}$. This is smallest σ-algebra.
2. **Power Set** := all subsets of σ, denoted by $\mathcal{P}(\Omega)$. This is the largest σ-algebra.
3. **The smallest σ-algebra containing** $A \in \Omega := \{\emptyset, A, A^C, \Omega\}$.

It is easy to define (Lesbegue) measure on the semi-algebra $\mathcal{S}$, and then easily

to extend it to the algebra $\overline{\mathcal{S}}$, finally, we can extend it further t o some σ-algebra (mostly consider the smallest one containing $\mathcal{S}$).

Lemma 4.1.1

If $\mathcal{S}$ is a semi-algebra, then $\overline{\mathcal{S}} = \{\text{finite disjoint unions of sets in } \mathcal{S}\}$ is an algebra, denoted by $\mathcal{A}(\mathcal{S})$, called **the algebra generated by** $\mathcal{S}$.

Proof. Let $A, B \in \overline{\mathcal{S}}$, then $A = \sum_{i=1}^{n} A_i, B = \sum_{j=1}^{m} B_j$ with $A_i, B_i \in \mathcal{S}$.

Intersection: For $A_i \cap B_j \in \mathcal{S}$ by the definition of semi-algebra $\mathcal{S}$, thus

$$A \cap B = \sum_{i=1}^{n} \sum_{j=1}^{m} A_i \cap B_j \in \overline{\mathcal{S}}.$$

So $\overline{\mathcal{S}}$ is closed under (finite) intersection.

Complement: For DeMorgan's Law, $A_i^C \in \mathcal{S}$ by the definition of semi-algebra $\mathcal{S}$ and $\overline{\mathcal{S}}$ closed under (finite) intersection that we just shown, thus

$$A^C = (\sum_{i=1}^{n} A_i)^C = \cap_{i=1}^{n} A_i^C \in \overline{\mathcal{S}}.$$

So $\overline{\mathcal{S}}$ is closed under complement.

Union: For DeMorgan's Law and $\overline{\mathcal{S}}$ closed under (finite) intersection and complement that we just shown, thus

$$A \cup B = (A^C \cap B^C)^C \in \overline{\mathcal{S}}.$$

So $\overline{\mathcal{S}}$ is closed under (finite) union.

Hence, $\overline{\mathcal{S}}$ is an algebra. □

Theorem 4.1.1

For any class $\mathcal{A}$, there exists a unique minimal σ-algebra containing $\mathcal{A}$, denoted by $\sigma(\mathcal{A})$, called **the σ-algebra generated by $\mathcal{A}$**. In other words,

1. $\mathcal{A} \subset \sigma(\mathcal{A})$.
2. For any σ-algebra $\mathcal{B}$ with $\mathcal{A} \subset \mathcal{B}$, $\sigma(\mathcal{A}) \subset \mathcal{B}$.

and $\sigma(\mathcal{A})$ is unique.

Proof. **Existence**:

Uniqueness:

□

Example (Borel σ-algebras generated from semi-algebras)**.** 1.

4.2 Measure

Definition 4.2.1 (Measure)

Measure is a nonnegative countably additive set function, that is, a function $\mu : \mathcal{A} \to \mathbb{R}$ with

1. $\mu(A) \geq \mu(\emptyset) = 0$ for all $A \in \mathcal{A}$.
2. if $A_i \in \mathcal{A}$ is a countable sequence of disjoint sets, then $\mu(\cup_i A_i) = \sum_i \mu(A_i)$.

Definition 4.2.2 (Measure Space)

If μ is a measure on a σ-algebra $\mathcal{A}$ of subsets of Ω, the triplet $(\Omega, \mathcal{A}, \mu)$ is a **measure space**.

Remark. A measure space $(\Omega, \mathcal{A}, \mu)$ is a **probability space**, if $P(\Omega) = 1$.

Property. Let μ be a measure on a σ-algebra $\mathcal{A}$

1. **monotonieity** if $A \subset B$, then $\mu(A) \leq \mu(B)$.
2. **subadditivity** if $A \subset \cup_{m=1}^{\infty} A_m$, then $\mu(A) \leq \sum_{m=1}^{\infty} u(A_m)$.
3. **continuity from below** if $A_i \uparrow A$ (i.e. $A_1 \subset A_2 \subset \ldots$ and $\cup_i A_i = A$), then $\mu(A_i) \uparrow \mu(A)$.
4. **continuity from above** if $A_i \downarrow A$ (i.e. $A_1 \supset A_2 \supset \ldots$ and $\cap_i A_i = A$), then $\mu(A_i) \downarrow \mu(A)$.

Proof. □

Chapter 5

Lebesgue Integration

5.1 Properties of the Integral

Theorem 5.1.1 (Jensen's Inequality)

Let (Ω, A, μ) be a probability space. If f is a real-valued function that is μ-integrable, and if φ is a convex function on the real line, then:

$$\varphi\left(\int_\Omega f \,\mathrm{d}\mu\right) \leq \int_\Omega \varphi(f) \,\mathrm{d}\mu. \tag{5.1}$$

Proof. Let $x_0 = \int_\Omega f \,\mathrm{d}\mu$. Since the existence of subderivatives for convex functions, $\exists a, b \in R$, such that,

$$\forall x \in R, \varphi(x) \geq ax + b \text{ and } ax_0 + b = \varphi(x_0).$$

Then, we got

$$\int_\Omega \varphi(f) \,\mathrm{d}\mu \geq \int_\Omega af + b \,\mathrm{d}\mu = a\int_\Omega f \,\mathrm{d}\mu + b = ax_0 + b = \varphi\left(\int_\Omega f \,\mathrm{d}\mu\right).$$

□

Theorem 5.1.2 (Hölder's Inequality)

Let $(\Omega, \mathcal{F}, \mu)$ be a measure space and let $p, q \in [1, \infty]$ with $1/p + 1/q = 1$. Then, for all measurable functions f and g on Ω,

$$\int_\Omega |f \cdot g| \,\mathrm{d}\mu \leq \|f\|_p \|g\|_q. \tag{5.2}$$

Proof. □

Theorem 5.1.3 (Minkowski's Inequality)

Let $(\Omega, \mathcal{F}, \mu)$ be a measure space and let $p \in [1, \infty]$. Then, for all measurable functions f and g on Ω,

$$\|f + g\|_p \leq \|f\|_p + \|g\|_p. \tag{5.3}$$

Proof. Since $\varphi(x) = x^p$ is a convex function for $p \in [1, \infty)$. By it's definition,

$$|f+g|^p = \left|2 \cdot \frac{f}{2} + 2 \cdot \frac{g}{2}\right|^p \leq \frac{1}{2}|2f|^p + \frac{1}{2}|2g|^p = 2^{p-1}\left(|f|^p + |g|^p\right).$$

Therefore,

$$|f+g|^p < 2^{p-1}\left(|f|^p + |g|^p\right) < \infty.$$

By Hölder's Inequality (5.1.2),

$$\begin{aligned}
\|f+g\|_p^p &= \int |f+g|^p \,\mathrm{d}\mu \\
&= \int |f+g| \cdot |f+g|^{p-1} \,\mathrm{d}\mu \\
&\leq \int (|f| + |g|)|f+g|^{p-1} \,\mathrm{d}\mu \\
&= \int |f||f+g|^{p-1} \,\mathrm{d}\mu + \int |g||f+g|^{p-1} \,\mathrm{d}\mu \\
&\leq \left(\left(\int |f|^p \,\mathrm{d}\mu\right)^{\frac{1}{p}} + \left(\int |g|^p \,\mathrm{d}\mu\right)^{\frac{1}{p}}\right)\left(\int |f+g|^{(p-1)\left(\frac{p}{p-1}\right)} \,\mathrm{d}\mu\right)^{1-\frac{1}{p}} \\
&= \left(\|f\|_p + \|g\|_p\right)\frac{\|f+g\|_p^p}{\|f+g\|_p}
\end{aligned}$$

which means, as $p \in [1, \infty)$,

$$\|f + g\|_p \leq \|f\|_p + \|g\|_p.$$

When $p = \infty$,

$$a$$

□

Theorem 5.1.4 (Bounded Convergence Theorem)

Theorem 5.1.5 (Fatou's Lemma)

Theorem 5.1.6 (Monotone Convergence Theorem)

5.2 Product Measures

Theorem 5.2.1 (Fubini's Theorem)

Part III

Functional Analysis

Part IV

Matrix Theory

Chapter 6

Matrix Norms

6.1 Matrix Norms Induced by Vector Norms

Chapter 7

Matrix Decompositions

7.1 Spectral Decomposition

Definition 7.1.1 (Eigenvectors and Eigenvalues)

A (non-zero) vector $\mathbf{v}$ of dimension n is an **eigenvector** of a square $n \times n$ matrix $\mathbf{A}$, if

$$\mathbf{A}\mathbf{v} = \lambda\mathbf{v} \tag{7.1}$$

where λ is a scalar, termed the **eigenvalue** corresponding to $\mathbf{v}$.

Definition 7.1.2 (Spectral Decomposition)

For any $n \times n$ matrix with n linearly independent eigenvectors $q_i, i = 1, \ldots, n$. Then $\mathbf{A}$ can be factorized as

$$\mathbf{A} = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^{-1}$$

where $\mathbf{Q}$ is the square $n \times n$ matrix whose i-th column is the eigenvector $\mathbf{q}_i$ of $\mathbf{A}$, and $\mathbf{\Lambda}$ is the diagonal matrix whose diagonal elements are the corresponding eigenvalues, $\mathbf{\Lambda} = \lambda_i$. This factorization is called eigendecomposition or sometimes spectral decomposition.

Example (Real Symmetric Matrices)**.** As a special case, for every $n \times n$ real symmetric matrix, the eigenvalues are real and the eigenvectors can be chosen as real and orthonormal. Thus a real symmetric matrix $\mathbf{A}$ can be decomposed as

$$\mathbf{A} = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}' \tag{7.2}$$

where $\mathbf{Q}$ is an orthogonal matrix whose columns are eigenvectors of $\mathbf{A}$, and $\mathbf{\Lambda}$ is a diagonal matrix whose entries are the eigenvalues of $\mathbf{A}$.

7.2 Singular Value Decomposition

Definition 7.2.1 (Singular Value Decomposition)

For any matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$, we have

$$\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}' \tag{7.3}$$

where

- $\mathbf{U} \in \mathbb{R}^{m \times m}$ is an orthogonal matrix whose columns are the eigenvectors of $\mathbf{A}\mathbf{A}'$
- $\mathbf{V} \in \mathbb{R}^{n \times n}$ is an orthogonal matrix whose columns are the eigenvectors of $\mathbf{A}'\mathbf{A}$
- $\mathbf{\Sigma} \in \mathbb{R}^{m \times n}$ is an all-zero matrix except for the first r diagonal elements

  $$\sigma_i = \mathbf{\Sigma}_{ii}, \quad i = 1, 2, \ldots, r$$

  which are called singular values, which are the square roots of the eigenvalues of $\mathbf{A}'\mathbf{A}$ and of $\mathbf{A}\mathbf{A}'$ (these two matrices have the same eigenvalues)

Remark. We assume above that the singular values are sorted in descending order and the eigenvectors are sorted according to descending order of their eigenvalues.

Proof. Without loss of generality, we assume $m \geq n$. Since for the case $n > m$, can then be established by transposing the SVD of $\mathbf{A}'$,

$$\mathbf{A} = (\mathbf{A}')' = (\mathbf{U}'\mathbf{\Sigma}\mathbf{V})' = \mathbf{V}'(\mathbf{U}'\mathbf{\Sigma})' = \mathbf{V}'\mathbf{\Sigma}\mathbf{U}$$

For $m \geq n$, suppose $\text{rank}(A) = r$, and then $\text{rank}(\mathbf{A}'\mathbf{A}) = r$ and the spectral decomposition of $\mathbf{A}'\mathbf{A}$ be

$$\mathbf{A}'\mathbf{A}\mathbf{V} = \mathbf{V}\,\text{diag}\left(\sigma_1^2, \ldots, \sigma_r^2, 0, \ldots, 0\right)$$

where σ_i^2 are the eigenvalues of $\mathbf{A}'\mathbf{A}$ and the columns of $\mathbf{V}$, denoted $\boldsymbol{v}^{(i)}$, are the corresponding orthonormal eigenvectors.

Let

$$\mathbf{u}^{(i)} = \frac{\mathbf{A}\boldsymbol{v}^{(i)}}{\sigma_i}$$

then

$$\mathbf{A}'\mathbf{u}^{(i)} = \frac{\mathbf{A}'\mathbf{A}\boldsymbol{v}^{(i)}}{\sigma_i} = \sigma_i \boldsymbol{v}^{(i)} \Rightarrow$$
$$\mathbf{A}\mathbf{A}'\mathbf{u}^{(i)} = \sigma_i \mathbf{A}\boldsymbol{v}^{(i)} = \sigma_i^2 \mathbf{u}^{(i)}$$

implying that $\mathbf{u}^{(i)}$ are eigenvectors of $\mathbf{A}\mathbf{A}'$ corresponding to eigenvalues σ_i^2.

Since the eigenvectors $\boldsymbol{v}^{(i)}$ are orthonormal, then so are the eigenvectors $\mathbf{u}^{(i)}$

$$\left(\mathbf{u}^{(i)}\right)' \mathbf{u}^{(j)} = \frac{\left(\boldsymbol{v}^{(i)}\right)' \mathbf{A}'\mathbf{A}\boldsymbol{v}^{(j)}}{\sigma_i^2} = \left(\boldsymbol{v}^{(i)}\right)' \boldsymbol{v}^{(j)} = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases}$$

We have thus far a matrix $\mathbf{V}$ whose columns are eigenvectors of $\mathbf{A}'\mathbf{A}$ with eigenvalues σ_i^2, and a matrix $\mathbf{U}$ whose columns are r eigenvectors of $\mathbf{A}\mathbf{A}'$ corresponding to eigenvalues σ_i^2.

We augment the eigenvectors $\mathbf{u}^{(i)}, i = 1, \ldots, r$ with orthonormal vectors $\mathbf{u}^{(i)}, i = r+1, \ldots, m$ that span null $(\mathbf{A}\mathbf{A}')$, and together $\mathbf{u}^{(i)}, i = 1, \ldots, n$ are a full orthonormal set of eigenvectors of $\mathbf{A}\mathbf{A}'$ with eigenvalues σ_i^2 (with $\sigma_i = 0$ for $i > r$).

Since

$$[\mathbf{U}'\mathbf{A}\mathbf{V}]_{ij} = \left(\mathbf{u}^{(i)}\right)' \mathbf{A}\boldsymbol{v}^{(j)} = \begin{cases} \sigma_j \left(\mathbf{u}^{(i)}\right)' \mathbf{u}^{(j)} & i \leq r \\ 0 & i > r \end{cases}$$

we get

$$\mathbf{U}'\mathbf{A}\mathbf{V} = \boldsymbol{\Sigma}$$

where

$$\boldsymbol{\Sigma} = \begin{pmatrix} \operatorname{diag}(\sigma_1, \ldots, \sigma_n) \\ \mathbf{0} \end{pmatrix}, \quad \sigma_i = 0 \forall r < i \leq n$$

Consequentially, we get the desired decompositions

$$\mathbf{A} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}'$$

□

7.2.1 Relationship to Matrix Norm

Theorem 7.2.1

For any matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$,

$$\|\mathbf{A}\|_2 = \sigma_{\max}(\mathbf{A}) \tag{7.4}$$

Proof. For any matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$, the SVD implies that,

$$\|\mathbf{A}\|_2 = \sup_{\mathbf{x} \neq 0} \frac{\|\mathbf{A}\mathbf{x}\|_2}{\|\mathbf{x}\|_2} = \sup_{\mathbf{x} \neq 0} \frac{\|\mathbf{U}\boldsymbol{\Sigma}\mathbf{V}'\mathbf{x}\|_2}{\|\mathbf{x}\|_2}$$

Since $\mathbf{U}$ is unitary, that is,

$$\|\mathbf{U}\mathbf{x}\|_2^2 = \mathbf{x}^\top \mathbf{U}'\mathbf{U}\mathbf{x}^T\mathbf{x} = \|\mathbf{x}\|_2^2, \quad \forall \mathbf{x} \in \mathbb{R}^m$$

thus,

$$= \sup_{\mathbf{x}\neq 0} \frac{\|\mathbf{\Sigma V'x}\|_2}{\|\mathbf{x}\|_2}$$

Let $\mathbf{y} = \mathbf{V'x}$, and since $\mathbf{V}$ is unitary, we have

$$\|\mathbf{y}\|_2 = \|\mathbf{V'x}\|_2 = \|\mathbf{x}\|_2 = 1$$

thus,

$$= \sup_{\mathbf{y}\neq 0} \frac{\|\mathbf{\Sigma y}\|_2}{\|\mathbf{Vy}\|_2} = \sup_{\mathbf{y}\neq 0} \frac{\left(\sum_{i=1}^{r} \sigma_i^2 \left|y_i\right|^2\right)^{\frac{1}{2}}}{\left(\sum_{i=1}^{r} \left|y_i\right|^2\right)^{\frac{1}{2}}} \leq \sigma_{\max}(\mathbf{A})$$

which takes "=", if $\boldsymbol{y} = (1, 0, \ldots, 0)'$. □

Theorem 7.2.2

For any matrix $\mathbf{A} \in \mathbb{R}^{m\times n}$, suppose $\operatorname{rank}(A) = n$, then

$$\min_{\|\mathbf{x}\|_2=1} \|\mathbf{Ax}\|_2 = \sigma_n(\mathbf{A}) \tag{7.5}$$

Proof. The proof process is analogous to the above theorem. □

Remark. If $\operatorname{rank}(\mathbf{A}) < n$, then there is an $\mathbf{x}$ such that the minimum is zero.

Part V

Convex Optimization

Chapter 8

Convex Sets

8.1 Affine and Convex Sets

8.1.1 Affine Sets

A nonempty set C is said to be **affine set**, if

$$\forall x_1, x_2 \in C, \theta \in \mathbb{R}, \theta x_1 + (1-\theta)x_2 \in C.$$

Definition 8.1.1 (Affine Set)

8.1.2 Convex Sets

Definition 8.1.2 (Convex Set)

A nonempty set C is said to be **convex set**, if

$$\forall x_1, x_2 \in C, \theta \in [0,1], \theta x_1 + (1-\theta)x_2 \in C.$$

Definition 8.1.3 (Convex Hull)

The **convex hull** of said to be set C, denoted by conv C is a set of all convex combinations of points in C,

$$\text{conv } C = \{\theta_1 x_1 + \cdots + \theta_k x_k | x_i \in C; \theta_i \geq 0, i = 1, \ldots, k; \theta_1 + \cdots + \theta_k = 1\}.$$

Remark. The convex hull conv C is always convex, which is the minimal convex set that contains C.

8.1.3 Cones

Definition 8.1.4 (Cone)

A nonempty set C is said to be **cone**, if

$$\forall x \in C, \theta \geq 0, \theta x \in C.$$

Definition 8.1.5 (Convex Cone)

A nonempty set C is said to be **convex cone**, if

$$\forall x_1, x_2 \in C, \theta_1, \theta_2 \geq 0, \theta_1 x_1 + \theta_2 x_2 \in C.$$

8.2 Some Important Examples

Definition 8.2.1 (Hyperplane)

A hyperplane is defined to be $\{x | a^\top x = b\}$, where $a \in \mathbb{R}^n, a \neq 0, b \in \mathbb{R}$.

Definition 8.2.2 (Halfspace)

A hyperplane is defined to be $\{x | a^\top x \leq b\}$, where $a \in \mathbb{R}^n, a \neq 0, b \in \mathbb{R}$.

Definition 8.2.3 ((Euclidean) Ball)

A (Euclidean) ball in $\mathbb{R}^n$ with center x_c and radius r is defined to be

$$B(x_c, r) = \{x | \|x - x_c\|_2 \leq r\} = \{x_c + ru | \|u\|_2 \leq 1\},$$

where $r > 0$.

Definition 8.2.4 (Ellipsoid)

A Ellipsoid in $\mathbb{R}^n$ with center x_c is defined to be

$$\mathcal{E} = \{x | (x - x_c)^\top P^{-1} (x - x_c) \leq 1\} = \{x_c + Au | \|u_2\| \leq 1\},$$

where $P \in \mathbb{S}^n_{++}$ (symmetric positive definite).

8.3 Generalized Inequalities

8.3.1 Definition of Generalized Inequalities

Definition 8.3.1 (Proper Cone)

A cone $K \subseteq \mathbb{R}^n$ is said to be a proper cone, if

- K is convex.
- K is closed.
- K is solid (nonempty interior).
- K is pointed (contains no line).

Definition 8.3.2 (Generalized Inequalities)

The partial ordering on $\mathbb{R}^n$ defined by proper cone K, if

$$y - x \in K, \tag{8.1}$$

which can be denoted by

$$x \preceq_K y \text{ or } y \succeq_K x. \tag{8.2}$$

The strict partial ordering on $\mathbb{R}^n$ defined by proper cone K, if

$$y - x \in \operatorname{int} K, \tag{8.3}$$

which can be denoted by

$$x \prec_K y \text{ or } y \succ_K x. \tag{8.4}$$

Remark. When $K = \mathbb{R}_+$, the partial ordering $\preceq_K$ is the usual ordering $\leq$ on $\mathbb{R}$, and the strict partial ordering $\prec_K$ is the usual strict ordering $<$ on $\mathbb{R}$.

8.3.2 Properties of Generalized Inequalities

Theorem 8.3.1 (Properties of Generalized Inequalities)

A generalized inequality $\preceq_K$ has the following properties:

- Preserved under addition:
- Transitive:
- Preserved under nonnegative scaling:
- Reflexive:
- Antisymmetric:
- Preserved under limits:

A strict generalized inequality $\prec_K$ has the following properties:

Chapter 9

Convex Optimization Problems

9.1 Generalized Inequality Constraints

Definition 9.1.1 (With Generalized Inequality Constraints)

A convex optimization problem with generalized inequality constraints is defined to be

$$\begin{array}{ll} \min_x & f_0(x) \\ \text{s.t.} & f_i(x) \preceq_{K_i} 0, \quad i = 1, \ldots, m \\ & Ax = b \end{array} \tag{9.1}$$

where $f_0 : \mathbb{R}^n \to \mathbb{R}$, $K_i \in \mathbb{R}^{k_i}$ are proper conves, and $f_i : \mathbb{R}^n \to \mathbb{R}^{k_i}$ are K_i-convex.

9.1.1 Conic Form Problems

Definition 9.1.2 (Conic Form Problem)

A conic form problem is defined to be

$$\begin{array}{ll} \min & c^T x \\ \text{s.t.} & Fx + g \preceq_K 0 \\ & Ax = b \end{array} \tag{9.2}$$

9.1.2 Semidefinite Programming

9.2 Vector Optimization

Chapter 10

Unconstrained Minimization

10.1 Definition of Unconstrained Minimization

Definition 10.1.1 (Unconstrained Minimization Problem)

The unconstrained minimization problem is defined to be

$$\min_{\mathbf{x}} f(\mathbf{x}) \tag{10.1}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is convex and twice continuously differentiable.

Assume the problem is solvable, i.e., there exists an optimal point $\mathbf{x}^*$, such that,

$$f(\mathbf{x}^*) = \inf_{\mathbf{x}} f(\mathbf{x})$$

and denote it by p^*. Since f is differentiable and convex, the point $\mathbf{x}^*$ is optimal, if and only if

$$\nabla f(\mathbf{x}^*) = 0 \tag{10.2}$$

Solving (10.1) is equal to finding the solution of (10.2), thus (10.1) can be solved by analytic solution of (10.2) in a few cases, but usually can be solved by an iterative algorithm, i.e.,

$$\exists \mathbf{x}^{(0)}, \mathbf{x}^{(1)}, \ldots \in \operatorname{dom} f, \text{ s.t. } f(\mathbf{x}^{(k)}) \to p^*, \text{ as } \quad k \to \infty$$

This algorithm is terminated when $f\left(x^{(k)}\right) - p^\star \leq \epsilon$, where $\epsilon > 0$ is some specified tolerance.

Remark. The initial point $\mathbf{x}^{(0)}$ must lie in $\operatorname{dom} f$, and the sublevel set

$$S = \left\{\mathbf{x} \in \operatorname{dom} f \mid f(\mathbf{x}) \leq f(\mathbf{x}^{(0)})\right\}$$

must be closed. Any closed function (Definition 1.1.2)

Example (Quadratic Minimization)**.** The general convex quadratic minimization problem has the form

$$\min_{\mathbf{x}} \frac{1}{2}\mathbf{x}^\top \mathbf{P}\mathbf{x} + \mathbf{q}'\mathbf{x} + r \tag{10.3}$$

where $\mathbf{P} \in \mathbb{S}_+^n, \mathbf{q} \in \mathbb{R}^n$, and $r \in \mathbb{R}$. The optimality condition is

$$\mathbf{P}\mathbf{x}^* + \mathbf{q} = \mathbf{0} \tag{10.4}$$

which is a set of linear equations.

1. If $\mathbf{P} \succ 0$, exists a unique solution $\mathbf{x}^* = -\mathbf{P}^{-1}\mathbf{q}$.
2. If $\mathbf{P}$ is not positive definite, any solution of (10.4) is optimal for (10.3).
3. If (10.4) does not have a solution, then (10.3) is unbounded.

Proof.

1. Obviously.
2. Since $\mathbf{P} \not\succeq 0$, i.e.,
$$\exists \mathbf{v}, \text{ s.t. } \mathbf{v}'\mathbf{P}\mathbf{v} < 0$$
Let $\mathbf{x} = t\mathbf{v}$, we have
$$f(\mathbf{x}) = t^2 (\mathbf{v}'\mathbf{P}\mathbf{v}/2) + t(\mathbf{q}'\mathbf{v}) + r$$
which converges to $-\infty$ as $t \to \infty$.
3. Since (10.4) does not have a solution, i.e.,
$$\mathbf{q} \notin \mathcal{R}(\mathbf{P})$$
Let
$$\mathbf{q} = \tilde{\mathbf{q}} + \mathbf{v}$$
where $\tilde{\mathbf{q}}$ is the Euclidean projection of $\mathbf{q}$ onto $\mathcal{R}(\mathbf{P})$, and $\mathbf{v} = \mathbf{q} - \tilde{\mathbf{q}}$. And $\mathbf{v}$ is nonzero and orthogonal to $\mathcal{R}(\mathbf{P})$, i.e., $\mathbf{v}'\mathbf{P}\mathbf{v} = 0$. If we take $\mathbf{x} = t\mathbf{v}$, we have
$$f(\mathbf{x}) = t\mathbf{q}'\mathbf{v} + r = t(\tilde{\mathbf{q}} + \mathbf{v})'\mathbf{v} + r = t(\mathbf{v}'\mathbf{v}) + r$$
which is unbounded below.

□

Remark. The least-squares problem is a special case of quadratic minimization, that,

$$\min_{\mathbf{x}} \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2 = \mathbf{x}^\top (\mathbf{A}'\mathbf{A})\mathbf{x} - 2(\mathbf{A}'\mathbf{b})'\mathbf{x} + \mathbf{b}'\mathbf{b}$$

The optimality condition is

$$\mathbf{A}'\mathbf{A}\mathbf{x}^* = \mathbf{A}'\mathbf{b}$$

are called the normal equations of the least-squares problem.

Example (Unconstrained Geometric Programming)**.** The unconstrained geometric program in convex form

$$\min_{\mathbf{x}} f(\mathbf{x}) = \log\left(\sum_{i=1}^{m} \exp\left(\mathbf{a}_i'\mathbf{x} + b_i\right)\right)$$

The optimality condition is

$$\nabla f\left(x^*\right) = \frac{\sum_{i=1}^{m} \exp\left(\mathbf{a}_i'\mathbf{x}^* + b_i\right)\mathbf{a}_i}{\sum_{j=1}^{m} \exp\left(\mathbf{a}_j'\mathbf{x}^* + b_j\right)} = \mathbf{0}$$

which has no analytical solution, so we must resort to an iterative algorithm. For this problem, $\operatorname{dom} f = \mathbb{R}^n$, so any point can be chosen as the initial point $\mathbf{x}^{(0)}$.

Example (Analytic Center of Linear Inequalities)**.** Consider the optimization problem

$$\min_{\mathbf{x}} f(x) = -\sum_{i=1}^{m} \log\left(\mathbf{b}_i - \mathbf{a}_i^T\mathbf{x}\right)$$

where the domain of f is the open set

$$\operatorname{dom} f = \{\mathbf{x} \mid \mathbf{a}_i'\mathbf{x} < \mathbf{b}_i, i = 1, \ldots, m\}$$

Definition 10.1.2 (Strong Convexity)

10.2 General Descent Method

10.3 Gradient Descent Method

10.4 Steepest Descent Method

10.5 Newton's Method

(Smoothness Hessian) Suppose the Hessians of f are Lipschitz continuous, i.e.,

$$\|\nabla^2 f(\mathbf{x}) - \nabla^2 f(\mathbf{y})\|_2 \leq L\|\mathbf{x} - \mathbf{y}\|_2 \tag{10.5}$$

Algorithm 1: Damped Newton Method

Input: Initial point $\mathbf{x}_0 \in \mathrm{dom} f$, tolerance $\epsilon > 0$

Output:

repeat

  Compute the Newton step and decrement

$$\Delta \mathbf{x}_{\mathrm{nt}} := -\nabla^2$$

until;

Theorem 10.5.1

Under the condition, there exist $0 < \eta < m^2/L$ and $\gamma > 0$, for the damped Newton method, we have

- If $\|\nabla^2 f(\mathbf{x}^{(k)})\| \geq \eta$, then

$$f(\mathbf{x}^{(k+1)}) - f(\mathbf{x}^{(k)}) \leq -\gamma.$$

- If $\|\nabla^2 f(\mathbf{x}^{(k)})\| < \eta$, then the backtracking line search select $t^{(k)} = 1$, and

$$\frac{L}{2m^2}\|\nabla f(\mathbf{x}^{(k+1)})\|_2 \leq \left(\frac{L}{2m^2}\|\nabla f(\mathbf{x}^{(k)})\|_2\right)^2.$$

Method	Descent Direction	Step Length	Features
Steepest			
Steepest (MG)			
Steepest (CD)			
Steepest (BB)			
Newton			
Newton (LM)			
Newton (Mixed)			
Quasi-Newton (SR1)			
Quasi-Newton (DFP)			
Quasi-Newton (BFGS)			
Quasi-Newton (LBFGS)			

Example (Extended Rosenbrock Function)**.**

$$\min_{\mathbf{x}} f(\mathbf{x}) = \sum_{i=1}^{n} r_i^2(\mathbf{x}) \tag{10.6}$$

where n is even, and

$$r_i(\mathbf{x}) = \begin{cases} 10(x_{2k} - x_{2k-1}^2), & i = 2k-1 \\ 1 - x_{2k-1}, & i = 2k \end{cases} \tag{10.7}$$

The minimum point is $\mathbf{x}^* = (1, 1, \ldots, 1)'$, the initial point is $\mathbf{x}_0 = (-1.2, 1, \ldots, -1.2, 1)'$.

Chapter 11

Exercises for Convex Optimization

11.1 Convex Sets

Exercise. Solution set of a quadratic inequality Let $C \subseteq \mathbb{R}^n$ be the solution set of a quadratic inequality,

$$C = \{x \in \mathbb{R}^n | x^T A x + b^T x + c \leq 0\}$$

with $A \in \mathbb{S}^n$, $b \in \mathbb{R}^n$, and $c \in \mathbb{R}$.

1. Show that C is convex if $A \succeq 0$.

Proof. 1. We have to show that $\theta x + (1-\theta)y \in C$ for all $\theta \in [0,1]$ and $x, y \in C$.

$$\begin{aligned}
&(\theta x + (1-\theta)y)^T A(\theta x + (1-\theta)y) + b^T(\theta x + (1-\theta)y) + c \\
=&\theta^2 x^\top A x + \theta(1-\theta)(y^\top A x + x^\top A y) + (1-\theta)^2 y^\top A y + \theta b^\top x + (1-\theta) b^\top y + c \\
=&\theta^2(x^T A x + b^T x + c) + (1-\theta)^2(y^T A y + b^T y + c) - \theta^2(b^T x + c) \\
&- (1-\theta)^2(b^T y + c) + \theta(1-\theta)(y^\top A x + x^\top A y) + \theta b^\top x + (1-\theta) b^\top y + c \\
\leq& - \theta^2(b^T x + c) - (1-\theta)^2(b^T y + c) + \theta(1-\theta)(y^\top A x + x^\top A y) \\
&+ \theta b^\top x + (1-\theta) b^\top y + c \\
=&\theta(1-\theta)[(b^\top x + c) + (b^\top y + c) + x^\top A x + y^\top A y] \\
\leq&\theta(1-\theta)(-x^\top A x - y^\top A y + x^\top A x + y^\top A y) \leq 0
\end{aligned}$$

Therefore, $\theta x + (1-\theta)y \in C$, which shows that C is convex if $A \succeq 0$.

□

Part VI

Probability Theory

Chapter 12

Random Variables

12.1 Probability Space

Definition 12.1.1 (Probability Space)

A probability space is a triple $(\Omega, \mathcal{F}, P)$ consisting of:

1. the sample space Ω: an arbitrary non-empty set.
2. the σ-algebra $\mathcal{F} \subseteq 2^{\Omega}$: a set of subsets of Ω, called events.
3. the probability measure $P : \mathcal{F} \to [0, 1]$: a function on $\mathcal{F}$ which is a measure function.

12.2 Random Variables

Definition 12.2.1 (Random Variable)

A random variable is a measurable function $X : \Omega \to S$ from a set of possible outcomes $(\Omega, \mathcal{F})$ to a measurable space $(S, \mathcal{S})$, that is,

$$X^{-1}(B) \equiv \{\omega : X(\omega) \in B\} \in \mathcal{F} \quad \forall B \in \mathcal{S}. \tag{12.1}$$

Typically, $(S, \mathcal{S}) = (R^d, \mathcal{R}^d) \quad (d > 1)$.

How to prove that functions are measurable?

Theorem 12.2.1

If $\{\omega : X(\omega) \in A\} \in \mathcal{F}$ for all $A \in \mathcal{A}$ and $\mathcal{A}$ generates $\mathcal{S}$, then X is measurable.

1.

12.3 Distributions

12.3.1 Definition of Distributions

Definition 12.3.1 (Distribution)

A distribution of random variable X is a probability function $P : \mathcal{R} \to \mathbb{R}$ by setting

$$\mu(A) = P(X \in A) = P\left(X^{-1}(A)\right), \quad \text{for } A \in \mathcal{R}. \tag{12.2}$$

Definition 12.3.2 (Distribution Function)

The distribution of a random variable X is usually described by giving its **distribution function**,

$$F(x) = P(X \leq x). \tag{12.3}$$

Definition 12.3.3 (Density Function)

If the distribution function $F(x) = P(X \leq x)$ has the form

$$F(x) = \int_{-\infty}^{x} f(y)\,\mathrm{d}y,$$

that X has density function f.

12.3.2 Properties of Distributions

Theorem 12.3.1 (Properties of Distribution Function)

Any distribution function F has the following properties,

1. F is nondecreasing.
2. $\lim_{x\to\infty} F(x) = 1, \lim_{x\to-\infty} F(x) = 0$.
3. F is right continuous, i.e., $\lim_{y\downarrow x} F(y) = F(x)$.
4. If $F(x-) = \lim_{y\uparrow x} F(y)$, then $F(x-) = P(X < x)$.
5. $P(X = x) = F(x) - F(x-)$.

Proof. □

Theorem 12.3.2

If F satisfies (1), (2), and (3) in Theorem 12.3.1, then it is the distribution function of some random variable.

Proof. □

Theorem 12.3.3

A distribution function has at most countably many discontinuities

Proof. □

12.3.3 Families of Distributions

Exponential Family

Definition 12.3.4 (Exponential Family)

An exponential family of probability distributions is those distributions whose density is defined to be

$$f\left(y \mid \theta, \phi\right) = \exp\left[\frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi)\right] \tag{12.4}$$

Property. The exponential family has the following properties,

$$E(Y) = b'(\theta) \quad \mathrm{Var}(Y) = b''(\theta)a(\phi).$$

Proof. □

Table 12.1: Common Distributions of Exponential Family

Distribution	Parameter (s)	θ	ϕ	$b(\theta)$	$a(\phi)$	$c(y,\phi)$	$E(Y)$	$\text{Var}(Y)$
Normal	$N(\mu,\sigma^2)$	μ	σ^2	$\frac{\theta^2}{2}$	ϕ	$-\frac{1}{2}\left[\frac{y^2}{\phi}+\log(2\pi\phi)\right]$	θ	ϕ
Bernoulli	Bern(p)	$\log\left(\frac{p}{1-p}\right)$	1	$\log\left(1+e^{\theta}\right)$	1	0	$\frac{e^{\theta}}{1+e^{\theta}}$	$\frac{e^{\theta}}{\left(1+e^{\theta}\right)^2}$
Poisson	$P(\mu)$	$\log(\mu)$	1	e^{θ}	1	$-\log(y!)$	e^{θ}	e^{θ}
Gamma	$\Gamma(\alpha,\beta)$	$\log\left(\frac{\alpha}{\beta}\right)$	1	$-\log(-\theta)$	1	$-\log\left(\Gamma(\alpha)\right)+(\alpha-1)\log(y)-y$	$\frac{\alpha}{\beta}$	$\frac{\alpha}{\beta^2}$

12.4 Expected Value

Definition 12.4.1 (Expectation)

Theorem 12.4.1 (Bounded Convergence theorem)

Theorem 12.4.2 (Fatou's Lemma)

If $X_n \geq 0$, then

$$\liminf_{n\to\infty} EX_n \geq E\left(\liminf_{n\to\infty} X_n\right). \tag{12.5}$$

Theorem 12.4.3 (Monotone Convergence theorem)

If $0 \leq X_n \uparrow X$, then

$$EX_n \uparrow EX. \tag{12.6}$$

Theorem 12.4.4 (Dominated Convergence theorem)

If $X_n \to X$ a.s., $|X_n| \leq Y$ for all n, and $EY < \infty$, then

$$EX_n \to EX. \tag{12.7}$$

12.5 Independence

12.5.1 Definition of Independence

Definition 12.5.1 (Independence)

1. Two events A and B are independent if $P(A \cap B) = P(A)P(B)$.
2. Two random variables X and Y are independent if for all $C, D \in \mathcal{R}$

$$P(X \in C, Y \in D) = P(X \in C)P(Y \in D). \tag{12.8}$$

3. Two σ-fields $\mathcal{F}$ and $\mathcal{G}$ are independent if for all $A \in \mathcal{F}$ and $B \in \mathcal{G}$ the events A and B are independent.

The second definition is a special case of the third.

Theorem 12.5.1

1. If X and Y are independent then $\sigma(X)$ and $\sigma(Y)$ are independent.
2. Conversely, if $\mathcal{F}$ and $\mathcal{G}$ are independent, $X \in \mathcal{F}$ and $Y \in \mathcal{G}$, then X and Y are independent.

The first definition is, in turn, a special case of the second.

Theorem 12.5.2

1. If A and B are independent, then so are A^c and B, A and B^c, and A^c and B^c.
2. Conversely, events A and B are independent if and only if their indicator random variables 1_A and 1_B are independent.

The definition of independence can be extended to the infinite collection.

Definition 12.5.2

An infinite collection of objects (σ-fields, random variables, or sets) is said to be independent if every finite subcollection is,

1. σ-fields $\mathcal{F}_1, \mathcal{F}_2, \ldots, \mathcal{F}_n$ are independent if whenever $A_i \in \mathcal{F}_i$ for $i = 1, \ldots, n$, we have
$$P\left(\cap_{i=1}^n A_i\right) = \prod_{i=1}^n P\left(A_i\right). \tag{12.9}$$
2. Random variables $X_1, \ldots, X_n$ are independent if whenever $B_i \in \mathcal{R}$ for $i = 1, \ldots, n$ we have
$$P\left(\cap_{i=1}^n \left\{X_i \in B_i\right\}\right) = \prod_{i=1}^n P\left(X_i \in B_i\right). \tag{12.10}$$
3. Sets $A_1, \ldots, A_n$ are independent if whenever $I \subset \{1, \ldots, n\}$ we have
$$P\left(\cap_{i \in I} A_i\right) = \prod_{i \in I} P\left(A_i\right). \tag{12.11}$$

12.5.2 Sufficient Conditions for Independence

12.5.3 Independence, Distribution, and Expectation

Theorem 12.5.3

Suppose $X_1, \ldots, X_n$ are independent random variables and X_i has distribution μ_i, then $(X_1, \ldots, X_n)$ has distribution $\mu_1 \times \cdots \times \mu_n$.

Theorem 12.5.4

If $X_1, \ldots, X_n$ are independent and have

1. $X_i \geq 0$ for all i, or
2. $E|X_i| < \infty$ for all i.

then

$$E\left(\prod_{i=1}^{n} X_i\right) = \prod_{i=1}^{n} EX_i \tag{12.12}$$

12.5.4 Sums of Independent Random Variables

Theorem 12.5.5 (Convolution for Random Variables)

1. If X and Y are independent, $F(x) = P(X \leq x)$, and $G(y) = P(Y \leq y)$, then
$$P(X + Y \leq z) = \int F(z - y)\, \mathrm{d}G(y). \tag{12.13}$$
2. If X and Y are independent, X with density f and Y with distribution function G, then $X + Y$ has density
$$h(x) = \int f(x - y)\, \mathrm{d}G(y). \tag{12.14}$$
Suppose Y has density g, the last formula can be written as
$$h(x) = \int f(x - y)g(y)\, \mathrm{d}y. \tag{12.15}$$
3. If X and Y are independent, integral-valued random variables, then
$$P(X + Y = n) = \sum_{m} P(X = m)P(Y = n - m). \tag{12.16}$$

12.6 Moments

Lemma 12.6.1

If $Y > 0$ and $p > 0$, then

$$E(Y^p) = \int_0^\infty py^{p-1}P(Y > y)\, \mathrm{d}y. \tag{12.17}$$

12.7 Characteristic Functions

12.7.1 Definition of Characteristic Functions

Definition 12.7.1 (Characteristic Function)

If X is a random variable, we define its characteristic function (ch.f) by

$$\varphi(t) = E\left(e^{itX}\right) = E\left(\cos tX\right) + iE\left(\sin tX\right). \tag{12.18}$$

Remark. Euler Equation.

12.7.2 Properties of Characteristic Functions

Theorem 12.7.1 (Properties of Characteristic Function)

Any characteristic function has the following properties:

1. $\varphi(0) = 1$,
2. $\varphi(-t) = \overline{\varphi(t)}$,
3. $|\varphi(t)| = |Ee^{itX}| \leq E|e^{itX}| = 1$,
4. $\varphi(t)$ is uniformly continuous on $(-\infty, \infty)$,
5. $Ee^{it(aX+b)} = e^{itb}\varphi(at)$,
6. If X_1 and X_2 are independent and have ch.f. φ_1 and φ_2, then $X_1 + X_2$ has ch.f. $\varphi_1(t)\varphi_2(t)$.

Proof. □

12.7.3 The Inversion Formula

The characteristic function uniquely determines the distribution. This and more is provided by:

Theorem 12.7.2 (The Inversion Formula)

Let $\varphi(t) = \int e^{itx}\mu(\mathrm{d}x)$ where μ is a probability measure. If $a < b$, then

$$\lim_{T\to\infty}(2\pi)^{-1}\int_{-T}^{T}\frac{e^{-ita} - e^{-itb}}{it}\varphi(t)dt = \mu(a,b) + \frac{1}{2}\mu(\{a,b\}) \tag{12.19}$$

Proof. □

Theorem 12.7.3

If $\int |\varphi(t)|\,\mathrm{d}t < \infty$, then μ has bounded continuous density

$$f(y) = \frac{1}{2\pi}\int e^{-ity}\varphi(t)\,\mathrm{d}t. \tag{12.20}$$

Proof. □

12.7.4 Moments and Derivatives

Theorem 12.7.4

If $\int |x|^n \mu(dx) < \infty$, then its characteristic function φ has a continuous derivative of order n given by

$$\varphi^{(n)}(t) = \int (ix)^n e^{itx}\mu(\mathrm{d}x). \tag{12.21}$$

Theorem 12.7.5

If $E|X|^2 < \infty$ then

$$\varphi(t) = 1 + itEX - t^2 E\left(X^2\right)/2 + o\left(t^2\right). \tag{12.22}$$

Theorem 12.7.6

If $\limsup_{h\downarrow 0}\{\varphi(h) - 2\varphi(0) + \varphi(-h)\}/h^2 > -\infty$, then

$$E|X|^2 < \infty. \tag{12.23}$$

Chapter 13

Convergence of Random Variables

13.1 Modes of Convergence

13.1.1 Convergence in Mean

Definition 13.1.1 (Convergence in Mean)

A sequence $\{X_n\}$ of real-valued random variables **converges in the r-th mean** $(r \geq 1)$ towards the random variable X, if

1. The r-th absolute moments $E(|X_n|^r)$ and $E(|X|^r)$ of $\{X_n\}$ and X exist,
2. $\lim_{n\to\infty} E\left(|X_n - X|^r\right) = 0$.

Convergence in the r-th mean is denoted by

$$X_n \xrightarrow{L^r} X. \tag{13.1}$$

13.1.2 Convergence in Probability

Definition 13.1.2 (Convergence in Probability)

A sequence $\{X_n\}$ of real-valued random variables **converges in probability** towards the random variable X, if

$$\forall \varepsilon > 0, \quad \lim_{n\to\infty} P\left(|X_n - X| > \varepsilon\right) = 0. \tag{13.2}$$

Convergence in probability is denoted by

$$X_n \xrightarrow{p} X. \tag{13.3}$$

13.1.3 Convergence in Distribution

Definition 13.1.3 (Convergence in Distribution)

A sequence $\{X_n\}$ of real-valued random variables is said to **converge in distribution**, or **converge weakly**, or **converge in law** to a random variable X, if

$$\lim_{n\to\infty} F_n(x) = F(x), \tag{13.4}$$

for every number at $x \in \mathbb{R}$ which F is continuous. Here F_n and F are the cumulative distribution functions of random variables X_n and X, respectively. Convergence in distribution is denoted as

$$X_n \xrightarrow{d} X, \text{ or } X_n \Rightarrow X. \tag{13.5}$$

- Convergence in Distribution is the weakest form of convergence typically discussed since it is implied by all other types of convergence mentioned in this chapter.
- Convergence in Distribution does not imply that the sequence of corresponding probability density functions will also converge. However, according to Scheff's theorem, convergence of the probability density functions implies convergence in distribution.

Theorem 13.1.1 (Portmanteau Lemma)

$\{X_n\}$ converges in distribution to X, if and only if any of the following statements are true,

- $P(X_n \leq x) \to P(X \leq x)$, for all continuity points of the distribution of X.
- $Ef(X_n) \to Ef(X)$, for all bounded, continuous (Lipschitz) functions f.
- $\liminf_{n\to\infty} P\left(X_n \in G\right) \geq P\left(X_\infty \in G\right)$, for all open sets G.
- $\limsup_{n\to\infty} P\left(X_n \in K\right) \leq P\left(X_\infty \in K\right)$, for all closed sets K.
- $\lim_{n\to\infty} P\left(X_n \in A\right) = P\left(X_\infty \in A\right)$, for all Borel sets A with $P\left(X_\infty \in \partial A\right) = 0$.

Proof. □

Continuous Mapping Theorem

Theorem 13.1.2 (Continuous Mapping Theorem)

Let g be a measurable function and $D_g = \{x : g \text{ is discontinuous at } x\}$ with $P(X \in D_g) = 0$, then,

$$\begin{aligned} X_n \xrightarrow{d} X &\Rightarrow g(X_n) \xrightarrow{\mathrm{d}} g(X), \\ X_n \xrightarrow{p} X &\Rightarrow g(X_n) \xrightarrow{p} g(X), \\ X_n \xrightarrow{a.s.} X &\Rightarrow g(X_n) \xrightarrow{a.s.} g(X). \end{aligned} \tag{13.6}$$

If in addition g is bounded, then

$$Eg(X_n) \to Eg(X). \tag{13.7}$$

Proof. □

Slutsky's Theorem

Theorem 13.1.3 (Slutsky's Theorem)

Let X_n, Y_n be sequences of random variables. If $X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{p} c$, then

1. $X_n + Y_n \xrightarrow{d} X + c$.
2. $X_n Y_n \xrightarrow{d} cX$.
3. $X_n / Y_n \xrightarrow{d} X/c$, provided that c is invertible.

Proof. □

Remark. However that convergence in distribution of $X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{d} Y$ does in general not imply convergence in distribution of $X_n + Y_n \xrightarrow{d} X + Y$ or of $X_n Y_n \xrightarrow{d} XY$.

The Delta Methods

Theorem 13.1.4 (Delta Method)

Let $\{X_n\}$ be a sequence of random variables with

$$\sqrt{n}\left(X_n - \theta\right) \xrightarrow{d} N\left(0, \sigma^2\right) \tag{13.8}$$

where θ and σ are finite, then for any function g with the property that $g'(\theta)$ exists and is non-zero valued,

$$\sqrt{n}\left[g\left(X_n\right) - g(\theta)\right] \xrightarrow{d} N\left(0, \sigma^2 \cdot \left[g'(\theta)\right]^2\right) \tag{13.9}$$

Proof. Under the assumption that $g'(\theta)$ is continuous.

Since, $g'(\theta)$ exists, with the first-order Taylor Approximation, that

$$g(X_n) = g(\theta) + g'(\tilde{\theta})(X_n - \theta)$$

where $\tilde{\theta}$ lies between X_n and θ. Since $X_n \xrightarrow{p} \theta$, and $|\tilde{\theta} - \theta| < |X_n - \theta|$, then

$$\tilde{\theta} \xrightarrow{p} \theta$$

Since $g'(\theta)$ is continuous, by Continuous Mapping Theorem (13.1.2),

$$g'(\tilde{\theta}) \xrightarrow{p} g'(\theta)$$

and,

$$\sqrt{n}\left(g(X_n) - g(\theta)\right) = \sqrt{n} g'(\tilde{\theta})(X_n - \theta)$$
$$\sqrt{n}\left(X_n - \theta\right) \xrightarrow{d} N\left(0, \sigma^2\right)$$

by Slutsky's Theorem (13.1.3),

$$\sqrt{n}\left[g\left(X_n\right) - g(\theta)\right] \xrightarrow{d} N\left(0, \sigma^2 \cdot \left[g'(\theta)\right]^2\right)$$

□

Theorem 13.1.5 (Second-order Delta Method)

Remark. We can approximate the moments of a function $f(\cdot)$ of a random variable X using Taylor expansions, provided that $f(\cdot)$ is sufficiently differentiable and that the moments of X are finite. Suppose $\mu = \mathbb{E}(X)$, and $\sigma^2 = \mathrm{Var}(X)$, with the Taylor expansions for the functions of random variables,

$$f(X) = f[\mu + (X - \mu)] \approx f(\mu) + f'(\mu)(X - \mu) \tag{13.10}$$

Thus,

$$\mathbb{E}[f(X)] \approx \mathbb{E}[f(\mu)], \quad \mathrm{Var}[f(X)] \approx \left[f'(\mu)\right]^2 \cdot \sigma^2 \tag{13.11}$$

Lèvy' s Continuity Theorem

Theorem 13.1.6 (Lèvy' s Continuity Theorem)

Let $\mu_n, 1 \leq n \leq \infty$ be probability measures with ch.f. φ_n.

1. If $\mu_n \xrightarrow{d} \mu_\infty$, then $\varphi_n(t) \to \varphi_\infty(t)$ for all t.
2. If $\varphi_n(t)$ converges pointwise to a limit $\varphi(t)$ that is continuous at 0, then the associated sequence of distributions μ_n is tight and converges weakly to the measure μ with characteristic function φ.

Proof. □

Cramér-Wold Theorem

Theorem 13.1.7 (Cramér-Wold Theorem)

13.1.4 Almost Sure Convergence

Definition 13.1.4 (Almost Sure Convergence)

A sequence $\{X_n\}$ of real-valued random variables converges **almost sure** or **almost everywhere** or **with probability 1** or **strongly** towards the random variable X, if

$$P\left(\lim_{n\to\infty} X_n = X\right) = 1. \tag{13.12}$$

Almost sure convergence is denoted by

$$X_n \xrightarrow{a.s.} X. \tag{13.13}$$

Remark.

13.1.5 Convergence in Uninform

Definition 13.1.5 (Convergence in Uninform)

13.1.6 Asymptotic Notation

Definition 13.1.6

A sequence $\{A_n\}$ of real-valued random variables is of smaller order in probability than a sequence $\{B_n\}$, if

$$\frac{A_n}{B_n} \xrightarrow{p} 0. \tag{13.14}$$

Smaller order in probability is denoted by

$$A_n = o_p(B_n). \tag{13.15}$$

Particularly,

$$A_n = o_p(1) \iff A_n \xrightarrow{p} 0. \tag{13.16}$$

Definition 13.1.7

A sequence $\{A_n\}$ of real-valued random variables is of smaller order than or equal to a sequence $\{B_n\}$ in probability, if

$$\forall \varepsilon > 0 \; \exists M_\varepsilon, \quad \lim_{n\to\infty} P\left(|A_n| \leq M_\varepsilon |B_n|\right) \geq 1 - \varepsilon. \tag{13.17}$$

Smaller order than or equal to in probability is denoted by

$$A_n = O_p(B_n). \tag{13.18}$$

Definition 13.1.8

A sequence $\{A_n\}$ of real-valued random variables is of the same order as a sequence $\{B_n\}$ in probability, if

$$\forall \varepsilon > 0 \; \exists m_\varepsilon < M_\varepsilon, \quad \lim_{n\to\infty} P\left(m_\varepsilon < \frac{|A_n|}{|B_n|} < M_\varepsilon\right) \geq 1 - \varepsilon. \tag{13.19}$$

The same order in probability is denoted by

$$A_n \asymp_p B_n. \tag{13.20}$$

13.2 Relationships of Modes

Lemma 13.2.1

If $p > 0$ and $E|Z_n|^p \to 0$, then

$$Z_n \xrightarrow{p} 0. \tag{13.21}$$

Proof. □

Theorem 13.2.1

If $X_n \xrightarrow{p} X$, then

$$X_n \xrightarrow{d} X, \tag{13.22}$$

and that, conversely, if $X_n \xrightarrow{d} c$, where c is a constant, then

$$X_n \xrightarrow{p} c. \tag{13.23}$$

Proof. 1. $\forall \varepsilon > 0$, at fixed point x, since if $X_n \leq x$ and $|X_n - X| \leq \varepsilon$, then $X \leq x + \varepsilon$, then

$$\{X \leq x + \varepsilon\} \subset \{X_n \leq x\} \cup \{|X_n - X| > \varepsilon\},$$

similarly, if $X \leq x - \varepsilon$ and $|X_n - X| \leq \varepsilon$, then $X_n \leq x$, then

$$\{X_n \leq x\} \subset \{X \leq x - \varepsilon\} \cup \{|X_n - X| > \varepsilon\},$$

then, by the union bound,

$$\begin{aligned} P(X \leq x + \varepsilon) &\leq P(X_n \leq x) + P(|X_n - X| > \varepsilon), \\ P(X_n \leq x) &\leq P(X \leq x - \varepsilon) + P(|X_n - X| > \varepsilon). \end{aligned}$$

So, we got

$$\begin{aligned} P(X \leq x + \varepsilon) - P(|X_n - X| > \varepsilon) &\leq P(X_n \leq x) \\ &\leq P(X \leq x - \varepsilon) + P(|X_n - X| > \varepsilon) \end{aligned}$$

As $n \to \infty$, $P(|X_n - X| > \varepsilon) \to 0$, then

$$\begin{aligned} P(X \leq x - \varepsilon) \leq \lim_{n\to\infty} P(X_n \leq x) \leq P(X \leq x + \varepsilon) \\ \Rightarrow F(x - \varepsilon) \leq \lim_{n\to\infty} F_n(x) \leq F(x + \varepsilon) \end{aligned}.$$

By the property of distribution (Theorem 12.3.1), as $\varepsilon \to 0$, then

$$\lim_{n\to\infty} F_n(x) = F(x),$$

which means,

$$X_n \xrightarrow{d} X.$$

2. Since $X_n \xrightarrow{d} c$, where c is a constant, then $\forall \varepsilon > 0$,

$$\lim_{n\to\infty} P(X_n \leq c+\varepsilon) = 1 \Rightarrow \lim_{n\to\infty} P(X_n > c+\varepsilon) = 0$$
$$\lim_{n\to\infty} P(X_n \leq c-\varepsilon) = 0.$$

Therefore,

$$P\left(|X_n - c| < \varepsilon\right) = 0,$$

which means

$$X_n \xrightarrow{p} c.$$

□

Theorem 13.2.2

If $X_n \xrightarrow{a.s.} X$, then

$$X_n \xrightarrow{p} X. \tag{13.24}$$

Proof. □

Theorem 13.2.3

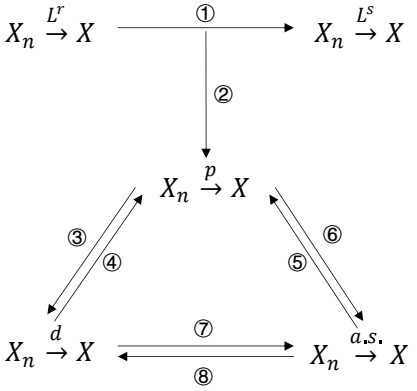
$X_n \xrightarrow{p} X$ if and only if for all subsequence $X_{n(m)}$ exists a further subsequence $X_{n(m_k)}$, such that

$$X_{n(m_k)} \xrightarrow{a.s.} X. \tag{13.25}$$

Lemma 13.2.2

If $F_n \xrightarrow{d} F_\infty$, then there are random variables $Y_n, 1 \leq n \leq \infty$, with distribution F_n so that

$$Y_n \xrightarrow{a.s.} Y_\infty. \tag{13.26}$$



① $if\ \mathrm{r} \geq \mathrm{s} \geq 1, X_n \xrightarrow{L^r} X \Rightarrow X_n \xrightarrow{L^s} X$

② $X_n \xrightarrow{L^r} X \Rightarrow X_n \xrightarrow{p} X$

③ $X_n \xrightarrow{p} X \Rightarrow X_n \xrightarrow{d} X$

④ $if\ X = c, X_n \xrightarrow{d} X \Rightarrow X_n \xrightarrow{p} X$

⑤ $X_n \xrightarrow{a.s.} X \Rightarrow X_n \xrightarrow{p} X$

⑥ $X_n \xrightarrow{p} X \Rightarrow X_{n(m_k)} \xrightarrow{a.s.} X$

⑦ $X_n \xrightarrow{d} X \Rightarrow \exists Y_n \overset{d}{=} X_n, Y \overset{d}{=} X, s.t.\ Y_n \xrightarrow{a.s.} Y$

⑧ $X_n \xrightarrow{a.s.} X \Rightarrow X_n \xrightarrow{d} X$

Figure 13.1: Relations of Convergence of Random Variables

Chapter 14

Law of Large Numbers

14.1 Weak Law of Large Numbers

Theorem 14.1.1 (Weak Law of Large Numbers with Finite Variances)

Let $X_1, X_2, \ldots$ be iid random variables with $EX_i = \mu$ and $\mathrm{Var}(X_i) \le C < \infty$. Suppose $S_n = X_1 + X_2 + \cdots + X_n$, then

$$S_n/n \xrightarrow{L^2} \mu, \quad S_n/n \xrightarrow{p} \mu. \tag{14.1}$$

Proof. □

Theorem 14.1.2 (Weak Law of Large Numbers without iid)

Let $X_1, X_2, \ldots$ be random variables, Suppose $S_n = X_1 + X_2 + \cdots + X_n$, $\mu_n = ES_n$, $\sigma_n^2 = \mathrm{Var}(S_n)$, if $\sigma_n^2/b_n^2 \to 0$, then

$$\frac{S_n - \mu_n}{b_n} \xrightarrow{p} 0. \tag{14.2}$$

Proof. □

Theorem 14.1.3 (Weak Law of Large Numbers for Triangular Arrays)

For each n, let $X_{n,m}, 1 \le m \le n$, be independent random variables. Suppose $b_n > 0$ with $b_n \to \infty$, $\bar{X}_{n,m} = X_{n,m} I_{(X_{n,m} \le b_n)}$, if

1. $\sum_{m=1}^{n} P\left(|X_{n,m}| > b_n\right) \to 0$, and
2. $b_n^{-2} \sum_{m=1}^{n} E\bar{X}_{n,m}^2 \to 0$.

Suppose $S_n = X_{n,1} + \cdots + X_{n,n}$ and $a_n = \sum_{m=1}^{n} E\bar{X}_{n,m}$, then

$$\frac{S_n - a_n}{b_n} \xrightarrow{p} 0. \tag{14.3}$$

Proof. □

Theorem 14.1.4 (Weak Law of Large Numbers by Feller)

Let $X_1, X_2, \ldots$ be iid random variables with

$$\lim_{x \to 0} xP(|X_i| > x) = 0. \tag{14.4}$$

Suppose $S_n = X_1 + X_2 + \cdots + X_n$, $\mu_n = E\left(X_1 I_{(|X_1| < n)}\right)$, then

$$S_n/n - \mu_n \xrightarrow{p} 0. \tag{14.5}$$

Proof. □

Theorem 14.1.5 (Weak Law of Large Numbers)

Let $X_1, X_2, \ldots$ be iid random variables with $E|X_i| < \infty$. Suppose $S_n = X_1 + X_2 + \cdots + X_n$, $\mu = EX_i$, then

$$S_n/n \xrightarrow{p} \mu. \tag{14.6}$$

Proof. □

Remark. $E|X_i| = \infty$

14.2 Strong Law of Large Numbers

14.2.1 Borel-Cantelli Lemmas

Lemma 14.2.1 (Borel-Cantelli Lemma)

If $\sum_{n=1}^{\infty} P(A_n) < \infty$, then

$$P(A_n \text{ i.o. }) = 0. \tag{14.7}$$

Lemma 14.2.2 (The Second Borel-Cantelli Lemma)

If $\{A_n\}$ are independent with $\sum_{n=1}^{\infty} P(A_n) = \infty$, then,

$$P(A_n \text{ i.o. }) = 1. \tag{14.8}$$

Corollary 14.2.1

Suppose $\{A_n\}$ are independent with $P(A_n) < 1, \forall n$. If $P(\cup_{n=1}^{\infty} A_n) = 1$ then

$$\sum_{n=1}^{\infty} P(A_n) = \infty, \tag{14.9}$$

and hence $P(A_n \text{ i.o. }) = 1$

Proof. □

14.2.2 Strong Law of Large Numbers

Theorem 14.2.1 (Strong Law of Large Numbers)

Let $X_1, X_2, \ldots$ be iid random variables with $E|X_i| < \infty$. Suppose $S_n = X_1 + X_2 + \cdots + X_n$, $\mu = EX_i$, then

$$S_n/n \overset{a.s.}{\to} \mu. \tag{14.10}$$

14.3 Uniform Law of Large Numbers

Theorem 14.3.1 (Uniform Law of Large Numbers)

Suppose

1. Θ is compact.
2. $g(X_i, \theta)$ is continuous at each $\theta \in \Theta$ almost sure.
3. $g(X_i, \theta)$ is dominated by a function $G(X_i)$, i.e. $|g(X_i, \theta)| \leq G(X_i)$.
4. $EG(X_i) < \infty$.

Then

$$\sup_{\theta \in \Theta} \left| n^{-1} \sum_{i=1}^{n} g(X_i, \theta) - Eg(X_i, \theta) \right| \xrightarrow{p} 0. \tag{14.11}$$

Proof. Suppose

$$\Delta_\delta(X_i, \theta_0) = \sup_{\theta \in B(\theta_0, \delta)} g(X_i, \theta) - \inf_{\theta \in B(\theta_0, \delta)} g(X_i, \theta).$$

Since (i) $\Delta_\delta(X_i, \theta_0) \xrightarrow{a.s.} 0$ by condition (2), (ii) $\Delta_\delta(X_i, \theta_0) \leq 2 \sup_{\theta \in \Theta} |g(X_i, \theta)| \leq 2G(X_i)$ by condition (3) and (4). Then

$$E\Delta_\delta(X_i, \theta_0) \to 0, \delta \to 0.$$

So, for all $\theta \in \Theta$ and $\varepsilon > 0$, there exists $\delta_\varepsilon(\theta)$ such that

$$E\left[\Delta_{\delta_\varepsilon(\theta)}(X_i, \theta)\right] < \varepsilon.$$

Since Θ is compact, we can find a finite subcover, such that Θ is covered by

$$\cup_{k=1}^{K} B(\theta_k, \delta_\varepsilon(\theta_k)).$$

$$\begin{aligned}
&\sup_{\theta \in \Theta} \left[n^{-1} \sum_{i=1}^{n} g(X_i, \theta) - Eg(X_i, \theta) \right] \\
=& \max_k \sup_{\theta \in B(\theta_k, \delta_\varepsilon(\theta_k))} \left[n^{-1} \sum_{i=1}^{n} g(X_i, \theta) - Eg(X_i, \theta) \right] \\
\leq& \max_k \left[n^{-1} \sum_{i=1}^{n} \sup_{\theta \in B(\theta_k, \delta_\varepsilon(\theta_k))} g(X_i, \theta) - E \inf_{\theta \in B(\theta_k, \delta_\varepsilon(\theta_k))} g(X_i, \theta) \right]
\end{aligned}.$$

Since

$$E \left| \sup_{\theta \in B(\theta_k, \delta_c(\theta_k))} g(X_i, \theta) \right| \leq EG(X_i) < \infty,$$

by the Weak Law of Large Numbers (Theorem 14.1.5),

$$
\begin{aligned}
=& o_p(1) + \max_k \left[E \sup_{\theta \in B(\theta_k, \delta_\varepsilon(\theta_k))} g(X_i, \theta) - E \inf_{\theta \in B(\theta_k, \delta_\varepsilon(\theta_k))} g(X_i, \theta) \right] \\
=& o_p(1) + \max_k E \Delta_{\delta_\varepsilon(\theta_k)}(X_i, \theta_k) \\
\leq& o_p(1) + \varepsilon
\end{aligned}
$$

By analogous argument,

$$
\inf_{\theta \in \Theta} \left[n^{-1} \sum_{i=1}^{n} g(X_i, \theta) - E g(X_i, \theta) \right] \geq o_p(1) - \varepsilon.
$$

The desired result follows from the above equation by the fact that ε is chosen arbitrarily. □

Chapter 15

Central Limit Theorems

15.1 Central Limit Theorem

15.1.1 The De Moivre-Laplace Theorem

Lemma 15.1.1 (Stirling's Formula)

$$n! \sim \sqrt{2\pi} n^{n+\frac{1}{2}} e^{-n} \quad n \to \infty. \tag{15.1}$$

Proof. □

Lemma 15.1.2

If $c_j \to 0$, $a_j \to \infty$ and $a_j c_j \to \lambda$, then

$$(1 + c_j)^{a_j} \to e^{\lambda}. \tag{15.2}$$

Proof. □

Theorem 15.1.1 (The De Moivre-Laplace Theorem)

Let $X_1, X_2, \ldots$ be iid with $P(X_1 = 1) = P(X_1 = -1) = 1/2$ and let $S_n = X_1 + \cdots + X_n$. If $a < b$, then as $m \to \infty$

$$P\left(a \leq S_m/\sqrt{m} \leq b\right) \to \int_a^b (2\pi)^{-1/2} e^{-x^2/2} \, \mathrm{d}x. \tag{15.3}$$

Proof. If n and k and integers

$$P(S_{2n} = 2k) = \begin{pmatrix} 2n \\ n+k \end{pmatrix} 2^{-2n}$$

By Lemma 15.1.1, we have

$$\begin{aligned}\binom{2n}{n+k} &= \frac{(2n)!}{(n+k)!(n-k)!} \\ &\sim \frac{(2n)^{2n}}{(n+k)^{n+k}(n-k)^{n-k}} \cdot \frac{(2\pi(2n))^{1/2}}{(2\pi(n+k))^{1/2}(2\pi(n-k))^{1/2}}\end{aligned}$$

Hence,

$$\begin{aligned}P\left(S_{2n}=2k\right) &= \binom{2n}{n+k} 2^{-2n} \\ &\sim \left(1+\frac{k}{n}\right)^{-n-k} \cdot \left(1-\frac{k}{n}\right)^{-n+k} \\ &\cdot (\pi n)^{-1/2} \cdot \left(1+\frac{k}{n}\right)^{-1/2} \cdot \left(1-\frac{k}{n}\right)^{-1/2} \\ &= \left(1-\frac{k^2}{n^2}\right)^{-n} \cdot \left(1+\frac{k}{n}\right)^{-k} \cdot \left(1-\frac{k}{n}\right)^{k} \\ &\cdot (\pi n)^{-1/2} \cdot \left(1+\frac{k}{n}\right)^{-1/2} \cdot \left(1-\frac{k}{n}\right)^{-1/2}\end{aligned}$$

Let $2k = x\sqrt{2n}$, i.e., $k = x\sqrt{\frac{n}{2}}$. By Lemma 15.1.2, we have

$$\begin{aligned}\left(1-\frac{k^2}{n^2}\right)^{-n} &= \left(1-x^2/2n\right)^{-n} \to e^{x^2/2} \\ \left(1+\frac{k}{n}\right)^{-k} &= (1+x/\sqrt{2n})^{-x\sqrt{n/2}} \to e^{-x^2/2} \\ \left(1-\frac{k}{n}\right)^{k} &= (1-x/\sqrt{2n})^{x\sqrt{n/2}} \to e^{-x^2/2}\end{aligned}$$

For this choice of k, $k/n \to 0$, so

$$\left(1+\frac{k}{n}\right)^{-1/2} \cdot \left(1-\frac{k}{n}\right)^{-1/2} \to 1.$$

Putting things together, we have

$$P\left(S_{2n}=2k\right) \sim (\pi n)^{-1/2} e^{-x^2/2}, \frac{2k}{\sqrt{2n}} \to x.$$

Therefore,

$$P\left(a\sqrt{2n} \le S_{2n} \le b\sqrt{2n}\right) = \sum_{m \in \left[a\sqrt{2n}, b\sqrt{2n}\right] \cap 2\mathbb{Z}} P\left(S_{2n}=m\right)$$

Let $m = x\sqrt{2n}$, we have that this is

$$\approx \sum_{x \in [a,b] \cap \left(2\mathbb{Z}/\sqrt{2n}\right)} (2\pi)^{-1/2} e^{-x^2/2} \cdot (2/n)^{1/2}$$

where $2\mathbb{Z}/\sqrt{2n} = \left\{2z/\sqrt{2n} : z \in \mathbb{Z}\right\}$. As $n \to \infty$, the sum just shown is

$$\approx \int_a^b (2\pi)^{-1/2} e^{-x^2/2} \, \mathrm{d}x.$$

To remove the restriction to even integers, observe $S_{2n+1} = S_{2n} \pm 1$.
Let $m = 2n$, as $m \to \infty$,

$$P\left(a \le S_m/\sqrt{m} \le b\right) \to \int_a^b (2\pi)^{-1/2} e^{-x^2/2} \, \mathrm{d}x.$$

□

15.1.2 Central Limit Theorem

Theorem 15.1.2 (Classic Central Limit Theorem)

Let $X_1, X_2, \ldots$ be iid with $\mathbb{E}X_i = \mu$, $\mathrm{Var}(X_i) = \sigma^2 < \infty$. Let $\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i$, then

$$n^{1/2} \frac{\bar{X} - \mu}{\sigma} \xrightarrow{d} \mathcal{N}(0,1). \tag{15.4}$$

Proof. □

Theorem 15.1.3 (The Linderberg-Feller Central Limit Theorem)

For each n, let $X_{n,m}, 1 \le m \le n$, be independent random variables with $EX_{n,m} = 0$. If

1. $\sum_{m=1}^n EX_{n,m}^2 \to \sigma^2 > 0$.
2. $\forall \epsilon > 0, \lim_{n\to\infty} \sum_{m=1}^n E\left(|X_{n,m}|^2 ; |X_{n,m}| > \epsilon\right) = 0$

Then $S_n = X_{n,1} + \cdots + X_{n,n} \xrightarrow{d} \sigma\chi$ as $n \to \infty$.

15.1.3 Berry-Esseen Theorem

Theorem 15.1.4 (Berry-Esseen Theorem)

Let $X_1, X_2, \ldots, X_n$ be iid with distribution F, which has a mean μ and a finite third moment σ^3, then there exists a constant C (independent of F),

$$|G_n(x) - \Phi(x)| \le \frac{C}{\sqrt{n}} \frac{E\,|X_1 - \mu|^3}{\sigma^3}, \quad \forall x. \tag{15.5}$$

Corollary 15.1.1

Under the assumptions of Theorem 51,

$$G_n(x) \to \Phi(x) \text{ as } n \to \infty$$

for any sequence F_n with mean ξ_n and variance σ_n^2 for which

$$\frac{E_n \left|X_1 - \xi_n\right|^3}{\sigma_n^3} = o(\sqrt{n})$$

and thus in particular if (72) is bounded. Here E_n denotes the expectation under F_n.

15.2 CLT for independent non-identical Random Variables

Theorem 15.2.1 (Liapounov CLT)

Theorem 15.2.2

Let $Y_1, Y_2, \ldots$ be iid with $\mathbb{E}(Y_i) = 0$, $\text{Var}(Y_i) = \sigma^2 > 0$, and $\mathbb{E}\left|Y_i^3\right| = \gamma < \infty$. If

$$\left(\sum_{i=1}^{n} |d_{ni}|^3\right)^2 = o\left(\sum_{i=1}^{n} d_{ni}^2\right)^3, \tag{15.6}$$

then

$$\frac{\sum_{i=1}^{n} d_{ni} Y_i}{\sigma \sqrt{\sum_{i=1}^{n} d_n^2}} \xrightarrow{d} N(0,1).$$

Corollary 15.2.1

The sufficient condition (15.6) is euivalent to

$$\max_{i=1,\ldots,n} \left(d_{ni}^2\right) = o\left(\sum d_{ni}^2\right). \tag{15.7}$$

15.3 CLT for Dependent Random Variables

Chapter 16

Multivariate Extensions

16.1 Multivariate Distributions

16.1.1 Multivariate Normal Distribution

Definition 16.1.1 (Multivariate Normal Distribution)

The multivariate normal distribution of a p-dimensional random vector $\mathbf{X}$ can be written as:

$$\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

where $\boldsymbol{\mu}$ is a p-dimensional mean vector and $\boldsymbol{\Sigma}$ is a $p \times p$ covariance matrix. Furthermore, the probability density function of $\mathbf{X}$ is:

$$p(\mathbf{X}) = \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{X}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{X}-\boldsymbol{\mu})\right).$$

Theorem 16.1.1

Suppose $\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then:

1. $\boldsymbol{\Sigma}^{-1/2}(\mathbf{X}-\boldsymbol{\mu}) \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$.
2. $(\mathbf{X}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{X}-\boldsymbol{\mu}) \sim \chi^2_p$.

16.1.2 Wishart Distribution

Definition 16.1.2 (Wishart Distribution)

The Wishart distribution is a generalization of the chi-squared distribution to multiple dimensions. If $\mathbf{Z}$ is a $p \times n$ matrix with each column drawn from a multivariate normal distribution $\mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$, then the quadratic form $\mathbf{X}$ has a Wishart distribution (with parameters $\mathbf{\Sigma}$, and n):

$$\mathbf{X} = \mathbf{Z}\mathbf{Z}^\top \sim W_p(\mathbf{\Sigma}, n).$$

Furthermore, the probability density function of $\mathbf{X}$ is:

$$p(\mathbf{X}) = \frac{|\mathbf{X}|^{(n-p-1)/2} \exp\left(-\frac{1}{2}\mathrm{tr}(\mathbf{\Sigma}^{-1}\mathbf{X})\right)}{2^{np/2}|\mathbf{\Sigma}|^{n/2}\Gamma_p(n/2)}$$

16.1.3 Hotelling's T-squared Distribution

Definition 16.1.3 (Hotelling's T^2 Distribution)

If the vector $\mathbf{d}$ is Gaussian multivariate-distributed with zero mean and unit covariance matrix $\mathcal{N}(\mathbf{0}_p, \mathbf{I}_p)$ and $\mathbf{M}$ is a $p \times p$ matrix with unit scale matrix and m degrees of freedom with a Wishart distribution $W(\mathbf{I}_p, m)$, then the quadratic form X has a Hotelling distribution (with parameters p and m):

$$X = m\mathbf{d}^\top \mathbf{M}^{-1}\mathbf{d} \sim T^2(p, m).$$

Furthermore, if a random variable X has Hotelling's T-squared distribution, $X \sim T^2_{p,m}$, then:

$$\frac{m-p+1}{pm} X \sim F_{p,m-p+1}$$

where $F_{p,m-p+1}$ is the F-distribution with parameters p and $m-p+1$.

16.2 Convergence of Random Vectors

Let $\mathbf{X}^{(n)}$ be a sequence of random vectors with cdf H_n converging in law to $\mathbf{X}$ with cdf H. One then often needs to know whether for some set S in R^k,

$$P\left(\mathbf{X}^{(n)} \in S\right) \to P(\mathbf{X} \in S). \tag{16.1}$$

That (16.1) need not be true for all S is seen from the case $k = 1$, $S = \{x : x \leq a\}$. Then (16.1) can only be guaranteed when a is a continuity point of H.

Theorem 16.2.1

A sufficient condition for (16.1) to hold is that

$$P(\mathbf{X} \in \partial S) = 0.$$

Example (Multinomial)**.**

Example (Difference of Means)**.** Let $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_n$ be independently distributed according to distributions F and G, with means ξ and η and finite variances σ^2 and τ^2, respectively. Then

$$\sqrt{m}(\bar{X} - \xi) \xrightarrow{d} \mathcal{N}\left(0, \sigma^2\right), \quad \sqrt{n}(\bar{Y} - \eta) \xrightarrow{d} \mathcal{N}\left(0, \tau^2\right).$$

If $\frac{m}{m+n} \to \lambda$ $(0 < \lambda < 1)$, it follows that

$$\sqrt{m+n}(\bar{X} - \xi) = \sqrt{\frac{m+n}{m}}\sqrt{m}(\bar{X} - \xi) \xrightarrow{d} N\left(0\frac{\sigma^2}{\lambda}\right),$$
$$\sqrt{m+n}(\bar{Y} - \eta) \xrightarrow{d} N\left(0, \frac{\tau^2}{1-\lambda}\right),$$

and hence that

$$(\sqrt{m+n}(\bar{X} - \xi), \sqrt{m+n}(\bar{Y} - \eta)) \xrightarrow{d} (X, Y),$$

where X and Y are independent random variables with distributions $N\left(0, \frac{\sigma^2}{\lambda}\right)$ and $N\left(0, \frac{\tau^2}{1-\lambda}\right)$, respectively. Since $Y - X$ is a continuous function of (X, Y), it follows that

$$\sqrt{m+n}[(\bar{Y} - \bar{X}) - (\eta - \xi)] \xrightarrow{d} N\left(0, \frac{\sigma^2}{\lambda} + \frac{\tau^2}{1-\lambda}\right),$$

or, equivalently, that

$$\frac{(\bar{Y} - \bar{X}) - (\eta - \xi)}{\sqrt{\frac{\sigma^2}{m} + \frac{\tau^2}{n}}} \xrightarrow{d} N(0, 1).$$

More specifically, consider the probability

$$P\{\sqrt{m+n}[(\bar{Y} - \bar{X}) - (\eta - \xi)] \leq z\}.$$

By Theorem 16.2.1, since $P(Y - X = z) = 0$, it follows that

$$P\{(Y - X) \leq z\} = \Phi\left(\frac{z}{\sqrt{\frac{\sigma^2}{\lambda} + \frac{\tau^2}{1-\lambda}}}\right).$$

Theorem 16.2.2

A necessary and sufficient condition for

$$(X_1^{(n)}, \ldots, X_k^{(n)}) \xrightarrow{d} (X_1, \ldots, X_k),$$

is that for any constants $c_1, \ldots, c_k$, we have

$$\sum_{i=1}^{k} c_i X_i^{(n)} \xrightarrow{d} \sum_{i=1}^{k} c_i X_i.$$

Example (Orthogonal Linear Combinations)**.** Let $Y_1, Y_2, \ldots$ be iid with mean $\mathbb{E}(Y_i) = 0$ and variance $\mathrm{Var}(Y_i) = \sigma^2$, and consider the joint distribution of the linear combinations

$$X_1^{(n)} = \sum_{j=1}^{n} a_{nj} Y_j, \quad X_2^{(n)} = \sum_{j=1}^{n} b_{nj} Y_j$$

satisfying the orthogonality conditions

$$\sum_{j=1}^{n} a_{nj}^2 = \sum_{j=1}^{n} b_{nj}^2 = 1, \quad \sum_{j=1}^{n} a_{nj} b_{nj} = 0. \tag{16.2}$$

Then if

$$\max_j a_{jn}^2 \to 0, \quad \max_j b_{jn}^2 \to 0, n \to \infty.$$

it follows that

$$(X_1^{(n)}, X_2^{(n)}) \xrightarrow{d} (X_1, X_2),$$

with (X_1, X_2) independently distributed, each according to the normal distribution $N(0, \sigma^2)$.

Proof. To prove this result, it is by Theorem 16.2.2 enough to show that

$$c_1 X_1^{(n)} + c_2 X_2^{(n)} = \sum (c_1 a_{nj} + c_2 b_{nj}) Y_j \xrightarrow{d} c_1 X_1 + c_2 X_2,$$

where the distribution of $c_1 X_1 + c_2 X_2$ is $N(0, [c_1^2 + c_2^2]\sigma^2)$. The sum on the left side of (5.1.24) is of the form $\sum d_{nj} Y_j$ with

$$d_{nj} = c_1 a_{nj} + c_2 b_{nj}$$

It follows from (16.2) that

$$\sum_{j=1}^{n} d_{nj}^2 = c_1^2 + c_2^2$$

furthermore

$$\max d_{nj}^2 \leq 2 \max \left[c_1^2 a_{nj}^2 + c_2^2 b_{nj}^2 \right] \leq 2 \left[c_1^2 \max a_{nj}^2 + c_2^2 \max b_{nj}^2 \right].$$

Thus,

$$\max_j d_{nj}^2 / \sum_{j=1}^{n} d_{nj}^2 \to 0.$$

According to Theorems 15.2.2, it thus follows that

$$\sum d_{nj} Y_j \xrightarrow{d} N\left(0, \left(c_1^2 + c_2^2\right) \sigma^2\right).$$

Thus, by Theorem 16.2.2, we have

$$(X_1^{(n)}, X_2^{(n)}) \xrightarrow{d} (X_1, X_2),$$

with (X_1, X_2) independently distributed, each according to the normal distribution $N(0, \sigma^2)$. □

Chapter 17

Exercises for Probability Theory and Examples

17.1 Measure Theory

Exercise. 1. Show that if $\mathcal{F}_1 \subset \mathcal{F}_2 \subset \ldots$ are σ -algebras, then $\cup_i \mathcal{F}_i$ is an algebra.

2. Give an example to show that $\cup_i \mathcal{F}_i$ need not be a σ -algebra.

Proof. 1. **Complement**: Suppose $A \in \cup_i \mathcal{F}_i$, since $\mathcal{F}_1 \subset \mathcal{F}_2 \subset \ldots$, assume $A \in \mathcal{F}_i$. And each $\mathcal{F}_i$ is σ-algebra,

$$A^c \in \mathcal{F}_i \subset \cup_i \mathcal{F}_i.$$

Finite Union: Suppose $A_1, A_2 \in \cup_i \mathcal{F}_i$, since $\mathcal{F}_1 \subset \mathcal{F}_2 \subset \ldots$, assume $A_1 \in \mathcal{F}_i, A_2 \in \mathcal{F}_j$, such that,

$$A_1, A_2 \in \mathcal{F}_{\max(i,j)}.$$

Since each $\mathcal{F}_i$ is σ-algebra,

$$A_1 \cup A_2 \in \mathcal{F}_i \subset \cup_i \mathcal{F}_i.$$

2. Let $\mathcal{F}_i$ be a Borel Set of $[1, 2 - \frac{1}{i}]$. Suppose $A_i = [1, 2 - \frac{1}{i}] \in \mathcal{F}_i$,

$$\cup_i A_i = [1, 2) \notin \cup_i \mathcal{F}_i.$$

□

17.2 Laws of Large Numbers

17.3 Central Limit Theorems

Exercise. Let $g \geq 0$ be continuous. If $X_n \xrightarrow{d} X_\infty$, then

$$\liminf_{n\to\infty} Eg\left(X_n\right) \geq Eg\left(X_\infty\right).$$

Proof. Let $Y_n \overset{d}{=} X_n, 1 \leq n \leq \infty$ with $Y_n \xrightarrow{a.s.} Y_\infty$ (Lemma 13.2.2). Since $g \geq 0$ be continuous, $g(Y_n) \xrightarrow{a.s.} g(Y_\infty)$ and $g(Y_n) \geq 0$ (Theorem 13.1.2), and the Fatou's Lemma (12.4.2) implies,

$$\begin{aligned}\liminf_{n\to\infty} Eg(X_n) =& \liminf_{n\to\infty} Eg(Y_n) \geq E\left(\liminf_{n\to\infty} g(Y_n)\right)\\ =& Eg(Y_\infty) = Eg(X_\infty).\end{aligned}$$

□

Exercise. Suppose g, h are continuous with $g(x) > 0$, and $|h(x)|/g(x) \to 0$ as $|x| \to \infty$. If $F_n \xrightarrow{d} F$ and $\int g(x)\,\mathrm{d}F_n(x) \leq C < \infty$, then

$$\int h(x)\,\mathrm{d}F_n(x) \to \int h(x)\,\mathrm{d}F(x).$$

Proof.

$$\begin{aligned}\left|\int h(x)\,\mathrm{d}F_n(x) - \int h(x)\,\mathrm{d}F(x)\right| =& \left|\int_{x\in[-M,M]} h(x)\,\mathrm{d}F_n(x) + \int_{x\notin[-M,M]} h(x)\,\mathrm{d}F_n(x)\right.\\ &\left. - \int_{x\in[-M,M]} h(x)\,\mathrm{d}F(x) - \int_{x\notin[-M,M]} h(x)\,\mathrm{d}F(x)\right|\\ \leq& \left|\int_{x\in[-M,M]} h(x)\,\mathrm{d}F_n(x) - \int_{x\in[-M,M]} h(x)\,\mathrm{d}F(x)\right|\\ &+ \left|\int_{x\notin[-M,M]} h(x)\,\mathrm{d}F_n(x) - \int_{x\notin[-M,M]} h(x)\,\mathrm{d}F(x)\right|\end{aligned}.$$

Let $X_n, 1 \leq n < \infty$, with distribution F_n, so that $X_n \xrightarrow{a.s.} X$ (Lemma 13.2.2).

$$\left|\int_{x\in[-M,M]} h(x)\,\mathrm{d}F_n(x) - \int_{x\in[-M,M]} h(x)\,\mathrm{d}F(x)\right| = \left|E\left(h(X_n) - h(X)\right) I_{x\in[-M,M]}\right|.$$

By Continuity Mapping Theorem (13.1.2), $\lim_{n\to\infty}\left|E\left(h(X_n) - h(X)\right) I_{x\in[-M,M]}\right| = 0$.

Since

$$h(x)I_{x\notin[-M,M]} \le g(x) \sup_{x\notin[-M,M]} \frac{h(x)}{g(x)},$$

and by Exercise 17.3

$$Eg(X) \le \liminf_{n\to\infty} Eg(X_n) = \liminf_{n\to\infty} \int g(x)\,\mathrm{d}F_n(x) \le C < \infty,$$

$$\begin{aligned}\left|\int_{x\notin[-M,M]} h(x)\,\mathrm{d}F_n(x) - \int_{x\notin[-M,M]} h(x)\,\mathrm{d}F(x)\right| &= \left|E\left(h(X_n)-h(X)\right)I_{x\notin[-M,M]}\right| \\ &\le 2E\max(h(Xn),h(X))I_{x\notin[-M,M]} \le 2C \sup_{x\notin[-M,M]} \frac{h(x)}{g(x)}.\end{aligned}$$

Hence, let $M\to\infty$,

$$\lim_{n\to\infty}\left|\int h(x)\,\mathrm{d}F_n(x) - \int h(x)\,\mathrm{d}F(x)\right| \le 2C \sup_{x\notin[-M,M]} \frac{h(x)}{g(x)} \to 0,$$

which means,

$$\int h(x)\,\mathrm{d}F_n(x) \to \int h(x)\,\mathrm{d}F(x).$$

□

Exercise. Let $X_1, X_2, \ldots$ be iid with $EX_i = 0$ and $EX_i^2 = \sigma^2 \in (0,\infty)$. Then

$$\sum_{m=1}^{n} X_m \Big/ \left(\sum_{m=1}^{n} X_m^2\right)^{1/2} \xrightarrow{d} \chi.$$

Exercise. Show that if $|X_i| \le M$ and $\sum_n \mathrm{Var}\,(X_n) = \infty$, then

$$\left(S_n - ES_n\right)/\sqrt{\mathrm{Var}\,(S_n)} \xrightarrow{d} \chi.$$

Exercise. Suppose $EX_i = 0, EX_i^2 = 1$ and $E\left|X_i\right|^{2+\delta} \le C$ for some $0 < \delta, C < \infty$. Show that

$$S_n/\sqrt{n} \xrightarrow{d} \chi.$$

Part VII

Stochastic Process

Chapter 18

Martingales

18.1 Conditional Expectation

Definition 18.1.1 (Conditional Expectation)

Example. 1. If $X \in \mathcal{F}$, then

$$E(X \mid \mathcal{F}) = X.$$

2. If X is independent of $\mathcal{F}$, then

$$E(X \mid \mathcal{F}) = E(X).$$

3. If $\Omega_1, \Omega_2, \ldots$ is a finite or infinite partition of Ω into disjoint sets, each of which has a positive probability, and let $\mathcal{F} = \sigma(\Omega_1, \Omega_2, \ldots)$, then for each i,

$$E(X \mid \mathcal{F}) = \frac{E(X; \Omega_i)}{P(\Omega_i)}.$$

Property.

18.2 Martingales

Let $\mathcal{F}_n$ be a filtration, i.e., an increasing sequence of σ-fields.

Definition 18.2.1 (Martingale)

A sequence $\{X_n\}$ of real-valued random variables is said to be a martingale for $\mathcal{F}_n$, if

1. X_n is integrable, i.e., $E|X_n| < \infty$
2. X_n is adapted to $\mathcal{F}_n$, i.e., $\forall n, X_n \in \mathcal{F}_n$
3. X_n satisfies the martingale condition, i.e.,

$$E(X_{n+1} \mid \mathcal{F}_n) = X_n, \quad \forall n \tag{18.1}$$

Remark. If in the last definition, $=$ is replaced by $\leq$ or $\geq$, then X is said to be a supermartingale or submartingale, respectively.

Example (Linear Martingale)**.**

Example (Quadratic Martingale)**.**

Example (Exponential Martingale)**.**

Example (Random Walk)**.** Suppose $X_n = X_0 + \xi_1 + \cdots + \xi_n$, where X_0 is constant, ξ_m are independent and have $E\xi_m = 0, \sigma_m^2 = E\xi_m^2 < \infty$. Let $\mathcal{F}_n = \sigma(\xi_1, \ldots, \xi_n)$ for $n \geq 1$ and take $\mathcal{F}_0 = \{\emptyset, \Omega\}$. Show X_n is a martingale, and X_n^2 is a submartingale.

Proof. It is obvious that,

$$E|X_n| < \infty, \quad X_n \in \mathcal{F}_n$$

Since ξ_{n+1} is independent of $\mathcal{F}_n$, so using the linearity of conditional expectation, (4.1.1), and Example 4.1.4,

$$E(X_{n+1} \mid \mathcal{F}_n) = E(X_n \mid \mathcal{F}_n) + E(\xi_{n+1} \mid \mathcal{F}_n) = X_n + E\xi_{n+1} = X_n$$

So X_n is a martingale, and Theorem 4.2.6 implies X_n^2 is a submartingale. □

Remark. If we let $\lambda = x^2$ and apply Theorem 4.4.2 to X_n^2, we get Kolmogorov's maximal inequality, Theorem 2.5.5:

$$P\left(\max_{1 \leq m \leq n} |X_m| \geq x\right) \leq x^{-2} \operatorname{var}(X_n) \tag{18.2}$$

Theorem 18.2.1 (Orthogonality of Martingale Increments)

Theorem 18.2.2 (Conditional Variance Formula)

Definition 18.2.2 (Predictable Sequence)

Definition 18.2.3 (Stopping Time)

Theorem 18.2.3 (Martingale Convergence Theorem)

18.3 Doob's Inequality

Example (Doob's Decomposition)**.** Given a sequence of independent random variables $\{X_k\}_{k=1}^m$, recall the sequence $Y_k = \mathbb{E}[f(X) \mid X_1, \ldots, X_k]$ previously defined, and suppose that $\mathbb{E}|f(X)| < \infty$. We claim that $\{Y_k\}_{k=1}^n$ is a martingale with respect to $\{X_k\}_{k=1}^n$. Indeed, in terms of the shorthand $X_1^k = (X_1, \ldots, X_k)$, we have

$$\mathbb{E}|Y_k| = \mathbb{E}\left|\mathbb{E}\left[f(X) \mid X_1^k\right]\right| \leq \mathbb{E}|f(X)| < \infty$$

where the bound follows from Jensen's inequality. Tuning in to the martingale property, we have

$$\mathbb{E}\left[Y_{k+1} \mid X_1^k\right] = \mathbb{E}\left[\mathbb{E}\left[f(X) \mid X_1^{k+1}\right] \mid X_1^k\right] = \mathbb{E}\left[f(X) \mid X_1^k\right] = Y_k$$

where the second equality follows from the tower property of conditional expectation. This establishes the martingale property of $\{Y_k\}_{k=1}^n$ with respect to $\{X_k\}_{k=1}^n$.

Example (Likelihood ratio)**.** Let f and g be two mutuall absolutely continuous probability densities, and let $\{X_k\}_{k=1}^n$ be a sequence of random variables drawn i.i.d. from f. For each $k > 1$, let $Y_k := \prod_{l=1}^k \frac{g(X_l)}{f(X_l)}$ be the likelihood ratio based on the first k samples. Then the sequence $\{Y_k\}_{k=1}^n$ is a martingale with respect to $\{X_k\}_{k=1}^n$. Indeed, we have

$$\mathbb{E}\left[Y_{k+1} \mid X_1^k\right] = \mathbb{E}\left[\frac{g(X_{k+1})}{f(X_{k+1})}\right] \prod_{l=1}^k \frac{g(X_l)}{f(X_l)} = Y_k,$$

using the fact that $\mathbb{E}\left[\frac{g(X_{k+1})}{f(X_{k+1})}\right] = 1$.

A closely related notion is that of *martingale difference sequences*, meaning an adapted sequence $\{(D_k, \mathcal{F}_k)\}_{k=1}^\infty$ such that

$$\mathbb{E}[|D_k|] < \infty \quad \text{and} \quad \mathbb{E}[D_{k+1} \mid \mathcal{F}_k] = 0.$$

As suggested by the name, the martingale difference sequence arises in a natural way from martingales, given by $D_k = Y_k - Y_{k-1}$ for a martingale $\{(Y_k, \mathcal{F}_k)\}_{k=1}^\infty$.

Martingale difference sequences is a martingale. Sum of martingale difference sequences is a martingale.

Theorem 18.3.1

Let $\{(D_k, \mathcal{F}_k)\}_{k=1}^{\infty}$ be a martingale difference sequence, and suppose that $\mathbb{E}\left[e^{\lambda D_k} \mid \mathcal{F}_{k-1}\right] \leq e^{\frac{\lambda^2 v_k^2}{2}}$ almost surely for any $|\lambda| < 1/\alpha_k/$ Then the following hold:

1. The sum $S_n := \sum_{k=1}^n D_k$ is sub-exponential with parameter $\left(\sqrt{\sum_{k=1}^n v_k^2}, \alpha_*\right)$, where $\alpha_* = \max_k \alpha_k$.
2. The sum satisfies the concentration inequality

$$\mathbb{P}\left(|S_n| \geq t\right) \leq \begin{cases} 2\exp\left(-\frac{t^2}{2\sum_{k=1}^n v_k^2}\right) & \text{if } 0 \leq t \leq \frac{\sqrt{2\sum_{k=1}^n v_k^2}}{\alpha_*} \\ 2\exp\left(-\frac{t}{2\alpha_*}\right) & \text{if } t > \frac{\sqrt{2\sum_{k=1}^n v_k^2}}{\alpha_*} \end{cases}$$

Proof. We follow the standard approach of controling the moment generating function of S_n, and then applying the Chernoff bound. FOr any scalar λ such that $|\lambda| < 1/\alpha_*$, conditioning on $\mathcal{F}_{n-1}$, and applying the iterated expectation yields

$$\mathbb{E}\left[e^{\lambda S_n}\right] = \mathbb{E}\left[\mathbb{E}\left[e^{\lambda S_{n-1}}\right]\mathbb{E}\left[e^{\lambda D_n} \mid \mathcal{F}_{n-1}\right]\right] \leq \mathbb{E}\left[e^{\lambda S_{n-1}}\right] e^{\frac{\lambda^2 v_n^2}{2}},$$

where the inequality follows from the stated assumption. Iterating this inequality, we obtain

$$\mathbb{E}\left[e^{\lambda S_n}\right] \leq e^{\frac{\lambda^2}{2}\sum_{k=1}^n v_k^2},$$

valid for any $|\lambda| < 1/\alpha_*$. This implies that S_n is sub-exponential with parameter $(\sqrt{\sum_{k=1}^n v_k^2}, \alpha_*)$, and the Chernoff bound follows from the definition of sub-exponentiality. □

Corollary 18.3.1 (Azuma-Hoeffding)

Let $\{(D_k, \mathcal{F}_k)\}_{k=1}^{\infty}$ be a martingale difference sequence for which there are constants $\{(a_k, b_k)\}_{k=1}^{\infty}$ such that $D_k \in [a_k, b_k]$ almost surely for all $k = 1, 2, \ldots, n$. Then for any $t > 0$, we have

$$\mathbb{P}\left(|S_n| \geq t\right) \leq 2\exp\left(-\frac{t^2}{2\sum_{k=1}^n (b_k - a_k)^2}\right).$$

An important application of Corollary 18.3.1 concerns functions that satisfy a bounded differences property.

Definition 18.3.1 (Bounded Differences Property)

A function $f : \mathbb{R}^n \to \mathbb{R}$ is said to satisfy the *bounded differences property* with respect to a sequence of sets $\{S_k\}_{k=1}^n$ if for any $x_1, \ldots, x_n \in \mathbb{R}^n$ that differ only in the k-th coordinate, we have

$$|f(x_1, \ldots, x_{k-1}, x_k, x_{k+1}, \ldots, x_n) - f(x_1, \ldots, x_{k-1}, x'_k, x_{k+1}, \ldots, x_n)| \leq L_k$$

for any $x, x' \in \mathbb{R}^n$

For instance, if the function f is L-Lipschitz with respect to the Hamming distance, $d_H(x, x') = \sum_{k=1}^n \mathbb{I}\{x_k \neq x'_k\}$, then f satisfies the bounded differences property with parameters L uniformly over all $k = 1, \ldots, n$.

Theorem 18.3.2 (Bounded Differences Inequality)

Let $f : \mathbb{R}^n \to \mathbb{R}$ satisfy the bounded differences property with parameters $\{L_k\}_{k=1}^n$ and that the random variables $\mathbf{X} = (X_1, \ldots, X_n)$ has independent coordinates. Then for any $t > 0$, we have

$$\mathbb{P}\left(|f(\mathbf{X}) - \mathbb{E}[f(\mathbf{X})]| \geq t\right) \leq 2\exp\left(-\frac{t^2}{2\sum_{k=1}^n L_k^2}\right).$$

Proof. □

Example (Classical Hoeffding from Bounded Differences)**.**

Example (U-statistics)**.**

Example (Rademacher Complexity)**.**

Theorem 18.3.3 (Doob's Inequality)

Theorem 18.3.4 (L^p Maximum Inequality)

18.4 Uniform Integrability

18.5 Optional Stopping Theorems

Chapter 19

Markov Chains

19.1 Markov Chain

Definition 19.1.1 (Markov Chain, Simple)

A sequence $\{X_n\}$ of real-valued random variables is said to be a Markov chain, if for any states $i_0, \dots i_{n-1}, i$, and j

$$P\left(X_{n+1} = j \mid X_n = i, X_{n-1} = i_{n-1}, \dots X_0 = i_0\right) = P\left(X_{n+1} = j \mid X_n = i\right) \tag{19.1}$$

and the transition probability is

$$p(i,j) = P\left(X_{n+1} = j \mid X_n = i\right) \tag{19.2}$$

Example (Random Walk)**.** Suppose $X_n = X_0 + \xi_1 + \cdots + \xi_n$, where X_0 is constant, $\xi_m \in \mathbb{Z}^d$ are independent with distribution μ. Show X_n is a Markov chain with transition probability,

$$p\left(i,j\right) = \mu\left(\{j-i\}\right)$$

Proof. Since ξ_m are independent with distribution μ,

$$\begin{aligned} &P\left(X_{n+1} = j \mid X_n = i, X_{n-1} = i_{n-1}, \dots X_0 = i_0\right) \\ =&P\left(X_n + \xi_{n+1} = j \mid X_n = i\right) = P\left(\xi_{n+1} = j - i\right) = \mu\left(\{j-i\}\right) \end{aligned}$$

□

Definition 19.1.2 (Branching Processes)

Let $\xi_i^n, i, n \geq 1$, be iid nonnegative integer-valued random variables. Define a sequence $Z_n, n \geq 0$ by $Z_0 = 1$ and

$$Z_{n+1} = \begin{cases} \xi_1^{n+1} + \cdots + \xi_{Z_n}^{n+1} & Z_n > 0 \\ 0 & Z_n = 0 \end{cases} \tag{19.3}$$

Z_n is called a Branching process.

Remark. The idea behind the definitions is that Z_n is the number of individuals in the n-th generation, and each member of the n-th generation gives birth independently to an identically distributed number of children.

Example (Branching Processes)**.** Show branching process is a Markov chain with transition probability,

$$p(i,j) = P\left(\sum_{k=1}^{i} \xi_k = j\right)$$

Proof. Since ξ_k^n are independent with identically distribution,

$$\begin{aligned} &P\left(Z_{n+1} = j \mid Z_n = i, Z_{n-1} = i_{n-1}, \ldots Z_0 = i_0\right) \\ =&P\left(\sum_{k=1}^{Z_n} \xi_k^{n+1} = j \mid Z_n = i\right) = P\left(\sum_{k=1}^{i} \xi_k = j\right) \end{aligned}$$

□

Suppose $(S, \mathcal{S})$ is a measurable space, which will be the state space for our Markov chain.

Definition 19.1.3 (Transition Probability)

A function $p : S \times \mathcal{S} \to \mathbb{R}$ is said to be a transition probability, if

1. For each $x \in S$, $A \to p(x, A)$ is a probability measure on $(S, \mathcal{S})$
2. For each $A \in \mathcal{S}$, $x \to p(x, A)$ is a measurable function

Definition 19.1.4 (Markov Chain)

A sequence $\{X_n\}$ of real-valued random variables with transition probability p is said to be a Markov chain with respect to $\mathcal{F}_n$, if

$$P\left(X_{n+1} \in B \mid \mathcal{F}_n\right) = p\left(X_n, B\right) \tag{19.4}$$

Remark. Given a transition probability p and an initial distribution μ on $(S, \mathcal{S})$, the consistent set of finite dimensional distributions is

$$P\left(X_j \in B_j, 0 \leq j \leq n\right) = \int_{B_0} \mu\left(\mathrm{d}x_0\right) \int_{B_1} p\left(x_0, \mathrm{d}x_1\right) \cdots \int_{B_n} p\left(x_{n-1}, \mathrm{d}x_n\right) F \tag{19.5}$$

19.2 Markov Properties

Definition 19.2.1 (Shift Operator)

Theorem 19.2.1 (Markov Property)

Corollary 19.2.1 (Chapman-Kolmogorov Equation)

Theorem 19.2.2 (Strong Markov Property)

19.3 Recurrence and Transience

Let $T_y^0 = 0$, and for $k \geq 1$, and

$$T_y^k = \inf\left\{n > T_y^{k-1} : X_n = y\right\} \tag{19.6}$$

then T_y^k is the time of the k-th return to y, where $T_y^1 > 0$, so any visit at time 0 does not count.

Let

$$\rho_{xy} = P_x\left(T_y < \infty\right) \tag{19.7}$$

and we have

$$P_x\left(T_y^k < \infty\right) = \rho_{xy}\rho_{yy}^{k-1} \tag{19.8}$$

Proof. □

Let

$$N(y) = \sum_{n=1}^{\infty} 1_{(X_n = y)} \tag{19.9}$$

be the number of visits to y at positive times.

Definition 19.3.1 (Recurrent)

A state y is said to be recurrent if $\rho_{yy} = 1$.

Property. The recurrent state y has the following properties

1. y is recurrent if and only if

$$E_y N(y) = \infty.$$

2. If x is recurrent and $\rho_{xy} > 0$, then y is recurrent and $\rho_{yx} = 1$.

Definition 19.3.2

A state y is said to be transient if $\rho_{yy} < 1$.

Property. The transient state y has the following properties

1. If y is transient, then
$$E_x N(y) < \infty, \quad \forall x.$$

Proof.
$$\begin{aligned} E_x N(y) &= \sum_{k=1}^{\infty} P_x(N(y) \geq k) = \sum_{k=1}^{\infty} P_x\left(T_y^k < \infty\right) \\ &= \sum_{k=1}^{\infty} \rho_{xy}\rho_{yy}^{k-1} = \frac{\rho_{xy}}{1-\rho_{yy}} < \infty \end{aligned}$$

□

Definition 19.3.3 (Closed State Set)

A set C of states is said to be closed, if
$$x \in C, \rho_{xy} > 0 \Rightarrow y \in C. \tag{19.10}$$

Definition 19.3.4 (Irreducible State Set)

A set D of states is said to be irreducible, if
$$x, y \in D \Rightarrow \rho_{xy} > 0. \tag{19.11}$$

Theorem 19.3.1

Let C be a finite closed set, then

1. C contains a recurrent state.
2. If C is irreducible, then all states in C are recurrent.

Theorem 19.3.2

Suppose $C_x = \{y : \rho_{xy} > 0\}$, then C_x is an irreducible closed set.

Proof. If $y, z \in C_x$, then $\rho_{yz} \geq \rho_{yx}\rho_{xz} > 0$. If $\rho_{yw} > 0$, then $\rho_{xw} \geq \rho_{xy}\rho_{yw} > 0$, so $w \in C_x$. □

Example (A Seven-state Chain)**.** Consider the transition probability,

	1	**2**	**3**	**4**	**5**	**6**	**7**
1	0.3	0	0	0	0.7	0	0
2	0.1	0.2	0.3	0.4	0	0	0
3	0	0	0.5	0.5	0	0	0
4	0	0	0	0.5	0	0.5	0
5	0.6	0	0	0	0.4	0	0
6	0	0	0	0.1	0	0.1	0.8
7	0	0	0	1	0	0	0

try to identify the recurrent states and those that are transient.

Proof. $\{2, 3\}$ are transition states, and $\{1, 4, 5, 6, 7\}$ are recurrent states. □

Remark. Suppose S is finite, for $x \in S$,

1. x is transient, if
$$\exists y, \rho_{xy} > 0, \text{ s.t. } \rho_{yx} = 0$$
2. x is recurrent, if
$$\forall y, \rho_{xy} > 0, \text{ s.t. } \rho_{yx} > 0$$

19.4 Stationary Measures

19.5 Asymptotic Behavior

19.6 Ergodic Theorems

Definition 19.6.1 (Stationary Sequence)

Theorem 19.6.1 (Ergodic Theorem)

Example.

Chapter 20

Brownian Motion

Definition 20.0.1 (Brownian Motion (1))

A real-valued stochastic process $B(t), t \geq 0$ is said to be Brownian motion, if

1. for any $0 = t_0 \leq t_1 \leq \cdots \leq t_n$ the increments

$$B(t_1) - B(t_0), \ldots, B(t_n) - B(t_{n-1})$$

are independent

2. for any $s, t \geq 0$ and Borel sets $A \in \mathbb{R}$,

$$P(B(s+t) - B(s) \in A) = \int_A (2\pi t)^{-1/2} \exp\left(-x^2/2t\right) \mathrm{d}x \tag{20.1}$$

3. the sample paths $t \to B(t)$ are a.s. continuous.

Property. For a one-dimensional Brownian motion, if $B(0) = 0$, then we have the following properties

1. $EB_t = 0, \mathrm{Var}(B_t) = t, \quad t \geq 0.$
2. $\mathrm{Cov}(B_s, B_t) = s, \mathrm{Corr}(B_s, B_t) = \sqrt{\frac{s}{t}}, \quad \forall 0 \leq s \leq t.$

Proof. 1. Since $B_t = B_t - B_0 \sim N(0, t)$, then we have

$$EB_t = 0, \mathrm{Var}(B_t) = t$$

2. Suppose $0 \leq s \leq t$,

$$\mathrm{Cov}(B_s, B_t) = E[(B_s - EB_s)(B_t - EB_t)] = EB_sB_t$$

Let $B_t = (B_t - B_s) + B_s$, we have

$$\begin{aligned} EB_sB_t &= E[B_s \cdot ((B_t - B_s) + B_s)] \\ &= E[B_s \cdot (B_t - B_s)] + EB_s^2 \end{aligned}$$

Since $B_s = B_s - B_0$ and $B_t - B_s$ are independent,

$$E\left[B_s \cdot (B_t - B_s)\right] = EB_s \cdot E\left[B_t - B_s\right] = 0$$

Thus

$$\mathrm{Cov}\left(B_s, B_t\right) = EB_s^2 = s$$

And

$$\mathrm{Corr}\left(B_s, B_t\right) = \frac{\mathrm{Cov}\left(B_s, B_t\right)}{\sigma_{B_s}\sigma_{B_t}} = \frac{s}{\sqrt{st}} = \sqrt{\frac{s}{t}}$$

□

A second equivalent definition of Brownian motion is as follows,

Definition 20.0.2 (Brownian Motion (2))

A real-valued stochastic process $B(t), t \geq 0$, **starting from** 0, is said to be Brownian motion, if

1. $B(t)$ is a Gaussian process[a]
2. $\forall s, t \geq 0, EB_s = 0$ and $EB_sB_t = s \wedge t$
3. the sample paths $t \to B(t)$ are a.s. continuous

[a] Gaussian process, i.e., all its finite-dimensional distributions are multivariate normal.

20.1 Markov Properties

20.2 Martingales

Example (Quadratic Martingale)**.** Suppose B_t is a Brownian motion, then $B_t^2 - t$ is a martingale.

Proof. Let $B_t^2 = (B_s + B_t - B_s)^2$, we have

$$\begin{aligned} E_x\left(B_t^2 \mid \mathcal{F}_s\right) &= E_x\left(B_s^2 + 2B_s\left(B_t - B_s\right) + \left(B_t - B_s\right)^2 \mid \mathcal{F}_s\right) \\ &= B_s^2 + 2B_sE_x\left(B_t - B_s \mid \mathcal{F}_s\right) + E_x\left(\left(B_t - B_s\right)^2 \mid \mathcal{F}_s\right) \\ &= B_s^2 + 0 + (t - s) \end{aligned}$$

since $B_t - B_s$ is independent of $\mathcal{F}_s$ and has mean 0 and variance $t - s$. □

Example (Exponential Martingale)**.** Suppose B_t is a Brownian motion, then $\exp\left(\theta B_t - (\theta^2 t/2)\right)$ is a martingale.

Proof. Let $B_t = B_t - B_s + B_s$, then

$$\begin{aligned} E_x\left(\exp\left(\theta B_t\right) \mid \mathcal{F}_s\right) &= \exp\left(\theta B_s\right) E\left(\exp\left(\theta\left(B_t - B_s\right)\right) \mid \mathcal{F}_s\right) \\ &= \exp\left(\theta B_s\right) \exp\left(\theta^2(t-s)/2\right) \end{aligned}$$

since $B_t - B_s$ is independent of $\mathcal{F}_s$ and has mean 0 and variance $t - s$. Thus

$$\begin{aligned} E_x\left(\exp\left(\theta B_t - \left(\theta^2 t/2\right)\right) \mid \mathcal{F}_s\right) &= E_x\left(\exp\left(\theta B_t\right) \mid \mathcal{F}_s\right) \cdot \exp\left(-\left(\theta^2 t/2\right)\right) \\ &= \exp\left(\theta B_s - \left(\theta^2 s/2\right)\right) \end{aligned}$$

□

Theorem 20.2.1 (Lévy's Martingale Characterization)

Let $B(t), t \geq 0$, be a real-valued stochastic process and let $\mathcal{F}_t = \sigma\left(B_s, s \leq t\right)$ be the filtration generated by it. Then $B(t)$ is a Brownian motion if and only if

1. $B(0) = 0$ a.s.
2. the sample paths $t \to B(t)$ are continuous a.s.
3. $B(t)$ is a martingale with respect to $\mathcal{F}_t$
4. $|B(t)|^2 - t$ is a martingale with respect to $\mathcal{F}_t$

20.3 Sample Paths

Let $0 = t_0^n < t_1^n < \cdots < t_n^n = T$, where $t_i^n = \frac{iT}{n}$ be a partition of the interval $[0, T]$ into n equal parts, and

$$\Delta_i^n B = B\left(t_{i+1}^n\right) - B\left(t_i^n\right) \tag{20.2}$$

be the corresponding increments of the Brownian motion $B(t)$.

Theorem 20.3.1

$$\lim_{n\to\infty} \sum_{i=0}^{n-1} \left(\Delta_i^n B\right)^2 = T \quad \text{in} \quad L^2 \tag{20.3}$$

Proof. Since the increments $\Delta_i^n B$ are independent and

$$E\left(\Delta_i^n B\right) = 0, \quad E\left(\left(\Delta_i^n B\right)^2\right) = \frac{T}{n}, \quad E\left(\left(\Delta_i^n B\right)^4\right) = \frac{3T^2}{n^2}$$

it follows that

$$\begin{aligned}
E\left(\left[\sum_{i=0}^{n-1}\left(\Delta_i^n B\right)^2 - T\right]^2\right) =& E\left(\left[\sum_{i=0}^{n-1}\left(\left(\Delta_i^n B\right)^2 - \frac{T}{n}\right)\right]^2\right) \\
=& \sum_{i=0}^{n-1} E\left[\left(\left(\Delta_i^n B\right)^2 - \frac{T}{n}\right)^2\right] \\
=& \sum_{i=0}^{n-1}\left[E\left(\left(\Delta_i^n B\right)^4\right) - \frac{2T}{n}E\left(\left(\Delta_i^n B\right)^2\right) + \frac{T^2}{n^2}\right] \\
=& \sum_{i=0}^{n-1}\left[\frac{3T^2}{n^2} - \frac{2T^2}{n^2} + \frac{T^2}{n^2}\right] \\
=& \frac{2T^2}{n} \to 0, \quad n \to \infty
\end{aligned}$$

□

Definition 20.3.1 (Variation)

The variation of a function $f : [0, T] \to \mathbb{R}$ is defined to be

$$\limsup_{\Delta t \to 0} \sum_{i=0}^{n-1} \left|f\left(t_{i+1}\right) - f\left(t_i\right)\right| \tag{20.4}$$

where $t = (t_0, t_1, \ldots, t_n)$ is a partition of $[0, T]$, i.e. $0 = t_0 < t_1 < \cdots < t_n = T$, and where

$$\Delta t = \max_{i=0,\ldots,n-1} |t_{i+1} - t_i| \tag{20.5}$$

Theorem 20.3.2

The variation of the paths of $B(t)$ is infinite a.s..

Proof. Consider the sequence of partitions $t^n = (t_0^n, t_1^n, \ldots, t_n^n)$ of $[0, T]$ into n equal parts. Then

$$\sum_{i=0}^{n-1} \left|\Delta_i^n B\right|^2 \leq \left(\max_{i=0,\ldots,n-1} \left|\Delta_i^n B\right|\right) \sum_{i=0}^{n-1} \left|\Delta_i^n B\right|$$

Since the paths of $B(t)$ are a.s. continuous on $[0, T]$,

$$\lim_{n\to\infty}\left(\max_{i=0,\ldots,n-1} \left|\Delta_i^n B\right|\right) = 0 \quad \text{a.s.}$$

By Theorem 20.3.1, we have

$$\lim_{n\to\infty} \sum_{i=0}^{n-1} \left(\Delta_i^n B\right)^2 = T \text{ in } L^2$$

Since every sequence of random variables convergent in L^2 has a subsequence convergent a.s. There is a subsequence $t^{n_k} = \left(t_0^{n_k}, t_1^{n_k}, \ldots, t_{n_k}^{n_k}\right)$ of partitions such that

$$\lim_{k\to\infty} \sum_{i=0}^{n_k-1} \left|\Delta_i^{n_k} B\right|^2 = T \quad \text{a.s.}$$

Since

$$\sum_{i=0}^{n_k-1} \left|\Delta_i^{n_k} B\right| \geq \frac{\sum_{i=0}^{n_k-1} \left|\Delta_i^{n_k} B\right|^2}{\max_{i=0,\ldots,n_k-1} \left|\Delta_i^{n_k} B\right|}$$

hence,

$$\lim_{k\to\infty} \sum_{i=0}^{n_k-1} \left|\Delta_i^{n_k} B\right| = \infty \quad \text{a.s.}$$

while

$$\lim_{k\to\infty} \Delta t^{n_k} = \lim_{k\to\infty} \frac{T}{n_k} = 0$$

□

20.4 Itô Stochastic Calculus

Definition 20.4.1 (Itô Stochastic Integral)

For any $T > 0$ we shall denote by M_T^2 the space of all stochastic processes $f(t), t \geq 0$ such that

$$1_{[0,T)} f \in M^2$$

The Itô stochastic integral (from 0 to T) of $f \in M_T^2$ is defined by

$$I_T(f) = I\left(1_{[0,T)} f\right) \tag{20.6}$$

which can be denoted by

$$\int_0^T f(t)\,\mathrm{d}B(t) \tag{20.7}$$

Property. The Itô Stochastic Integral has the following properties:

1. Linearity: For $\forall f, g \in M_t^2$, $\forall \alpha, \beta \in \mathbb{R}$,

$$\int_0^t (\alpha f(r) + \beta g(r))\,\mathrm{d}B(r) = \alpha \int_0^t f(r)\,\mathrm{d}B(r) + \beta \int_0^t g(r)\,\mathrm{d}B(r) \tag{20.8}$$

2. Isometry: For $\forall f \in M_t^2$,

$$E\left(\left|\int_0^t f(r)\,\mathrm{d}B(r)\right|^2\right) = E\left(\int_0^t |f(r)|^2\,\mathrm{d}r\right) \tag{20.9}$$

3. Martingale Property: For $\forall f \in M_t^2$ and $\forall 0 \leq s < t$,

$$E\left(\int_0^t f(r)\,\mathrm{d}B(r) \mid \mathcal{F}_s\right) = \int_0^s f(r)\,\mathrm{d}B(r) \tag{20.10}$$

Proof. □

Definition 20.4.2 (Itô Process)

A stochastic process $\xi(t), t \geq 0$ is said to be an Itô process if it has a.s. continuous paths and can be represented as

$$\xi(T) = \xi(0) + \int_0^T a(t)\,\mathrm{d}t + \int_0^T b(t)\,\mathrm{d}B(t) \quad \text{a.s.} \tag{20.11}$$

where $b(t)$ is a process belonging to M_T^2 for all $T > 0$ and $a(t)$ is a process adapted to the filtration $\mathcal{F}_t$ such that

$$\int_0^T |a(t)|\,\mathrm{d}t < \infty \quad \text{a.s.} \tag{20.12}$$

for all $T \geq 0$. The Itô process is denoted by

$$\mathrm{d}\xi(t) = a(t)\,\mathrm{d}t + b(t)\,\mathrm{d}B(t) \tag{20.13}$$

Remark. The class of all adapted processes $a(t)$ satisfying 20.12 for some $T > 0$ will be denoted by $\mathcal{L}_T^1$.

Theorem 20.4.1 (Itô Formula)

Suppose $F(t,x)$ is a real-valued function with continuous partial derivatives $F'_t(t,x)$, $F'_x(t,x)$ and $F''_{xx}(t,x)$ for all $t \geq 0$ and $x \in \mathbb{R}$.

1. If $\xi(t)$ be an Itô process

$$\xi(t) = \xi(0) + \int_0^t a(s)\,\mathrm{d}s + \int_0^t b(s)\,\mathrm{d}B(s)$$

and the process $b(t)F'_x(t,\xi(t))$ belongs to M_T^2 for all $T \geq 0$. Then $F(t,\xi(t))$ is an Itô process such that

$$\begin{aligned}\mathrm{d}F(t,\xi(t)) = &\left(F'_t(t,\xi(t)) + F'_x(t,\xi(t))a(t) + \frac{1}{2}F''_{xx}(t,\xi(t))b(t)^2\right)\mathrm{d}t \\ &+ F'_x(t,\xi(t))b(t)\,\mathrm{d}B(t)\end{aligned} \tag{20.14}$$

2. If $\xi(t)$ be an Brownian Motion, such that $\xi(t) = B(t)$, and the process $F'_x(t,B(t))$ belongs to M_T^2 for all $T \geq 0$. Then $F(t,B(t))$ is an Itô process such that

$$\mathrm{d}F(t,B(t)) = \left(F'_t(t,B(t)) + \frac{1}{2}F''_{xx}(t,B(t))\right)\mathrm{d}t + F'_x(t,B(t))\,\mathrm{d}B(t) \tag{20.15}$$

Example (Exponential Martingale)**.** Show that the exponential martingale

$$X(t) = \mathrm{e}^{B(t)}\mathrm{e}^{-\frac{t}{2}}$$

is an Itô process, and satisfies the equation

$$\mathrm{d}X(t) = X(t)\,\mathrm{d}B(t)$$

Proof. Let $F(t,x) = \mathrm{e}^x\mathrm{e}^{-\frac{t}{2}}$, then we have

$$F'_t(t,x) = -\frac{1}{2}F(t,x), \quad F'_x(t,x) = F(t,x), \quad F''_{xx}(t,x) = F(t,x)$$

thus, by Itô Formula, we have

$$\begin{aligned}\mathrm{d}X(t) =& \mathrm{d}F(t,B(t)) = \left(F'_t(t,B(t)) + \frac{1}{2}F''_{xx}(t,B(t))\right)\mathrm{d}t + F'_x(t,B(t))\,\mathrm{d}B(t) \\ =& \left(-\frac{1}{2}F(t,B(t)) + \frac{1}{2}F(t,B(t))\right)\mathrm{d}t + F(t,B(t))\,\mathrm{d}B(t) \\ =& X(t)\,\mathrm{d}B(t)\end{aligned}$$

□

Example.

Example.

Chapter 21

Exercises for Probability Theory and Examples

Part VIII

Empirical Process

Chapter 22

22.1 Concentration by Entropic Techniques

Definition 22.1.1 (Entropy)

The entropy of a random variable X for the convex function $\phi(\cdot)$ is defined as

$$H_\phi(X) = \mathbb{E}[\phi(X)] - \phi(\mathbb{E}[X]) \tag{22.1}$$

Example.
1. For $\phi(u) = u^2$, $H_\phi(X) = \mathrm{Var}(X)$
2. For $\phi(u) = -\log(u)$ $(u > 0)$, and for X real-valued random variable, we have that $Z := \exp(\lambda X)$ is a non-negative random variable, and

$$H_\phi(Z) = -\lambda\mathbb{E}[X] + \log(\mathbb{E}[\exp(\lambda X)]) = \log \mathbb{E}e^{\lambda(X - \mathbb{E}[X])}. \tag{22.2}$$

Definition 22.1.2

Let Ω be a finate sample space and denote $\mathcal{M}(\Omega)$ as the set of all probability measures (vectors) on Ω.

1. The **relative entropy** with respect to $q \in \mathcal{M}(\Omega)$ is defined as the mapping $H(\cdot \mid q) : \mathcal{M}(\Omega) \to [0, \infty] : p \mapsto H(p \mid q)$, where

$$H(p \mid q) = \begin{cases} \sum_{\omega\in\Omega} p(\omega) \log\left(\frac{p(\omega)}{q(\omega)}\right) & \text{if } p \ll q \\ +\infty & \text{otherwise} \end{cases} \tag{22.3}$$

2. The **Shannon entropy** of a Ω-valued random variable X with distribution $p \in \mathcal{M}(\Omega)$ is defined as

$$\mathcal{H}(p) = -\sum_{\omega\in\Omega} p(\omega) \log(p(\omega)). \tag{22.4}$$

Proposition 22.1.1 (Duality formula of the Entropy)

Lemma 22.1.1

Let Ω be a finite sample space and denote $\mathcal{M}(\Omega)$ as the set of probability measures on Ω. Let $q \in \mathcal{M}(\Omega)$, then the relative entropy $H(\cdot \mid q)$ is strictly convex, continuous and

$$H(p \mid q) = 0 \iff p = q. \tag{22.5}$$

Proof. To prove this theorem, we analyze the properties of the relative entropy (also known as the Kullback-Leibler divergence). Let Ω be a finite sample space, $\mathcal{M}(\Omega)$ denote the set of probability measures on Ω, and let $q \in \mathcal{M}(\Omega)$ be a fixed probability measure. For any $p \in \mathcal{M}(\Omega)$, the relative entropy $H(p \mid q)$ is defined as

$$H(p \mid q) = \sum_{\omega \in \Omega} p(\omega) \log \frac{p(\omega)}{q(\omega)},$$

where we assume $q(\omega) > 0$ for all $\omega \in \Omega$ (otherwise, the term is taken as 0).

Non-negativity of Relative Entropy First, observe that for any $p \in \mathcal{M}(\Omega)$, by Jensen's inequality, we have

$$\sum_{\omega \in \Omega} p(\omega) \log \frac{p(\omega)}{q(\omega)} \geq 0,$$

with equality if and only if $p = q$. Thus, we obtain $H(p \mid q) \geq 0$, and $H(p \mid q) = 0 \iff p = q$. This proves

$$H(p \mid q) = 0 \iff p = q.$$

Strict Convexity of Relative Entropy To prove that $H(\cdot \mid q)$ is strictly convex, we consider the dependence of the relative entropy $H(p \mid q)$ on p. Let $p_1, p_2 \in \mathcal{M}(\Omega)$ with $p_1 \neq p_2$. For $0 < \lambda < 1$, define

$$p_\lambda = \lambda p_1 + (1 - \lambda) p_2.$$

Using the definition of relative entropy, we have

$$H(p_\lambda \mid q) = \sum_{\omega \in \Omega} p_\lambda(\omega) \log \frac{p_\lambda(\omega)}{q(\omega)}.$$

Applying Jensen's inequality for strictly convex functions, we obtain

$$H(p_\lambda \mid q) < \lambda H(p_1 \mid q) + (1 - \lambda) H(p_2 \mid q).$$

This shows that $H(\cdot \mid q)$ is strictly convex.

Continuity of Relative Entropy Finally, we prove that the relative entropy $H(p \mid q)$ is continuous with respect to p. Since Ω is finite, in this finite-dimensional vector space, $p \mapsto H(p \mid q)$ is a sum of a finite number of terms, each of which is a continuous function of $p(\omega)\log(p(\omega)/q(\omega))$. Thus, $H(p \mid q)$ is continuous with respect to p.

□

$$H(\exp(\lambda X)) = \lambda M_X'(\lambda) - M_X(\lambda)\log M_X(\lambda). \tag{22.6}$$

Proposition 22.1.2 (Herbst argument)

Let X be a random variable and suppose that for $\sigma > 0$,

$$\mathcal{H}(\mathrm{e}^{\lambda X}) \leq \frac{\lambda^2\sigma^2}{2} M_X(\lambda), \tag{22.7}$$

for $\lambda \in I$ with interval I being either $\mathbb{R}$ or $[0, \infty)$. Then,

$$\log \mathbb{E}\mathrm{e}^{\lambda(X-\mathbb{E}[X])} \leq \frac{\lambda^2\sigma^2}{2}, \quad \forall \lambda \in I. \tag{22.8}$$

Proposition 22.1.3 (Bernstein entropy bound)

Suppose there exists $B > 0$ and $\sigma > 0$ such that

Definition 22.1.3 (Separately convex)

A function $f : \mathbb{R}^n \to \mathbb{R}$ is said to be **separately convex** if for all $k \in [n]$, the function $f(\cdot, x_{-k})$ is convex for all $x_{-k} \in \mathbb{R}^{n-1}$.

Definition 22.1.4 (Lipschitz continuous)

Definition 22.1.5 (Globally Lipschitz continuous)

Lemma 22.1.2 (Entropy bound for univariate functions)

Let X and Y two independent, identically distributed $\mathbb{R}$-valued random variables. Denote by $\mathbb{E}_{X,Y}$ the expectation with respect to X and Y. For any function $g : \mathbb{R} \to \mathbb{R}$ the following statements hold:

1. $\forall \lambda > 0$, $\mathcal{H}\left(\mathrm{e}^{\lambda g(X)}\right) \leq \lambda^2 \mathbb{E}_{X,Y}\left[(g(X) - g(Y))^2 \mathrm{e}^{\lambda g(X)} \mathbf{I}\{g(X) \geq g(Y)\}\right]$.
2. If in addition the random variable X is supported on $[a, b]$, $a < b$, and the function g is convex and Lipschitz continuous, then

$$\mathcal{H}\left(\mathrm{e}^{\lambda g(X)}\right) \leq \lambda^2 (b-a)^2 \mathbb{E}\left[(g'(X))^2 \mathrm{e}^{\lambda g(X)}\right], \quad \forall \lambda > 0.$$

Proof. Using the fact that X and Y are independent and identically distributed, we have

$$\mathcal{H}\left(\mathrm{e}^{\lambda g(X)}\right) = \mathbb{E}_X\left[\lambda g(X) \mathrm{e}^{\lambda g(X)}\right] - \mathbb{E}_X\left[\lambda g(X)\right] \log \mathbb{E}_Y \mathrm{e}^{\lambda g(Y)}.$$

□

Lemma 22.1.3 (Tensorisation of the entropy)

Let $X_1, \ldots, X_n$ be independent real-valued random variables and $f : \mathbb{R}^n \to \mathbb{R}$ be a given function. Then, for all $\lambda > 0$,

$$\mathcal{H}(\mathrm{e}^{\lambda f(X_1,\ldots,X_n)}) \leq \sum_{i=1}^{n} \mathcal{H}(\mathrm{e}^{\lambda f_i(X_i)} \mid \mathbf{X}_{-i}),$$

where $f_i : \mathbb{R} \to \mathbb{R}$ is defined as $f_i(x) = f(x_1, \ldots, x_{i-1}, x, x_{i+1}, \ldots, x_n)$.

Proof. According to the variational representation of the entropy, we have

$$\mathcal{H}(\mathrm{e}^{\lambda f(\mathbf{X})}) = \sup_{g \in \mathcal{G}} \left\{\mathbb{E}[g(\mathbf{X}) \mathrm{e}^{\lambda f(\mathbf{X})}]\right\},$$

where $\mathcal{G} = \{g : \Omega \to \mathbb{R} : \mathrm{e}^g \leq 1\}$.

For each $i \in [n]$, define $\mathbf{X}_i = (X_i, \ldots, X_n)$ and for any $g \in \mathcal{G}$ define g^i, $i \in [n]$:

$$\begin{aligned} g^1(\mathbf{X}) =& g(\mathbf{X}) - \log \mathbb{E}\left[\mathrm{e}^{g(\mathbf{X})} \mid \mathbf{X}_2\right], \\ g^i(\mathbf{X}_i) =& \log \frac{\mathbb{E}\left[\mathrm{e}^{g(\mathbf{X})} \mid \mathbf{X}_i\right]}{\mathbb{E}\left[\mathrm{e}^{g(\mathbf{X})} \mid \mathbf{X}_{i+1}\right]}, \quad i \in [n-1]. \end{aligned}$$

It is easy to see that by the above construction, we have

$$\sum_{i=1}^{n} g^i(\mathbf{X}) = g(\mathbf{X}) - \log \mathbb{E}\left[\mathrm{e}^{g(\mathbf{X})}\right] \geq g(\mathbf{X}), \tag{22.9}$$

and

$$\mathbb{E}\left[\exp\left(g^i(\mathbf{X}_i) \mid X_{i+1}\right)\right] = 1.$$

Within the variational representation of the entropy, we have

$$\begin{aligned}\mathbb{E}\left[g(\mathbf{X})\mathrm{e}^{\lambda f(\mathbf{X})}\right] &\underbrace{\leq}_{(22.9)} \mathbb{E}\left[\sum_{i=1}^{n} g^i(\mathbf{X})\mathrm{e}^{\lambda f(\mathbf{X})}\right] \\ &= \sum_{i=1}^{n} \mathbb{E}\left[g^i(\mathbf{X})\mathrm{e}^{\lambda f(\mathbf{X})}\right] \\ &\leq\end{aligned}$$

□

Theorem 22.1.1 (Tail-bound for Lipschitz functions)

Let $\mathbf{X} \in \mathbb{R}^n$ be a random vector with independent coordinates X_i supporrted on the interval $[a, b], a < b$ and let $f : \mathbb{R}^n \to \mathbb{R}$ be separately convex and L-Lipschitz continuous with respect to the Euclidean norm. Then, for all $t > 0$,

$$\mathbb{P}\left(f(\mathbf{X}) - \mathbb{E}[f(\mathbf{X})] \geq t\right) \leq \exp\left(-\frac{t^2}{4L^2(b-a)^2}\right). \tag{22.10}$$

Proof. For $i \in [n]$, and every $\mathbf{x}_{-i} \in \mathbb{R}^{n-1}$, the function $f_i(x)$ is convex, and thus by Lemma 22.1.2, for all $\lambda > 0$ that for every fixed $\mathbf{x}_{-i}$, we have

$$\mathcal{H}\left(\mathrm{e}^{\lambda f_i(X_i)} \mid \mathbf{X}_{-i}\right) \leq \lambda^2(b-a)^2\mathbb{E}\left[\left(f_i'(X_i)\right)^2 \mathrm{e}^{\lambda f_i(X_i)} \mid \mathbf{X}_{-i}\right].$$

□

Example (Operator norm of a random matrix)**.** Let $M \in \mathbb{R}^{n\times d}$ be a random matrix with independent identically distributed mean-zero random entries M_{ij} supported on the interval $[-1, 1]$.

$$\|M\| = \max_{v\in\mathbb{S}^{d-1}} \|Mv\|_2 = \max_{u\in\mathbb{S}^{n-1}, v\in\mathbb{S}^{d-1}} \langle u, Mv\rangle. \tag{22.11}$$

The operator norm is maximin/supermum of functions that are linear in the entries of M, and thus any such function is convex and as such separately convex.

Moreover, for any $M, M' \in \mathbb{R}^{n\times d}$, we have

$$\begin{aligned}
|\|M\| - \|M'\|| &= \left| \max_{u\in\mathbb{S}^{n-1}, v\in\mathbb{S}^{d-1}} \langle u, Mv\rangle - \max_{u\in\mathbb{S}^{n-1}, v\in\mathbb{S}^{d-1}} \langle u, M'v\rangle \right| \\
&\le \max_{u\in\mathbb{S}^{n-1}, v\in\mathbb{S}^{d-1}} |\langle u, Mv\rangle - \langle u, M'v\rangle| \\
&\le \max_{u\in\mathbb{S}^{n-1}, v\in\mathbb{S}^{d-1}} \|u\|_2\|v\|_2\|M - M'\|_F \\
&\le \|M - M'\|_F,
\end{aligned}$$

where $\|\cdot\|_F$ denotes the Frobenius norm, which is equivalent to the Euclidean norm for matrices. Thus, the operator norm is 1-Lipschitz continuous with respect to the Euclidean norm. Then by Theorem 22.1.1, we have

$$\mathbb{P}\left(\|M\| - \mathbb{E}[\|M\|] \ge t\right) \le \exp\left(-\frac{t^2}{16}\right). \tag{22.12}$$

Proof. □

22.2 Some Matrix Calculus and Covariance Estimation

Theorem 22.2.1 (Matrix Bernstein Inequality)

Let $X_1, \dots, X_n$ be independent, mean-zero random systematic matrices in $\mathbb{R}^{d\times d}$ such that $\|X_i\| \le K$ almost surely for all $i \in [n]$. Then for all $t > 0$, we have

$$\mathbb{P}\left(\left\|\sum_{i=1}^n X_i\right\| \ge t\right) \le 2d\exp\left(-\frac{t^2}{\sigma^2 + Kt/3}\right),$$

where $\sigma^2 = \|\sum_{i=1}^n \mathbb{E}[X_i^2]\|$ is the norm of the matrix variance of the sum.

Proof. Denote $S_n = \sum_{i=1}^n X_i$ and $\lambda_{\max}(S_n)$ as the largest eigenvalue of S_n. Then, we have

$$\|S_n\| = \max\{\lambda_{\max}(S_n), -\lambda_{\min}(S_n)\}.$$

Since

$$\mathbb{P}\left(\lambda_{\max}(S_n) \ge t\right) = \mathbb{P}\left(\exp(\lambda\lambda_{\max}(S_n)) \ge \exp(\lambda t)\right) \le \frac{\mathbb{E}[\exp(\lambda\lambda_{\max}(S_n))]}{\exp(\lambda t)}.$$

□

Lemma 22.2.1 (Bound on MGF)

Let X be an $d \times d$ symmetric mean-zero random matrix such that $\|X\| \leq K$ almost surely. Then, for $|\lambda| < 3/K$, we have

$$\mathbb{E}[\exp(\lambda X)] \preceq \exp\left(g(\lambda)\mathbb{E}[X^2]\right),$$

where $g(\lambda) = \frac{\lambda^2/2}{1-|\lambda|K/3}$.

Proof. □

Proposition 22.2.1 (Expectation Bound via the Bernstein Inequality)

Under the conditions of Theorem 22.2.1, we have the tail bound

$$\mathbb{P}\left(\left\|\sum_{i=1}^{n} X_i\right\| \geq t\right) \leq 2d \exp\left(-\frac{t^2}{\sigma^2 + Kt/3}\right).$$

Then,

$$\mathbb{E}\left[\left\|\sum_{i=1}^{n} X_i\right\|\right] \leq$$

Chapter 23

Basic Tools in High-dimensional Probability

23.1 Decoupling

Definition 23.1.1 (Chaos)

Let $X_1, \dots, X_n$ be independent real-valued random variables, and $a_{ij} \in \mathbb{R}$, $i, j \in [n]$. The random quadratic form

$$\sum_{i,j=1}^{n} a_{ij} X_i X_j = X^\top A X, \quad X = (X_1, \dots, X_n)^\top \in \mathbb{R}^n, \quad A = (a_{ij}) \in \mathbb{R}^{n \times n},$$

is called **chaos** in probability theory.

For simplicity, we assume that the random variables X_i have mean zero and unit variance, i.e., $\mathbb{E}[X_i] = 0$ and $\mathbb{E}[X_i^2] = 1$, for all $i \in [n]$. Then,

$$\mathbb{E}\left[X^\top A X\right] = \sum_{i,j=1}^{n} a_{ij} \mathbb{E}[X_i X_j] = \mathrm{Tr}(A).$$

We shall study concentration properties for chaos. This time we need to develop tools to overcome the fact that we have sums of not necessarily independent random variables. The idea is to use the decoupling technique. The idea is to study the following random quadratic form,

$$\sum_{i,j=1}^{n} a_{ij} X_i X_j' = X^\top A X' = \langle X, A X' \rangle,$$

where $X' = (X_1', \dots, X_n')^\top$ is an independent copy of X, and condition on X. Obvious, the bilinear form is easier to handle, e.g., when we condition on X', we

simply obtain a linear form in X, and vice versa, i.e.,

$$\langle X, AX' \rangle = \sum_{i=1}^{n} c_i X_i, \quad c_i = \sum_{j=1}^{n} a_{ij} X'_j,$$

is a random linear form in X depending on the condition of the independent copy X'.

Theorem 23.1.1 (Decoupling)

Let A be an $n \times n$ diagonal free matrix, $X = (X_1, \ldots, X_n) \in \mathbb{R}^n$ be a random vector with independent mean-zero coordinates X_i, and X' be an independent copy of X. Then, for every convex function $f : \mathbb{R} \to \mathbb{R}$, we have

$$\mathbb{E}[f(\langle X, AX \rangle)] \leq \mathbb{E}[f(4\langle X, AX' \rangle)].$$

Proof. The idea is to study the partial chaos

$$\langle X, AX \rangle = \sum_{(i,j) \in I \times I^{\complement}} a_{ij} X_i X_j,$$

with a random subset $I \subset [n]$. Let δ_i be independent Bernoulli random variables with $\mathbb{P}[\delta_i = 0] = \mathbb{P}[\delta_i = 1] = 1/2$, and define the random subset $I = \{i \in [n] : \delta_i = 1\}$. Then, we condition on X, and $\square$

Theorem 23.1.2 (Hanson-Wright Inequality)

Let $X = (X_1, \ldots, X_n) \in \mathbb{R}^n$ be a random vector with independent mean-zero sub-Gaussian coordinates X_i and let $A \in \mathbb{R}^{n \times n}$ be a deterministic matrix. Then, for all $t \geq 0$, we have

$$\mathbb{P}\left(|\langle X, AX \rangle - \mathbb{E}[\langle X, AX \rangle]| \geq t\right) \leq 2 \exp\left(-c \min\left(\frac{t^2}{K^4 \|A\|_F^2}, \frac{t}{K^2 \|A\|}\right)\right),$$

where $K := \max_{i \in [n]} \|X_i\|_{\psi_2}$, and $c > 0$ is a constant.

We prepare the proof of the Hanson-Wright inequality by proving the following lemmas.

Lemma 23.1.1 (MGF of Gaussian Chaos)

Let $X, X' \sim \mathcal{N}(0, I_n)$, X and X' be independent, and let $A \in \mathbb{R}^{n\times n}$ be a deterministic matrix. Then,

$$\mathbb{E}[\exp(\lambda\langle X, AX'\rangle)] \leq \exp\left(C\lambda^2\|A\|_F^2\right), \quad \forall|\lambda| \leq \frac{C}{\|A\|},$$

where $C > 0$ is a constant.

Proof. We use the singular value decomposition of the matrix A, i.e.,

$$A = \sum_{i=1}^{n} \sigma_i u_i v_i^\top,$$

then,

$$\langle X, AX'\rangle = \sum_{i=1}^{n} \sigma_i \langle X, u_i\rangle\langle X', v_i\rangle = \sum_{i=1}^{n} \sigma_i Y_i Y_i',$$

where $Y = (\langle X, u_1\rangle, \ldots, \langle X, u_n\rangle) \sim \mathcal{N}(0, I_n)$ and $Y' = (\langle X', v_1\rangle, \ldots, \langle X', v_n\rangle) \sim \mathcal{N}(0, I_n)$ are independent Gaussian vectors. Then independence of the Gaussian vectors Y and Y' implies that

$$\mathbb{E}[\exp(\lambda\langle X, AX'\rangle)] = \mathbb{E}[\exp(\lambda\sum_{i=1}^{n} \sigma_i Y_i Y_i')] = \prod_{i=1}^{n} \mathbb{E}[\exp(\lambda\sigma_i Y_i Y_i')].$$

For each $i \in [n]$, we compute the expectation with respect to Y', i.e., the conditional expectation holding the random vector Y fixed,

$$\mathbb{E}_Y\left[\mathbb{E}_{Y'}\left[\exp(\lambda\sigma_i Y_i Y_i') \mid Y\right]\right] = \mathbb{E}_Y\left[\exp\left(\frac{\lambda^2\sigma_i^2 Y_i^2}{2}\right)\right] = \exp\left(\frac{C\lambda^2\sigma_i^2}{2}\right), \quad \forall\lambda \leq \frac{C}{\sigma_i},$$

where the first equality follows from the fact that the MGF of a Gaussian random variable is given by $\mathbb{E}[\exp(\lambda Y)] = \exp(\lambda^2/2)$, and the second equality follows from the fact that the random variable Y_i is Gaussian and thus sub-Gaussian, and therefore Y_i^2 is sub-exponential, and thus gives the bound. Then, we obtain

$$\mathbb{E}[\exp(\lambda\langle X, AX'\rangle)] = \prod_{i=1}^{n} \exp\left(\frac{C\lambda^2\sigma_i^2}{2}\right) = \exp\left(C\lambda^2\sum_{i=1}^{n}\sigma_i^2\right), \quad \forall\lambda \leq \frac{C}{\max_{i\in[n]}\sigma_i}.$$

Finally, we use the fact that the Frobenius norm of the matrix A is given by $\|A\|_F^2 = \sum_{i=1}^{n}\sigma_i^2$, and the operator norm of the matrix A is given by $\|A\| = \max_{i\in[n]}\sigma_i$, and thus we obtain the desired bound. □

Lemma 23.1.2 (Comparison)

Let X and X' be independent mean-zero sub-Gaussian random vectors in $\mathbb{R}^n$ with $\|X\|_{\psi_2} \le K$ and $\|X'\|_{\psi_2} \le K$. Furthermore, let Y and Y' be independent mean-zero Gaussian random vectors in $\mathbb{R}^n$ with $Y, Y' \sim \mathcal{N}(0, I_n)$, and let $A \in \mathbb{R}^{n\times n}$ be a deterministic matrix. Then, for all $\lambda \in \mathbb{R}$, we have

$$\mathbb{E}[\exp(\lambda\langle X, AX\rangle)] \le \mathbb{E}[\exp(CK^2\lambda\langle Y, AY\rangle)].$$

Proof. We condtion on X' and take the expectation with respect to X, then, $\langle X, AX'\rangle$ is conditionally sub-Gaussian and we have

$$\mathbb{E}_X[\exp(\lambda\langle X, AX'\rangle)] \le \mathbb{E}[\exp(CK^2\lambda\|AX'\|_2^2)], \quad \forall\lambda \in \mathbb{R}.$$

We now replace X by Y but still condition on X', then we have

$$\mathbb{E}_Y[\exp(\mu\langle Y, AX'\rangle)] \le \mathbb{E}[\exp(\mu^2\|AX'\|_2^2/2)], \quad \forall\mu \in \mathbb{R}.$$

Choosing $\mu = \sqrt{2C\lambda}K$, we can match our estimates to get

$$\mathbb{E}_X[\exp(\lambda\langle X, AX'\rangle)] \le \mathbb{E}_Y[\exp(CK^2\lambda\langle Y, AX'\rangle)] = \exp(CK^2\lambda\|AX'\|_2^2).$$

We can now take the expectation with respect to X' on both sides, repeate the same procedure for the X' and Y' to obtain the desired bound. □

Proof. Wihtout loss of generality, $K = 1$. It suffices to show the one sided bound. Denote

$$p = \mathbb{P}\left(\langle X, AX\rangle - \mathbb{E}[\langle X, AX\rangle] \ge t\right).$$

We can write

$$\langle X, AX\rangle = \sum_{i=1}^n a_{ii}X_i^2 + \sum_{i\ne j} a_{ij}X_iX_j,$$

and thus the problem reduces to bounding the tail of the diagonal sums and the off-diagonal sums separately:

$$p \le \mathbb{P}\left(\sum_{i=1}^n a_{ii}\left(X_i^2 - \mathbb{E}\left[X_i^2\right]\right) \ge t/2\right) + \mathbb{P}\left(\sum_{i\ne j} a_{ij}\left(X_iX_j\right) \ge t/2\right) := p_1 + p_2.$$

We shall bound the two terms separately. For the diagonal terms, since $X_i^2 - \mathbb{E}[X_i^2]$ are independent mean-zero sub-Eponential random variables, we have

$$\|X_i^2 - \mathbb{E}[X_i^2]\|_{\psi_1} \le C\|X_i^2\|_{\psi_1} \le C\|X_i\|_{\psi_2}^2 \le C,$$

and thus by the Bernstein inequality, we have

$$p_1 \le \exp\left(-C\min\left(\frac{t^2}{\|A\|_F^2}, \frac{t}{\|A\|}\right)\right).$$

For the off-diagonal terms, denote $S := \sum_{i\ne j} a_{ij}X_iX_j$, then we have □

23.2 Concentration for Anisotropic Random Vectors

Lemma 23.2.1

Let $B \in \mathbb{R}^{m \times n}$ be a deterministic matrix, and let $X \in \mathbb{R}^n$ be an isotropic random vector $\in \mathbb{R}^n$, then

$$\mathbb{E}\left[\|BX\|_2^2\right] = \|B\|_F^2.$$

Proof. We have

$$\mathbb{E}\left[\|BX\|_2^2\right] = \mathbb{E}\left[\sum_{i=1}^{m}\left(\sum_{j=1}^{n} b_{ij}X_j\right)^2\right] = \sum_{i=1}^{m}\sum_{j=1}^{n} b_{ij}^2\mathbb{E}[X_j^2] = \|B\|_F^2.$$

□

23.3 Symmetrisation

Definition 23.3.1 (Symmetric Random Variables)

A real-valued random variable X is said to be symmetric if X and $-X$ have the same distribution.

Chapter 24

Random Processes

24.1 Introduction

Definition 24.1.1

A **random process** is a collection of random variables $\{X(t)\}_{t\in T}$ defined on a common probability space, which are indexed by the elements of some set T.

Example. Here are some examples of random processes:

1. If $T = \mathbb{N}$, then $\{X(t)\}_{t\in T}$ with $X_n = \sum_{i=1}^{n} Z_i$ is a discrete time **random walk**, where $\{Z_i\}_{i\in\mathbb{N}}$ is a sequence of i.i.d. random variables.
2. A **Wiener process** $X = \{X(t)\}_{t\geq 0}$, also known as **Brownian motion**, is a continuous-time random walk. It is a continuous-time stochastic process with stationary and independent increments. The Wiener process can be defined as follows:
   (a) The process has continuous paths, i.e., $X : [0,\infty) \to \mathbb{R}$, $t \mapsto X(t)$ is almost surely continuous.
   (b) The increments are independent and satisfy $X(t) - X(s) \sim \mathcal{N}(0, t-s)$ for all $0 \leq s < t$.

Definition 24.1.2

1. **Covariance function**: For a random process $\{X(t)\}_{t\in T}$, the covariance function is defined as

$$\mathbf{\Sigma}(s,t) = \operatorname{cov}(X(t), X(s)), \quad t, s \in T. \tag{24.1}$$

2. **Increment**: For a random process $\{X(t)\}_{t\in T}$, the increment is defined as

$$d(t,s) = \|X(t) - X(s)\|_{L^2} = \left(\mathbb{E}\left[X(t) - X(s)\right]^2\right)^{1/2}, \quad t, s \in T. \tag{24.2}$$

Definition 24.1.3 (Gaussian Process)

A random process $\{X(t)\}_{t\in T}$ is a **Gaussian process** if for any finite collection of indices $T_0 \subset T$, the random vector $\{X(t)\}_{t\in T_0}$ is multivariate Gaussian. Equivalently, $X(t)$ is a Gaussian process if every linear combination of the random variables $\sum_{t\in T_0} a_t X(t)$ is normally distributed, where $a_t \in \mathbb{R}$.

Definition 24.1.4 (Canonical Gaussian Process)

A **canonical Gaussian process** is a Gaussian process with mean zero and covariance function $\mathbf{\Sigma}(s,t) = \operatorname{cov}(X(t), X(s)) = t \wedge s$.

Remark. The canonical Gaussian process can also be defined as $T \in \mathbb{R}^n$ and let $Y \sim \mathcal{N}(0, \mathbf{I})$, then $X(t) = \langle t, Y\rangle$, $t \in T \subset \mathbb{R}^n$.

The increment of a canonical Gaussian process is

$$d(t,s) = \|X(t) - X(s)\|_{L^2} = \left(\mathbb{E}\left[X(t) - X(s)\right]^2\right)^{1/2} = \left(\mathbb{E}\left[\langle t - s, Y\rangle\right]^2\right)^{1/2} = \|t - s\|_2$$

where $t, s \in T$.

Lemma 24.1.1

Let $X \in \mathbb{R}^n$ be a mean-zero Gaussian random vector. Then there exists points $t_1, \ldots, t_n \in \mathbb{R}^n$ such that

$$X := (\langle t_i, Y\rangle)_{i=1}^n \in \mathbb{R}^n, \quad Y \sim \mathcal{N}(0, \mathbf{I}).$$

Proof. Let $\mathbf{\Sigma}$ be the covariance matrix of X. Then

$$X = \mathbf{\Sigma}^{1/2} Y, \quad Y \sim \mathcal{N}(0, \mathbf{I}).$$

The coordinates of $\mathbf{\Sigma}^{1/2} Y$ are $\langle t_i, Y\rangle$, where t_i are the columns of $\mathbf{\Sigma}^{1/2}$. □

24.2 Slepian's Inequality

In many applications, we are interested in the supremum of a Gaussian process, i.e., $\sup_{t\in T} X(t)$. Slepian's inequality provides an upper bound for the tail probability of the supremum of a Gaussian process.

Theorem 24.2.1 (Slepian's Inequality)

Let $\{X(t)\}_{t\in T}$ and $\{Y(t)\}_{t\in T}$ be two mean-zero Gaussian processes. Assume that for all $t, s \in T$, we have

$$\mathbb{E}\left[X(t)^2\right] = \mathbb{E}\left[Y(t)^2\right], \quad \mathbb{E}\left[(X(t)-X(s))^2\right] \le \mathbb{E}\left[(Y(t)-Y(s))^2\right].$$

Then for every $\tau \in \mathbb{R}$, we have

$$\Pr\left(\sup_{t\in T} X(t) \ge \tau\right) \le \Pr\left(\sup_{t\in T} Y(t) \ge \tau\right). \tag{24.3}$$

Remark. Whenever (24.3) holds, we say that $\{X(t)\}_{t\in T}$ is stochastically dominated by $\{Y(t)\}_{t\in T}$.

The proof of Theorem 24.2.1 follows by combining the two versions of Slepian's inequality, which are stated below. To present these versions, we need to introduce the method of Gaussian integration by parts.

Definition 24.2.1 (Gaussian Interpolation)

Suppose T is finate, and let $\{X(t)\}_{t\in T}$ and $\{Y(t)\}_{t\in T}$ be two Gaussian processes (withouth loss of generality, we assume that X and Y are independent). Then the **Gaussian interpolation** of $\{X(t)\}_{t\in T}$ and $\{Y(t)\}_{t\in T}$ is defined as

$$Z(u) = \sqrt{u}X(u) + (1-\sqrt{u})Y(u), \quad u \in [0,1].$$

It is easy to see that the covariance function of $Z(u)$ is the interpolation of the covariance functions of $X(t)$ and $Y(t)$:

$$\mathbf{\Sigma}_Z(u) = u\mathbf{\Sigma}_X + (1-u)\mathbf{\Sigma}_Y, \quad u \in [0,1].$$

For a function $f : \mathbb{R}^n \to \mathbb{R}$, we shall study how the expectation $\mathbb{E}[f(Z)]$ changes with respect to $u \in [0,1]$. We have the following lemma:

Lemma 24.2.1 (Gaussian integration by part)

Suppose $\mathbf{X} \sim$

Lemma 24.2.2 (Slepian's Inequality, Functional Form)

$X(t), Y_t \in \mathbb{R}^n$

Proof. The key idea is to use Lemma 24.2.2 fpr some approximation of the maximum. □

Theorem 24.2.2 (Sudakov-Fernique Inequality)

Let $\{X(t)\}_{t\in T}$ and $\{Y(t)\}_{t\in T}$ be a mean-zero Gaussian process. Assume that for all $t, s \in T$, we have

$$\mathbb{E}\left[(X(t) - X(s))^2\right] \leq \mathbb{E}\left[(Y(t) - Y(s))^2\right].$$

Then,

$$\mathbb{E}\left[\sup_{t\in T} X(t)\right] \leq \mathbb{E}\left[\sup_{t\in T} Y(t)\right].$$

Proof. The idea is to usa an approximation of the supremum itself and not for the indicator function. For $\beta > 0$, we have

$$f(x) := \frac{1}{\beta} \log \sum_{i=1}^{n} \exp(\beta x_i), \quad x \in \mathbb{R}^n.$$

One can easily show that

$$f(x) \underset{\beta\to\infty}{\longrightarrow} \max_{i\in[n]} x_i.$$

□

24.3 The Supermum of a Process

Definition 24.3.1 (Canonical Metric)

Suppose $X = \{X(t)\}_{t\in T}$ is a random process with index set T. The **canonical metric** of X is defined as

$$d(t, s) = \|X(t) - X(s)\|_{L^2} = \left(\mathbb{E}\left[X(t) - X(s)\right]^2\right)^{1/2}, \quad t, s \in T. \tag{24.4}$$

Theorem 24.3.1 (Sudakov's Minorisation Inequality)

Let $X = \{X(t)\}_{t\in T}$ be a mean-zero Gaussian process. Then, for any $\varepsilon \geq 0$, we have

$$\mathbb{E}\left[\sup_{t\in T} X(t)\right] \geq C\varepsilon\sqrt{\log \mathcal{N}(T, d, \varepsilon)},$$

where d is the canonical metric of the process, $\mathcal{N}(T, d, \varepsilon)$ is the ε-covering number of T with respect to the metric d, i.e., the smallest number of balls of radius ε needed to cover T, and $C > 0$ is a universal constant.

Proof. □

In many application this geometric qun

Definition 24.3.2 (Gaussian Width)

The **Gaussian width** of a set $T \subset \mathbb{R}^n$ is defined as

$$W(T) = \mathbb{E}\left[\sup_{t\in T}\langle t, Y\rangle\right], \tag{24.5}$$

where $Y \sim \mathcal{N}(0, \mathbf{I}_n)$.

Proposition 24.3.1 (Properties of Gaussian Width)

The Gaussian width of a set $T \subset \mathbb{R}^n$ satisfies the following properties:

1. $W(T)$ is finate if and only if T is bounded.
2. $W(T) = W(UT)$, for any orthogonal matrix $U \in \mathbb{R}^{n\times n}$.
3. $W(T+S) = W(T)+W(S)$, for any two sets $T, S \subset \mathbb{R}^n$, and $W(aT) = |a|W(T)$, for any $a \in \mathbb{R}$.
4. $W(T) = \frac{1}{2}W(T-T) = \frac{1}{2}\mathbb{E}\left[\sup_{t,t'\in T}\langle t-t', Y\rangle\right]$.
5. $(2\pi)^{-1/2}\operatorname{diam}(T) \leq W(T) \leq \sqrt{n}/2\operatorname{diam}(T)$.

Proof. 1. Cauchy-Schwarz inequality implies that

$$|\langle t, Y\rangle| \leq \|t\|_2\|Y\|_2, \quad \forall t \in T.$$

If $W(T) < \infty$, then $\|t\|_2\|_2 < \infty$, $\forall t \in T$, which implies that T is bounded. Conversely, if T is bounded, then we have $\|t\|_2 \leq M$, $\forall t \in T$, for some $M > 0$, which implies that

$$\mathbb{E}\left[\sup_{t\in T}\langle t, Y\rangle\right] \leq M\mathbb{E}\left[\sup_{t\in T}\|Y\|_2\right] \leq M\sqrt{n} < \infty.$$

2. Let U be an orthogonal matrix, then $UY \sim \mathcal{N}(0, \mathbf{I}_n)$.
3. For the addititivity property, we have

$$\mathbb{E}\left[\sup_{t\in T}\langle t, Y\rangle\right] + \mathbb{E}\left[\sup_{s\in S}\langle s, Y\rangle\right] = \mathbb{E}\left[\sup_{t\in T}\langle t, Y\rangle + \sup_{s\in S}\langle s, Y\rangle\right] = \mathbb{E}\left[\sup_{t\in T+s}\langle t, Y\rangle\right].$$

   For the scaling property, we have
4. Using the additivity property, we have

$$W(T) = \frac{1}{2}(W(T) + W(T)) = \frac{1}{2}(W(T) - W(-T)) = \frac{1}{2}W(T - T).$$

5. According to the property 4, we have $W(T) = \frac{1}{2}W(T-T)$, then for a fixed pair $t, t' \in T$, we have

$$W(T) \geq \frac{1}{2}\mathbb{E}\left[\max\{\langle t - t', Y\rangle, \langle t' - t, Y\rangle\}\right] = \frac{1}{2}\mathbb{E}\left[|\langle t - t', Y\rangle|\right] = \frac{1}{2}\sqrt{\frac{2}{\pi}}\|t - t'\|_2.$$

   As for the last equality, recall that $\langle t - t', Y\rangle \sim \mathcal{N}(0, \|t - t'\|_2^2)$, and therefore

$$\langle \frac{t - t'}{\|t - t'\|_2}, Y\rangle \sim \mathcal{N}(0, 1), \quad \mathbb{E}\left[|\langle \frac{t - t'}{\|t - t'\|_2}, Y\rangle|\right] = \sqrt{\frac{2}{\pi}}.$$

   Taking the supremum over all pairs $t, t' \in T$ gives the lower bound. The upper bound follows from the property 4 and the Cauchy-Schwarz inequality, i.e.,

$$\begin{aligned} W(T) &= \frac{1}{2}W(T - T) = \frac{1}{2}\mathbb{E}\left[\sup_{t,t'\in T}\langle t - t', Y\rangle\right] \\ &\leq \frac{1}{2}\mathbb{E}\left[\sup_{t,t'\in T}\|t - t'\|_2\|Y\|_2\right] \leq \frac{1}{2}\mathbb{E}\left[\|Y\|_2\right]\operatorname{diam}(T) = \frac{\sqrt{n}}{2}\operatorname{diam}(T). \end{aligned}$$

□

Example. We discuss the Gaussian width of some sets:

- **Unit Sphere**: The Gaussian width of the unit sphere $S^{n-1} = \{t \in \mathbb{R}^n : \|t\|_2 = 1\}$ is

$$W(S^{n-1}) = \mathbb{E}\left[\sup_{t\in S^{n-1}}\langle t, Y\rangle\right] = \mathbb{E}\left[\|Y\|_2\right] = \sqrt{n} \pm C$$

  for some constant $C > 0$.

- **Unit Cube w.r.t. ℓ_∞ norm**: The Gaussian width of the unit cube $C^n = [-1,1]^n$ with respect to the ℓ_∞ norm is

$$W(C^n) = \mathbb{E}\left[\|Y\|_1\right] = n\mathbb{E}\left[|Y_1|\right] = \sqrt{\frac{2}{\pi}}n.$$

- **Unit Ball w.r.t. ℓ_1 norm**: The Gaussian width of the unit ball $B^n = \{t \in \mathbb{R}^n : \|t\|_1 \leq 1\}$ with respect to the ℓ_1 norm is

$$W(B^n) = \mathbb{E}\left[\sup_{t\in B^n}\langle t, Y\rangle\right] = \mathbb{E}\left[\|Y\|_\infty\right],$$

such that,

$$c\sqrt{\log n} \leq W(B^n) \leq C\sqrt{\log n},$$

Definition 24.3.3 (Sub-Gaussian Increments)

Let $X = \{X(t)\}_{t\in T}$ be a random process on some metric space (T,d). We say that X has **sub-Gaussian increments** if there exists a constant $K > 0$ such that for all $t, s \in T$, we have

$$\|X(t) - X(s)\|_{\psi_2} \leq Kd(t,s).$$

Theorem 24.3.2 (Dudley's Inequality)

Let $X = \{X(t)\}_{t\in T}$ be a mean-zero random process on some metric space (T,d) with sub-Gaussian increments. Then

$$\mathbb{E}\left[\sup_{t\in T} X(t)\right] \leq CK\int_0^\infty \sqrt{\log \mathcal{N}(T,d,\varepsilon)}\,\mathrm{d}\varepsilon,$$

where $C > 0$ is a universal constant.

Theorem 24.3.3 (Discrete Dudley's Inequality)

Let $X = \{X(t)\}_{t\in T}$ be a mean-zero Gaussian process on a finite set $T = \{t_1, \ldots, t_n\}$ with sub-Gaussian increments. Then

$$\mathbb{E}\left[\sup_{t\in T} X(t)\right] \leq CK\sum_{k\in\mathbb{Z}} 2^{-k}\sqrt{\log \mathcal{N}(T,d,2^{-k})},$$

where $C > 0$ is a universal constant.

We discretize T by choosing an ε-net $\mathcal{N}$ of T, i.e., for every $t \in T$, there exists a point $\pi(t) \in \mathcal{N}$ such that $d(t,\pi(t)) \leq \varepsilon$. Then, the increment condition implies

that

$$\|X(t) - X(\pi(t))\|_{\psi_2} \leq Kd(t, \pi(t)) \leq K\varepsilon.$$

Proof. **Step 1: Chaining set-up.** Without loss of generality, we assume that $K = 1$ and T is finite. Define the dyadic scales

$$\varepsilon_k = 2^{-k}, \quad k \in \mathbb{Z},$$

and choose an ε_k-net $\mathcal{N}_k$ of T, such that

$$|T_k| = \mathcal{N}(T, d, \varepsilon_k).$$

Only part of the dyadic scales will be used in the chaining argument. Indeed, since T is finite, there exists a smallest integer $\kappa \in \mathbb{Z}$ such that $T_\kappa = \{t_0\}$, for some $t_0 \in T$, and a largest integer $K \in \mathbb{Z}$ such that $T_K = T$. Then, for a point $t \in T$, denote $\pi_k(t)$ the closest point in the ε_k-net $\mathcal{N}_k$ to t, i.e.,

$$d(t, \pi_k(t)) \leq \varepsilon_k.$$

Since $\mathbb{E}X(t_0) = 0$, we have

$$\mathbb{E}\left[\sup_{t\in T} X(t)\right] = \mathbb{E}\left[\sup_{t\in T} (X(t) - X(t_0))\right].$$

We can express $X(t) - X(t_0)$ as a telescopic sum, i.e., walk from t_0 to t through a chain of points $\pi_k(t)$, $k \in [\kappa, K]$, that mark progressively finer approximations of t:

$$X(t) - X(t_0) = \sum_{k=\kappa+1}^{K} \left(X(\pi_k(t)) - X(\pi_{k-1}(t))\right).$$

Since the supremum of the sum is bounded by the sum of the supremums, we have

$$\mathbb{E}\left[\sup_{t\in T} (X(t) - X(t_0))\right] \leq \sum_{k=\kappa+1}^{K} \mathbb{E}\left[\sup_{t\in T} (X(\pi_k(t)) - X(\pi_{k-1}(t)))\right]. \tag{24.6}$$

Step 2: Controling the increments. Although each term in the bound (24.6) still has a supremum over T, but it is actually a maximum over a much smaller set, namely the set all posible pairs $(\pi_k(t), \pi_{k-1}(t))$, $t \in T$, and the number of such pairs is

$$|T_k| \cdot |T_{k-1}| \leq |T_k|^2.$$

For fixed $t \in T$, the increments in (24.6) can be bounded as follows:

$$\|X(\pi_k(t)) - X(\pi_{k-1}(t))\|_{\psi_2} \leq d(\pi_k(t), \pi_{k-1}(t)) \leq$$
$$d(\pi_k(t), t) + d(t, \pi_{k-1}(t)) \leq \varepsilon_k + \varepsilon_{k-1} = 2\varepsilon_k.$$

Recall that the expectation of N sub-Gaussian random variables is at most $CL\sqrt{\log N}$, where L is the maximal ψ_2 norm. Thus, we have

$$\mathbb{E}\left[\sup_{t\in T}\left(X(\pi_k(t)) - X(\pi_{k-1}(t))\right)\right] \leq C\varepsilon_{k-1}\sqrt{\log|T_k|}.$$

Step 3: Summing up. Substituting the bound into (24.6), we obtain

$$\mathbb{E}\left[\sup_{t\in T} X(t)\right] \leq C_1 \sum_{k=\kappa+1}^{K} \varepsilon_k \sqrt{\log|T_k|}.$$

Step 4: Covert to integral. To convert the sum into an integral, we express $2^{-k} = 2\int_{2^{-k-1}}^{2^{-k}} \mathrm{d}\varepsilon$, and then we have

$$\sum_{k\in\mathbb{Z}} 2^{-k}\sqrt{\log \mathcal{N}(T,d,2^{-k})} = 2\sum_{k\in\mathbb{Z}} \int_{2^{-k-1}}^{2^{-k}} \sqrt{\log \mathcal{N}(T,d,\varepsilon)}\,\mathrm{d}\varepsilon.$$

Within the limits of the integral, we have $2^{-k} \geq \varepsilon$, so $\log \mathcal{N}(T,d,2^{-k}) \leq \log \mathcal{N}(T,d,\varepsilon)$, and the sum is bounded by

$$2\sum_{k\in\mathbb{Z}} \int_{2^{-k-1}}^{2^{-k}} \sqrt{\log \mathcal{N}(T,d,\varepsilon)}\,\mathrm{d}\varepsilon = 2\int_0^\infty \sqrt{\log \mathcal{N}(T,d,\varepsilon)}\,\mathrm{d}\varepsilon,$$

which completes the proof. □

24.4 Uniform Law of Large Numbers

Definition 24.4.1 (Empirical Process)

Let $(\Omega, \mathcal{B}(\Omega), \mu)$ be a probability space, $\mathcal{F} = \{f : \Omega \to \mathbb{R}\}$ be a class of real-valued functions, and X be a Ω-valued random variable with law $\mu \in \mathcal{M}_1(\Omega)$, with $X_1, X_2, \ldots, X_n$ be i.i.d. copies of X. The random process $X = (X_f)_{f\in\mathcal{F}}$ is defined as

$$X_f = \frac{1}{n}\sum_{i=1}^{n} f(X_i) - \mathbb{E}f(X),$$

is called the **empirical process** associated with the class $\mathcal{F}$.

Theorem 24.4.1 (Uniform Law of Large Numbers)

Denote

$$\mathcal{F} = \{f : [0,1] \to \mathbb{R} : \|f\|_{\text{Lip}} \leq L\},$$

the class of Lipschitz functions on $[0,1]$ with Lipschitz constant $L > 0$. Let $X_1, X_2, \ldots, X_n$ be i.i.d. distributed $[0,1]$-valued random variables. Then,

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n} f(X_i) - \mathbb{E}f(X)\right|\right] \leq \frac{CL}{\sqrt{n}},$$

where $C > 0$ is a universal constant.

A function defined on a compact set and with bounded Lipschitz constant is uniformly continuous, and therefore the function is bounded by a constant.

If not compact, the function may be unbounded, we need truncate it.

Proof. Without loss of generality, it suffices to consider the case

$$\mathcal{F} = \{f : [0,1] \to [0,1] : \|f\|_{\text{Lip}} \leq 1\}.$$

We wound like to bound the magnitude $\mathbb{E}\sup_{f\in\mathcal{F}}|X_f|$ of the empirical process $(X_f)_{f\in\mathcal{F}}$ defined in Definition 24.4.1.

Step 1: checking sub-gaussian increments.

Step 2: applying Dudley's inequality. □

Lemma 24.4.1

Let $\mathcal{F} = \{f : [0,1] \to [0,1] : \|f\|_{\text{Lip}} \leq 1\}$, then the Covering number of $\mathcal{F}$ with respect to the $\|\cdot\|_\infty$ norm is

$$\mathcal{N}(\mathcal{F}, \|\cdot\|_\infty, \varepsilon) \leq \left(\frac{2}{\varepsilon}\right)^{2/\varepsilon}, \quad \forall \varepsilon > 0.$$

Proof. □

24.5 VC Dimension

Definition 24.5.1 (VC Dimension)

Let $\mathcal{F}$ be a class of functions from a set T to $\{0,1\}$. The **VC dimension** of $\mathcal{F}$ is the cardinality of the largest set $A \subset T$ such that the restriction of $\mathcal{F}$ to A is the set of all possible functions from A to $\{0,1\}$.

Example (Intervals)**.**

Example (Half-planes)**.**

Remark (VC dimension of classes of sets)*.*

24.5.1 Pajor's Lemma

A lower bound is trivial:

$$|\mathcal{F}| \geq 2^{\mathrm{vc}(\mathcal{F})}.$$

Lemma 24.5.1 (Pajor's Lemma)

Let $\mathcal{F}$ be a class of Boolean functions on a finite set Ω, then

$$|\mathcal{F}| \leq |\{\Lambda \subset \Omega : \Lambda \text{ is shattered by } \mathcal{F}\}|.$$

We include the empty set in the definition of shattering.

Theorem 24.5.1 (Sauer-Shelah Lemma)

Let $\mathcal{F}$ be a class of Boolean functions on an n-element set Ω. Then

$$|\mathcal{F}| \leq \sum_{k=0}^{d} \binom{n}{k} \leq \left(\frac{en}{d}\right)^d,$$

where $d = \mathrm{vc}(\mathcal{F})$.

Proof. The proof follows by applying Pajor's Lemma to the class of all subsets of Ω that are shattered by $\mathcal{F}$. The cardinality of each such set Λ is bounded by $d = \mathrm{vc}(\mathcal{F})$, according to the definition of VC dimension. Thus

$$|\mathcal{F}| \leq |\{\Lambda \subset \Omega : |\Lambda| \leq d\}| \leq \sum_{k=0}^{d} \binom{n}{k},$$

since the sum in right-hand side gives the total number of subsets of Ω of size at most d, which proves the first inequality. The second inequality follows from the fact that $\binom{n}{k} \leq \left(\frac{en}{k}\right)^k$. □

24.5.2 Covering Numbers via VC Dimension

Let $\mathcal{F}$ be a class of boolean functions on a set Ω, and let μ be any probability measure on Ω. Then $\mathcal{F}$ can be viewed as a metric space under the $L^2(\mu)$ norm on $\mathcal{F}$ given by

$$d(f,g) = \|f-g\|_{L^2(\mu)} = \left(\int_\Omega (f-g)^2 \,\mathrm{d}\mu\right)^{1/2},$$

and the covering number of $\mathcal{F}$ with respect to this norm is denoted by $\mathcal{N}(\mathcal{F}, L^2(\mu), \varepsilon)$.

Theorem 24.5.2 (Covering Numbers via VC Dimension)

Let $\mathcal{F}$ be a class of Boolean functions on a probability space (Ω, Σ, μ). Then for any $\varepsilon \in (0, 1)$, we have

$$\mathcal{N}(\mathcal{F}, L^2(\mu), \varepsilon) \leq \left(\frac{2}{\varepsilon}\right)^d,$$

where $d = \mathrm{vc}(\mathcal{F})$.

Lemma 24.5.2 (Dimension Reduction)

Then such that the uniform probability measure μ_n on Ω_n satisfies

.

Proof. Our argument will be based on the probabilistic method. We choose the subset Ω_n at random and show that if satisfies the conclusion of the theorem with positve probability.

Let $X, X_1, X_2, \ldots, X_n$ independent be random points in Ω with law μ. Define the empirical probability measure μ_n on Ω by assigning each X_i, allowing multiplicities. We would like to show that with positive probability, the following event occurs:

$$\|f - g\|_{L^2(\mu_n)} := \frac{1}{n}\sum_{i=1}^{n}(f(X_i) - g(X_i))^2 > 0, \quad \forall f \neq g \in \mathcal{F},$$

which guarantees that the restriction of functions $f \in \mathcal{F}$ onto $\Omega_n := (X_1, \ldots, X_n)$ are all distinct.

Fix a pair of distinct functions $f, g \in \mathcal{F}$, and denote $h := (f-g)^2$ for simplicity. Then, we would like to bound the deviation

$$\|f - g\|^2_{L^2(\mu_n)} - \|f - g\|^2_{L^2(\mu)} = \frac{1}{n}\sum_{i=1}^{n} h(X_i) - \mathbb{E}h(X).$$

□

Proof of Theorem 24.5.2. □

24.5.3 Empirical Process via VC Dimension

Theorem 24.5.3 (Empirical Process via VC Dimension)

Let $\mathcal{F}$ be a class of Boolean functions on a probability space (Ω, Σ, μ), with finite VC dimension $\mathrm{vc}(\mathcal{F}) \geq 1$. Let $X, X_1, X_2, \ldots, X_n$ be independent random points in Ω distributed according to law μ. Then

$$\mathbb{E} \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} f(X_i) - \mathbb{E} f(X) \right| \leq C \sqrt{\frac{\mathrm{vc}(\mathcal{F})}{n}}.$$

Proof of Theorem 24.5.3. First, we use symmetrization to bound the left-hand side of the inequality by

$$\frac{2}{\sqrt{n}} \mathbb{E} \sup_{f \in \mathcal{F}} |Z_f|, \quad Z_f = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \varepsilon_i f(X_i),$$

Next, we condition on (X_i), leaving all the randomness in the signs ε_i. Then, to apply Dudley's inequality, we need to check the sub-Gaussian increments of the process $(Z_f)_{f \in \mathcal{F}}$. For any $f, g \in \mathcal{F}$, we have

$$\|Z_f - Z_g\|_{\psi_2} = \frac{1}{\sqrt{n}} \left\| \sum_{i=1}^{n} \varepsilon_i (f(X_i) - g(X_i)) \right\|_{\psi_2} \lesssim \left[\frac{1}{n} \sum_{i=1}^{n} (f-g)(X_i)^2 \right]^{1/2},$$

where we used the fact that the $\|\varepsilon_i\|_{\psi_2} \lesssim 1$.

TODO □

Let us examine an important application of Theorem 24.5.3, which is called Glivenko-Cantelli Theorem.

Theorem 24.5.4 (Glivenko-Cantelli Theorem)

Let $X, X_1, X_2, \ldots, X_n$ be i.i.d. random variables with common cumulative distribution F. Then

$$\mathbb{E} \|F_n - F\|_\infty = \mathbb{E} \sup_{x \in \mathbb{R}} |F_n(x) - F(x)| \leq \frac{C}{\sqrt{n}}.$$

24.6 Application: Statistical Learning Theory

An important class of learning problems are classification problems, where the function T is a Boolean function, i.e., $T : \Omega \to \{0, 1\}$, and thus T classifies the elements of Ω into two classes.

The goal is to learn a classifier $f : \Omega \to \mathbb{R}$, which we wolud like to choose f that minimizes the risk

$$R(f) = \mathbb{E}\left(f(X) - T(X)\right)^2,$$

where X denotes the random variable with distribution $\mathbb{P}$, i.e., with the same distribution as the sample points $X_1, X_2, \ldots, X_n \in \Omega$.

Ideally, we would like to find a function f^* from the hypothesis class $\mathcal{F}$ that minimizes the risk $R(f) = \mathbb{E}(f(X) - T(X))^2$, that is

$$f^* = \arg\min_{f \in \mathcal{F}} R(f).$$

However, the true distribution $\mathbb{P}$ is unknown, and we only have access to the sample points $X_1, X_2, \ldots, X_n$ drawn from $\mathbb{P}$. Therefore, we need to estimate the risk $R(f)$ based on the empirical risk

$$R_n(f) = \frac{1}{n}\sum_{i=1}^{n}(f(X_i) - T(X_i))^2,$$

Define the empirical risk minimizer in the hypothesis class $\mathcal{F}$ as f_n^*, i.e.,

$$f_n^* = \arg\min_{f \in \mathcal{F}} R_n(f).$$

The goal of statistical learning theory is to provide bounds on the excess risk

$$R(f_n^*) - R(f^*),$$

provided by our having to learn from a finite sample of size n.

Let us specialize to the classification problems where the target function T is a Boolean function. In this case, the risk $R(f)$ is the probability of misclassification, i.e.,

$$R(f) = \mathbb{P}(f(X) \neq T(X)).$$

Theorem 24.6.1 (Excess Risk via VC Dimension)

Assume that the target function T is a Boolean function, and let $\mathcal{F}$ be the hypothesis class of Boolean functions with finite VC dimension $\mathrm{vc}(\mathcal{F}) \geq 1$. Then

$$\mathbb{E}R(f_n^*) \leq R(f^*) + C\sqrt{\frac{\mathrm{vc}(\mathcal{F})}{n}},$$

Proof. □

Part IX

Random Matrix Theory

Chapter 25

Preliminary

25.1 Empirical Spectral Measure

Definition 25.1.1 (Empirical Spectral Measure)

For a symmetric matrix $\mathbf{M} \in \mathbb{R}^{n\times n}$, the spectral measure or empirical spectral measure or empirical spectral distribution (ESD) $\mu_{\mathbf{M}}$ of $\mathbf{M}$ is defined as the normalized counting measure of the eigenvalues $\lambda_1(\mathbf{M}), \ldots, \lambda_n(\mathbf{M})$ of $\mathbf{M}$, i.e.,

$$\mu_{\mathbf{M}} := \frac{1}{n}\sum_{i=1}^{n} \delta_{\lambda_i(\mathbf{M})} \tag{25.1}$$

where δ_x is a Dirac measure for any (measurable) set, that

$$\delta_x(A) := \mathbf{1}_A(x) = \begin{cases} 0, & x \notin A \\ 1, & x \in A \end{cases}$$

Since $\int \mu_{\mathbf{M}}(\mathrm{d}x) = 1$, the spectral measure $\mu_{\mathbf{M}}$ of a matrix $\mathbf{M} \in \mathbb{R}^{n\times n}$ (random or not) is a probability measure.

Remark. Many important statistics in multivariate analysis can be expressed as functionals of the ESD, such as, for $\mathbf{M}$ be an $n \times n$ positive definite matrix, then

$$\det(\mathbf{M}) = \prod_{i=1}^{n} \lambda_i = \exp\left(n \int_0^{\infty} \log x \mu_{\mathbf{M}}(\mathrm{d}x)\right) \tag{25.2}$$

25.2 Stieltjes Transform

Definition 25.2.1 (Resolvent)

For a symmetric matrix $\mathbf{M} \in \mathbb{R}^{n\times n}$, the resolvent $\mathbf{Q}_{\mathbf{M}}(z)$ of $\mathbf{M}$ is defined as

$$\mathbf{Q}_{\mathbf{M}}(z) := (\mathbf{M} - z\mathbf{I}_n)^{-1} \tag{25.3}$$

where $z \in \mathbb{C}$ not eigenvalue of $\mathbf{M}$.

Definition 25.2.2 (Stieltjes Transform)

For a real probability measure μ with support $\mathrm{supp}(\mu)$, the Stieltjes transform $m_\mu(z)$ is defined as

$$m_\mu(z) := \int \frac{1}{t-z}\mu\,(\mathrm{d}t) \tag{25.4}$$

where $z \in \mathbb{C}\backslash\,\mathrm{supp}(\mu)$.

Property. The Stieltjes transform m_μ has numerous interesting properties:

1. it is complex analytic on its domain of definition $\mathbb{C}\backslash\,\mathrm{supp}(\mu)$.
2. it is bounded $|m_\mu(z)| \leq 1/\,\mathrm{dist}(z, \mathrm{supp}(\mu))$.
3. it satisfies $\Im[z] > 0 \Rightarrow \Im[m(z)] > 0$.
4. it is an increasing function on all connected components of its restriction to $\mathbb{R}\backslash\,\mathrm{supp}(\mu)$.
5. if $\mathrm{supp}(\mu)$ is bounded, $\lim_{x\to\pm\infty} m_\mu(x) = 0$.

Remark. Most of the results involve Stieltjes transforms $m_\mu(z)$ of a real probability measure with support $\mathrm{supp}(\mu) \subset \mathbb{R}$. Since Stieltjes transforms are such that

$$m_\mu(z) > 0, \forall z < \inf \mathrm{supp}(\mu), \quad m_\mu(z) < 0, \forall z > \sup \mathrm{supp}(\mu), \quad \Im[z]\Im\left[m_\mu(z)\right] > 0, \text{ if } z \in \mathbb{C}\backslash\mathbb{R}$$

it will be convenient in the following to consider the set of scalar pairs

$$\mathcal{Z}(\mathcal{A}) = \{(z,m) \in \mathcal{A}\times\mathbb{C}, (\Im[z]\Im[m] > 0 \text{ if } \Im[z] \neq 0) \text{ or } (m > 0 \text{ if } z < \inf \mathcal{A}^c \cap \mathbb{R}) \\ \text{or } (m < 0 \text{ if } z > \sup \mathcal{A}^c \cap \mathbb{R})\}$$

As a transform, m_μ has an inverse formula to recover μ, as per the following result.

Theorem 25.2.1 (Inverse Stieltjes Transform)

For a, b continuity points of the probability measure μ, we have

$$\mu([a,b]) = \frac{1}{\pi}\lim_{y\downarrow 0}\int_a^b \Im\left[m_\mu(x+\imath y)\right]\mathrm{d}x \tag{25.5}$$

Specially, if μ has a density f at x, then

$$f(x) = \frac{1}{\pi}\lim_{y\downarrow 0}\Im\left[m_\mu(x+\imath y)\right] \tag{25.6}$$

And, if μ has an isolated mass at x, then

$$\mu(\{x\}) = \lim_{y\downarrow 0} -\imath y m_\mu(x+\imath y) \tag{25.7}$$

Proof.

$$\begin{aligned}\frac{1}{\pi}\int_a^b \Im\left[m_\mu(x+\imath y)\right]\mathrm{d}x =& \frac{1}{\pi}\int_a^b \left\{\int \Im\left[\frac{1}{(t-x)-\imath y}\right]\mu(\mathrm{d}t)\right\}\mathrm{d}x\\ =& \frac{1}{\pi}\int_a^b\left[\int \frac{y}{(t-x)^2+y^2}\mu(\mathrm{d}t)\right]\mathrm{d}x\end{aligned}$$

By Fubini theorem,

$$\begin{aligned}=& \frac{1}{\pi}\int\left[\int_a^b \frac{y}{(t-x)^2+y^2}\,\mathrm{d}x\right]\mu(\mathrm{d}t)\\ =& \frac{1}{\pi}\int\left[\arctan\left(\frac{b-t}{y}\right)-\arctan\left(\frac{a-t}{y}\right)\right]\mu(\mathrm{d}t)\end{aligned}$$

Since

$$\left|\frac{y}{(t-x)^2+y^2}\right| \le \frac{1}{y}, \quad \forall y>0$$

by the dominated convergence theorem,

$$\frac{1}{\pi}\lim_{y\downarrow 0}\int_a^b \Im\left[m_\mu(x+\imath y)\right]\mathrm{d}x = \frac{1}{\pi}\int\lim_{y\downarrow 0}\left[\arctan\left(\frac{b-t}{y}\right)-\arctan\left(\frac{a-t}{y}\right)\right]\mu(\mathrm{d}t)$$

as $y\downarrow 0$, the difference in brackets converges either to $\pm\pi$ or 0 depending on the relative position of a, b and t, thus

$$= \int 1_{[a,b]}\mu(\mathrm{d}t) = \mu([a,b])$$

Thus, if μ has a density f at x, then

$$f(x) = \frac{1}{\pi} \lim_{y \downarrow 0} \Im \left[m_\mu(x + \imath y) \right]$$

When μ has an isolated mass at x, i.e., $\mu(dt) = a\delta_x(t)$, similarly, since

$$|y(t-x)| \leq \frac{1}{2} \left(y^2 + (t-x)^2 \right)$$

by dominated convergence theorem,

$$\lim_{y \downarrow 0} -\imath y m_\mu(x + \imath y) = -\lim_{y \downarrow 0} \int \frac{\imath y (t-x) \mu(\mathrm{d}t)}{(t-x)^2 + y^2} + \lim_{y \downarrow 0} \int \frac{y^2 \mu(\mathrm{d}t)}{(t-x)^2 + y^2} = a$$

□

Remark. The important relation between the empirical spectral measure $\mu_{\mathbf{M}}$ of $\mathbf{M} \in \mathbb{R}^{n \times n}$, the Stieltjes transform $m_{\mu_{\mathbf{M}}}(z)$ and the resolvent $\mathbf{Q}_{\mathbf{M}}(z)$ lies in the fact that

$$m_{\mu_{\mathbf{M}}}(z) = \frac{1}{n} \sum_{i=1}^{n} \int \frac{\delta_{\lambda_i(\mathbf{M})}(t)}{t - z} = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{\lambda_i(\mathbf{M}) - z} = \frac{1}{n} \operatorname{tr} \mathbf{Q}_{\mathbf{M}}(z) \tag{25.8}$$

The resolvent $\mathbf{Q}_{\mathbf{M}}$ provides access to scalar observations of the eigenspectrum of $\mathbf{M}$ through its linear functionals. Cauchy's integral formula provides a connection between the linear functionals of the eigenvalues of $\mathbf{M}$ and the Stieltjes transform $m_{\mu_{\mathbf{M}}}(z)$ through

$$\frac{1}{n} \sum_{i=1}^{n} f\left(\lambda_i(\mathbf{M})\right) = -\frac{1}{2\pi\imath n} \oint_\Gamma f(z) \operatorname{tr}\left(\mathbf{Q}_{\mathbf{M}}(z)\right) \mathrm{d}z = -\frac{1}{2\pi\imath} \oint_\Gamma f(z) m_{\mu_{\mathbf{M}}}(z) \, \mathrm{d}z \tag{25.9}$$

for all f complex analytic in a compact neighborhood of $\operatorname{supp}(\mu_{\mathbf{M}})$, by choosing the contour Γ to enclose $\operatorname{supp}(\mu_{\mathbf{M}})$ (i.e., all the eigenvalues $\lambda_i(\mathbf{M})$).

25.3 Matrix Equivalents

Definition 25.3.1 (Deterministic Equivalent)

$\overline{\mathbf{Q}} \in \mathbb{R}^{n \times n}$ is said to be deterministic equivalent for the symmetric random matrix $\mathbf{Q} \in \mathbb{R}^{n \times n}$, if for a (sequences of) deterministic matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$ and vectors $\mathbf{a}, \mathbf{b} \in \mathbb{R}^n$ of unit norms (operator and Euclidean, respectively),

$$\frac{1}{n} \operatorname{tr} \mathbf{A}(\mathbf{Q} - \overline{\mathbf{Q}}) \to 0, \quad \mathbf{a}'(\mathbf{Q} - \overline{\mathbf{Q}})\mathbf{b} \to 0, \text{ as } n \to \infty \tag{25.10}$$

where the convergence is either in probability or almost sure.

Remark. A practical use of deterministic equivalents is to establish that, for a random matrix $\mathbf{M}$ of interest, suppose

$$\frac{1}{n}\operatorname{tr}\left(\mathbf{Q}_{\mathbf{M}}(z) - \overline{\mathbf{Q}}(z)\right) \to 0, \quad \text{a.s.}, \quad \forall z \in \mathcal{C}, \mathcal{C} \subset \mathbb{C}$$

this convergence implies that the Stieltjes transform of μ_{M} "converges" in the sense that

$$m_{\mu_{\mathrm{M}}}(z) - \bar{m}_n(z) \to 0$$

where $\bar{m}_n(z) = \frac{1}{n}\operatorname{tr}\overline{\mathbf{Q}}(z)$.

Definition 25.3.2 (Matrix Equivalents)

For $\mathbf{x}, \mathbf{Y} \in \mathbb{R}^{n\times n}$ two random or deterministic matrices, we write

$$\mathbf{x} \leftrightarrow \mathbf{Y} \tag{25.11}$$

if, for all $\mathbf{A} \in \mathbb{R}^{n\times n}$ and $\mathbf{a}, \mathbf{b} \in \mathbb{R}^n$ of unit norms (respectively, operator and Euclidean), we have the simultaneous results

$$\frac{1}{n}\operatorname{tr}\mathbf{A}(\mathbf{x}-\mathbf{Y}) \to 0, \quad \mathbf{a}'(\mathbf{x}-\mathbf{Y})\mathbf{b} \to 0, \quad \|\mathbb{E}[\mathbf{x}-\mathbf{Y}]\| \to 0$$

where, for random quantities, the convergence is either in probability or almost sure.

25.4 Resolvent and Perturbation Identities

Lemma 25.4.1 (Resolvent Identity)

For invertible matrices $\mathbf{A}$ and $\mathbf{B}$, we have

$$\mathbf{A}^{-1} - \mathbf{B}^{-1} = \mathbf{A}^{-1}(\mathbf{B}-\mathbf{A})\mathbf{B}^{-1} \tag{25.12}$$

Lemma 25.4.2 (Sherman-Morrison)

For $\mathbf{A} \in \mathbb{R}^{n\times n}$ invertible and $\mathbf{u}, \mathbf{v} \in \mathbb{R}^n$, then $\mathbf{A} + \mathbf{u}\mathbf{v}'$ is invertible if and only if $1 + \mathbf{v}'\mathbf{A}^{-1}\mathbf{u} \neq 0$ and

$$(\mathbf{A}+\mathbf{u}\mathbf{v}')^{-1} = \mathbf{A}^{-1} - \frac{\mathbf{A}^{-1}\mathbf{u}\mathbf{v}'\mathbf{A}^{-1}}{1+\mathbf{v}'\mathbf{A}^{-1}\mathbf{u}} \tag{25.13}$$

or,

$$(\mathbf{A}+\mathbf{u}\mathbf{v}')^{-1}\mathbf{u} = \frac{\mathbf{A}^{-1}\mathbf{u}}{1+\mathbf{v}'\mathbf{A}^{-1}\mathbf{u}} \tag{25.14}$$

Lemma 25.4.3 (Quadratic-form-close-to-the-trace)

Let $\mathbf{x} \in \mathbb{R}^p$ have iid entries of zero mean, unit variance and $\mathbb{E}\left[|x_i|^K\right] \leq \nu_K$ for some $K \geq 1$. Then for $\mathbf{A} \in \mathbb{R}^{p\times p}$ and $k \geq 1$

$$\mathbb{E}\left[\left|\mathbf{x}^\top \mathbf{A}\mathbf{x} - \operatorname{tr}\mathbf{A}\right|^k\right] \leq C_k\left[\left(\nu_4 \operatorname{tr}\left(\mathbf{A}\mathbf{A}'\right)\right)^{k/2} + \nu_{2k}\operatorname{tr}\left(\mathbf{A}\mathbf{A}'\right)^{k/2}\right]$$

for some constant $C_k > 0$ independent of p. In particular, if $\|\mathbf{A}\| \leq 1$ and the entries of x have bounded eighth-order moment,

$$\mathbb{E}\left[\left(\mathbf{x}^\top \mathbf{A}\mathbf{x} - \operatorname{tr}\mathbf{A}\right)^4\right] \leq Cp^2$$

for some $C > 0$ independent of p, and consequently, as $p \to \infty$,

$$\frac{1}{p}\mathbf{x}^\top \mathbf{A}\mathbf{x} - \frac{1}{p}\operatorname{tr}\mathbf{A} \xrightarrow{\text{a.s.}} 0$$

Chapter 26

Wigner Matrix

Chapter 27

Sample Covariance Matrix

Suppose $\{\mathbf{x}\}$ be a sequence of random vectors defined in $\mathbb{R}^n$, and $(x_i)_{1\leq i\leq n}$ be the components of the random vector $\mathbf{x}$, such that

$$E(\mathbf{x}) = 0, \quad E(\mathbf{x}\otimes\mathbf{x}) = \mathbf{I}_n$$

where $\mathbf{x}$ is also called **isotropic** random vector.

Suppose $\{m_n\}$ be a sequence defined in $\mathbb{N}$ such that

$$0 < \underline{\rho} := \liminf_{n\to\infty} \frac{n}{m_n} \leq \limsup_{n\to\infty} \frac{n}{m_n} =: \bar{\rho} < \infty$$

Let $\mathbf{x}_1, \ldots, \mathbf{x}_{m_n}$ be iid copies of $\mathbf{x}$, and $\mathbb{X}$ be the $m_n \times n$ random matrix with iid rows $\mathbf{x}_1, \ldots, \mathbf{x}_{m_n}$, and their empirical covariance matrix is

$$\widehat{\mathbf{\Sigma}} := \frac{1}{m_n}\sum_{i=1}^{m_n} \mathbf{x}_i\otimes\mathbf{x}_i = \frac{1}{m_n}\mathbb{X}'\mathbb{X}$$

which is a $n\times n$ symmetric positive semidefinite random matrix, and

$$E\left(\widehat{\mathbf{\Sigma}}\right) = \mathbb{E}(\mathbf{x}\otimes\mathbf{x}) = \mathbf{I}_n$$

For convenience, we define the random matrix

$$\mathbf{A} := m_n\widehat{\mathbf{\Sigma}} = \mathbb{X}'\mathbb{X} = \sum_{i=1}^{m_n} \mathbf{x}_i\otimes\mathbf{x}_i$$

27.1 Eigenvalues and Singular Values

Theorem 27.1.1

The eigenvalues of $\mathbf{A}$ are squares of the singular values of $\mathbb{X}$, in particularly

$$\lambda_{\max}(\mathbf{A}) = s_{\max}(\mathbb{X})^2 = \max_{\|\mathbf{x}\|=1} \|\mathbb{X}\mathbf{x}\|^2 = \|\mathbb{X}\|_2^2$$

if $m_n \geq n$, then

$$\lambda_{\min}(\mathbf{A}) = s_{\min}(\mathbb{X})^2 = \min_{\|\mathbf{x}\|=1} \|\mathbb{X}\mathbf{x}\|^2 = \left\|\mathbb{X}^{-1}\right\|_2^{-2}$$

Proof. □

27.2 Laguerre Orthogonal Ensemble

Definition 27.2.1 (Wishart Distribution)

Suppose $\mathbb{X}$ be a $p \times n$ matrix, each column of which is independently drawn from a p-variate normal distribution with zero means:

$$\mathbf{x}_i = \left(x_i^1, \ldots, x_i^p\right)' \sim N_p(0, \mathbf{\Sigma})$$

Then the Wishart distribution is the probability distribution of the $p \times p$ random matrix,

$$\mathbf{M} = \mathbb{X}'\mathbb{X} = \sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^\top \tag{27.1}$$

and which can be denoted by

$$\mathbf{M} \sim W_p(\mathbf{\Sigma}, n)$$

If $p = \mathbf{\Sigma} = 1$, then this distribution is a chi-squared distribution with n degrees of freedom.

Theorem 27.2.1

If $n \geq p$, the probability density function of $\mathbf{M}$ is

$$f(\mathbf{M}) = \frac{1}{2^{np/2}\left[\det(\mathbf{\Sigma})\right]^{n/2}\Gamma_p\left(\frac{n}{2}\right)} \det(\mathbf{M})^{(n-p-1)/2} \exp\left[-\frac{1}{2}\operatorname{tr}\left(\mathbf{\Sigma}^{-1}\mathbf{M}\right)\right] \tag{27.2}$$

concerning the Lebesque measure on the cone of symmetric positive definite matrices. Here, Γ_p is the multivariate gamma function defined as

$$\Gamma_p\left(\frac{n}{2}\right) = \pi^{p(p-1)/4}\prod_{j=1}^{p}\Gamma\left(\frac{n}{2} - \frac{j-1}{2}\right)$$

Remark. Especially, if the random variables $(x_i)_{1\leq i\leq n}$ are iid standard Gaussians, then the distribution of the random matrix $\widehat{\mathbf{\Sigma}}$ can be derived from the Wishart distribution. The probability density function of $\widehat{\mathbf{\Sigma}}$ can be derived from (27.2), since

$$\mathbf{A} \sim W_n(\mathbf{I}_n, m_n), \quad \det\left(\widehat{\mathbf{\Sigma}}\right) = m_n^{-n}\det(\mathbf{A}), \quad \operatorname{tr}\left(\widehat{\mathbf{\Sigma}}\right) = m_n^{-1}\operatorname{tr}(\mathbf{A})$$

thus,

$$f\left(\widehat{\mathbf{\Sigma}}\right) = \frac{m_n^{-n(m_n-n-1)/2+1}}{2^{m_n n/2}\Gamma_n\left(\frac{m_n}{2}\right)} \det\left(\widehat{\mathbf{\Sigma}}\right)^{(m_n-n-1)/2} \exp\left[-\frac{m_n}{2}\operatorname{tr}\left(\widehat{\mathbf{\Sigma}}\right)\right] \tag{27.3}$$

Theorem 27.2.2

If the random variables $(x_i)_{1\leq i\leq n}$ are iid standard Gaussians, the joint probability density function of eigenvalues of $\widehat{\mathbf{\Sigma}}$ is

$$p(\mathbf{\Lambda}) = \widetilde{Q}_{m_n,n}^{-1} \exp\left(-\frac{m_n}{2}\sum_{k=1}^{n}\lambda_k\right)\prod_{k=1}^{n}\lambda_k^{(m_n-n-1)/2}\prod_{i<j}|\lambda_i - \lambda_j| \tag{27.4}$$

where

$$0 \leq \lambda_1 \leq \cdots \leq \lambda_n < \infty$$

and $\widetilde{Q}_{m_n,n}$ is the normalization constant.

Proof. Fisrt, we will give the characteristic function of $\widehat{\mathbf{\Sigma}}$, i.e.,

$$\varphi_{\widehat{\mathbf{\Sigma}}}(\mathbf{P}) = E\left[\exp\left(\imath\sum_{1\leq i\leq j\leq n}P_{ij}\widehat{\mathbf{\Sigma}}_{ji}\right)\right] = E\left[\exp\left(\imath\operatorname{tr}\left(\mathbf{P}\widehat{\mathbf{\Sigma}}\right)\right)\right]$$

where $\{P_{ij}\}_{1\le i\le j\le n} \in \mathbb{R}^{(n+1)n/2}$ and $\mathbf{P}$ is a real symmetric matrix, that

$$\mathbf{P} = \left\{\widehat{P}_{ij}, \widehat{P}_{ij} = \widehat{P}_{ji}\right\}_{i,j=1}^{n}, \quad \widehat{P}_{ij} = \begin{cases} P_{ii}, & i = j \\ P_{ij}/2, & i < j \end{cases}$$

Thus, we have

$$\begin{aligned} &= \int_{\mathbb{R}^{m_n\times n}} \exp\left(\imath \operatorname{tr}\left(\mathbf{P}\widehat{\mathbf{\Sigma}}\right)\right) \cdot (2\pi)^{-m_n n/2} \exp\left(-\frac{1}{2}\sum_{k=1}^{m_n}\sum_{i=1}^{n}\left(x_i^{(k)}\right)^2\right) \prod_{k=1}^{m_n}\prod_{i=1}^{n} \mathrm{d}x_i^{(k)} \\ &= \int_{\mathbb{R}^{m_n\times n}} (2\pi)^{-m_n n/2} \exp\left(-\frac{1}{2}\sum_{k=1}^{m_n}\sum_{i=1}^{n}\sum_{j=1}^{n}\mathbf{Q}_{ij}x_i^{(k)}x_j^{(k)}\right) \prod_{k=1}^{m_n}\prod_{i=1}^{n} \mathrm{d}x_i^{(k)} \end{aligned}$$

where

$$\mathbf{Q} = \mathbf{I}_n - \frac{2\imath}{m_n}\mathbf{P}$$

Since $\left(x_i^{(k)}\right)_{1\le i\le n}$ are iid standard Gaussians,

$$\begin{aligned} &= \left[\int_{\mathbb{R}^n} (2\pi)^{-n/2} \exp\left(-\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\mathbf{Q}_{ij}x_i x_j\right) \prod_{i=1}^{n}\mathrm{d}x_i\right]^{m_n} \\ &= \left[\int_{\mathbb{R}^n} (2\pi)^{-n/2} \exp\left(-\frac{1}{2}\mathbf{x}^\top\mathbf{Q}\mathbf{x}\right)\mathrm{d}\mathbf{x}\right]^{m_n} \\ &= \left[\det(\mathbf{Q})^{-\frac{1}{2}}\int_{\mathbb{R}^n} (2\pi)^{-n/2} \exp\left(-\frac{1}{2}\left(\mathbf{Q}^{\frac{1}{2}}\mathbf{x}\right)'\left(\mathbf{Q}^{\frac{1}{2}}\mathbf{x}\right)\right)\mathrm{d}\mathbf{Q}^{\frac{1}{2}}\mathbf{x}\right]^{m_n} \\ &= [\det(\mathbf{Q})]^{-m_n/2} \end{aligned}$$

thus,

$$[\det(\mathbf{Q})]^{-m_n/2} = \left[\det\left(\mathbf{I}_n - \frac{2\imath}{m_n}\mathbf{P}\right)\right]^{-m_n/2} = \prod_{k=1}^{n}\left(1 - \frac{2\imath}{m_n}p_k\right)^{-m_n/2} \tag{27.5}$$

where $\{p_k\}_{k=1}^n$ are the eigenvalues of $\mathbf{P}$.

Then, we will show that the characteristic function of (27.4) coincides with the above function. By the Wishart distribution, the probability density of the real symmetric and positive definite random matrix $\widehat{\mathbf{\Sigma}}$ is

$$\widetilde{Q}_{m_n,n}^{-1} \exp\left[-\frac{m_n}{2}\operatorname{tr}\left(\widehat{\mathbf{\Sigma}}\right)\right]\left[\det\left(\widehat{\mathbf{\Sigma}}\right)\right]^{(m_n-n-1)/2}\mathrm{d}\widehat{\mathbf{\Sigma}} \tag{27.6}$$

where $\widetilde{Q}_{m_n,n}$ is the normalization constant. Then, the characteristic function of (27.6), i.e.,

$$\widetilde{Q}_{m_n,n}^{-1}\int_{\mathcal{S}_n^+} \exp\left[\imath\operatorname{tr}\left(\mathbf{P}\widehat{\mathbf{\Sigma}}\right) - \frac{m_n}{2}\operatorname{tr}\left(\widehat{\mathbf{\Sigma}}\right)\right]\left[\det\left(\widehat{\mathbf{\Sigma}}\right)\right]^{(m_n-n-1)/2}\mathrm{d}\widehat{\mathbf{\Sigma}}$$

where the integration is over the set $\mathcal{S}_n^+$ of $n \times n$ real symmetric and positive definite matrices. Since

$$\sum_{k=1}^{n} \lambda_k = \operatorname{tr}\left(\widehat{\mathbf{\Sigma}}\right), \quad \prod_{k=1}^{n} \lambda_k^{(m_n-n-1)/2} = \left[\det\left(\widehat{\mathbf{\Sigma}}\right)\right]^{(m_n-n-1)/2}$$

and

$$\mathrm{d}\widehat{\mathbf{\Sigma}} = \prod_{i<j} |\lambda_i - \lambda_j| \, \mathrm{d}\mathbf{\Lambda} H_1 \left(\mathrm{d}O\right)$$

where H_1 is the normalized Haar measure of $O(n)$, and the integration over $\mathbf{\Lambda}$ and $O \in O(n)$ are independent. Since the orthogonal invariance of the density of (27.6), and the characteristic function is

$$Q_{m_n,n}^{-1} \int_{(\mathbb{R}_+)^n} \exp\left[\sum_{k=1}^{n} \left(\imath p_k - \frac{m_n}{2}\right) \lambda_k\right] \prod_{k=1}^{n} \lambda_k^{(m_n-n-1)/2} \prod_{i<j} |\lambda_i - \lambda_j| \, \mathrm{d}\mathbf{\Lambda} \tag{27.7}$$

where $Q_{m_n,n} = m_n! \widetilde{Q}_{m_n,n}$.

If we viewed (27.5) and (27.7) as the function of $\{p_k\}_{k=1}^n \in \mathbb{R}^n$, then they can be **analytic continuation** to the domain

$$\{p_k + \imath p_k', p_k' \geq 0\}_{k=1}^n$$

If we replace $\{p_k\}_{k=1}^n$ by $\{\imath p_k', p_k' \geq 0\}_{k=1}^n$ on (27.5) since this is a set of the uniqueness of both (27.5) and eqrefeq:characteristic-function-wishart analytic functions, we have

$$Q_{m_n,n}^{-1} \int_{(\mathbb{R}_+)^n} \exp\left[-\frac{m_n}{2} \sum_{k=1}^{n} q_k \lambda_k\right] \prod_{k=1}^{n} \lambda_k^{(m_n-n-1)/2} \prod_{i<j} |\lambda_i - \lambda_j| \, \mathrm{d}\mathbf{\Lambda}$$

where $q_k = 1 + \frac{2p_k'}{m_n} \geq 1, k = 1, \ldots, n$, and since

$$\forall i, j \quad \frac{q_i}{q_j} = \frac{1 + \frac{2p_i'}{m_n}}{1 + \frac{2p_j'}{m_n}} \to 1, \text{ as } \quad m_n \to \infty$$

we have

$$\prod_{i<j} |q_i\lambda_i - q_j\lambda_j| = \prod_{i<j} q_i \left|\lambda_i - \frac{q_j}{q_i}\lambda_j\right| \to \prod_{k=1}^{n} q_k^{(n-1)/2} \prod_{i<j} |\lambda_i - \lambda_j|, \text{ as } \quad m_n \to \infty$$

thus,

$$\prod_{k=1}^n q_k^{-m_n/2} \cdot Q_{m_n,n}^{-1} \int_{(\mathbb{R}_+)^n} \exp\left[-\frac{m_n}{2} \sum_{k=1}^n q_k\lambda_k\right] \prod_{k=1}^n (q_k\lambda_k)^{(m_n-n-1)/2} \cdot \prod_{i<j} |q_i\lambda_i - q_j\lambda_j| \, \mathrm{d}\mathbf{q}\mathbf{\Lambda}$$

Since

$$\forall k \quad q_k \lambda_k \to \lambda_k, \text{ as } \quad m_n \to \infty$$

we can "lifting" from $\{\lambda_k\}_{k=1}^n$ to $\mathcal{S}_n^+$ bring the integral to

$$\prod_{k=1}^n \left(1+\frac{2p_k'}{m_n}\right)^{-m_n/2} \widetilde{Q}_n^{-1} \int_{\mathcal{S}_n^+} \exp\left[-\frac{m_n}{2}\operatorname{tr}\left(\widehat{\mathbf{\Sigma}}\right)\right]\left[\det\left(\widehat{\mathbf{\Sigma}}\right)\right]^{(m_n-n-1)/2} \mathrm{d}\widehat{\mathbf{\Sigma}}$$

The integral here is equal to $\widetilde{Q}_n$, the normalization constant of the probability measure (27.6). If we replace $\{\imath p_k'\}_{k=1}^n$ back by $\{p_k\}_{k=1}^n$, then the above expression is

$$\prod_{k=1}^n \left(1-\frac{2\imath}{m_n}p_k\right)^{-m_n/2}$$

which coincides with (27.5). Thus the probability law of the Wishart matrices of $\mathbf{\Sigma}$ given by (27.6) implies that the corresponding joint probability density of eigenvalues is given by (27.4) for $\mathbf{\Sigma}$. □

Definition 27.2.2 (Laguerre Orthogonal Ensemble)

For the $n \times n$ Laguerre orthogonal ensembles of statistics, the joint probability density function of eigenvalues is for arbitrary parameter $\beta > 0$ and $\alpha > -\frac{2}{\beta}$, is

$$p\left(\mathbf{\Lambda}\right) = K_{\alpha,\beta} \exp\left(-\frac{\beta}{2}\sum_{k=1}^n \lambda_k\right) \prod_{k=1}^n \lambda_k^{\frac{\alpha\beta}{2}} \prod_{i<j} |\lambda_i - \lambda_j|^\beta \tag{27.8}$$

where

$$0 \le \lambda_1 \le \cdots \le \lambda_n < \infty$$

and $K_{n,m}$ are normalization constant.

And Equation (27.8) can be written in the standard Boltzmann-Gibbs form, that,

$$p\left(\mathbf{\Lambda}\right) \propto \exp\left[-\beta E\left(\mathbf{\Lambda}\right)\right]$$

where

$$E\left(\mathbf{\Lambda}\right) = \frac{1}{2}\sum_{k=1}^n \left(\lambda_k - \alpha \log \lambda_k\right) - \frac{1}{2}\sum_{i\neq j} |\lambda_i - \lambda_j| \tag{27.9}$$

Remark. For the (27.4), which can be written as (27.8) form, that,

$$p\left(\mathbf{\Lambda}\right) \propto \exp\left[-\beta m_n E\left(\mathbf{\Lambda}\right)\right]$$

where $\beta = 1$ and

$$E\left(\mathbf{\Lambda}\right) = \frac{m_n}{2}\sum_{k=1}^n \left[\lambda_k - \left(\frac{m_n - n - 1}{m_n}\right)\log \lambda_k\right] - \frac{1}{2m_n}\sum_{i\neq j} |\lambda_i - \lambda_j|$$

27.3 Marčenko-Pastur Theorem

In this section, we will investigate the empirical spectral measure of $\widehat{\mathbf{\Sigma}}$, which converges to a nonrandom distribution — Marčenko-Pastur distribution. Before further proof, we will introduce some basic concepts and tools.

With the above tools, we can prove the Marčenko-Pastur Theorem. Here, we only suppose $\mathbf{x}$ has some smooth tail condition.

Theorem 27.3.1 (Marčenko-Pastur Theorem)

Consider the resolvent

$$\mathbf{Q}(z) = \left(\widehat{\mathbf{\Sigma}} - z\mathbf{I}_n\right)^{-1}$$

Then, if

$$\frac{n}{m_n} \to \rho \text{ with } \rho \in (0, \infty), \text{ as } n \to \infty$$

we have

$$\mathbf{Q}(z) \leftrightarrow \overline{\mathbf{Q}}(z), \quad \overline{\mathbf{Q}}(z) = m(z)\mathbf{I}_n$$

with $(z, m(z))$ the unique solution in $\mathcal{Z}\left(\mathbb{C}\backslash\left[(1-\sqrt{\rho})^2, (1+\sqrt{\rho})^2\right]\right)$ be

$$zcm^2(z) - (1 - c - z)m(z) + 1 = 0$$

where the function $m(z)$ is the Stieltjes transform of the probability measure μ given explicitly by

$$\mu(\mathrm{d}x) = \left(1 - \rho^{-1}\right)^+ \delta_0(x) + \frac{\sqrt{(x-a_-)^+(a_+-x)^+}}{2\pi\rho x}\,\mathrm{d}x$$

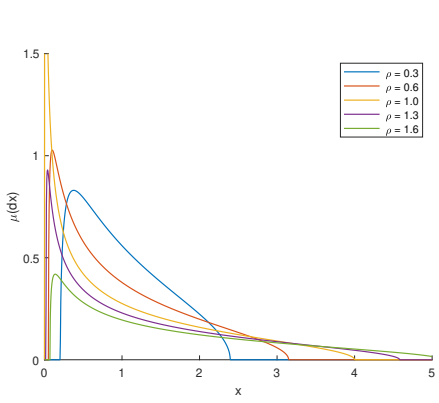
where $a_\pm = (1 \pm \sqrt{\rho})^2$ and $(x)^+ = \max(0, x)$, and is known as the Marčenko-Pastur distribution. In particular, with probability one, the empirical spectral measure $\mu_{\widehat{\mathbf{\Sigma}}}$ converges weakly to μ.

Proof. (Intuitive Proof) Suppose $\overline{\mathbf{Q}}(z) = \mathbf{F}(z)^{-1}$ for some matrix $\mathbf{F}(z)$. To prove $\overline{\mathbf{Q}}(z)$ to be a deterministic equivalent for $\mathbf{Q}(z)$, particularly,

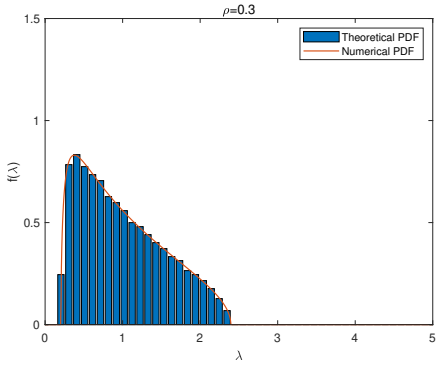
$$\frac{1}{n}\operatorname{tr}\mathbf{A}(\mathbf{Q}(z) - \overline{\mathbf{Q}}(z)) \to 0 \quad \text{a.s.}$$

where $\mathbf{A}$ is arbitrary, deterministic, and such that $\|\mathbf{A}\| = 1$. By Lemma 25.4.1, we have

$$\begin{aligned}\mathbf{Q}(z) - \overline{\mathbf{Q}}(z) =& \mathbf{Q}(z)\left(\mathbf{F}(z) + z\mathbf{I}_n - \widehat{\mathbf{\Sigma}}\right)\overline{\mathbf{Q}}(z) \\ =& \mathbf{Q}(z)\left(\mathbf{F}(z) + z\mathbf{I}_n - \frac{1}{m_n}\sum_{i=1}^{m_n}\mathbf{x}_i\mathbf{x}_i^\top\right)\overline{\mathbf{Q}}(z)\end{aligned}$$



(a) The Marčenko-Pastur Distribution for $\rho = 0.3, 0.6, 1, 1.3, 1.6$



(b) Simulation Results of the Marčenko-Pastur Theorem When $\rho = 0.3$

Figure 27.1: Illustrations of the Marčenko-Pastur Theorem

Thus, we turn to prove that,

$$\frac{1}{n}\operatorname{tr}\left[\left(\mathbf{F}(z)+z\mathbf{I}_n\right)\overline{\mathbf{Q}}(z)\mathbf{A}\mathbf{Q}(z)\right]-\frac{1}{n}\cdot\frac{1}{m_n}\sum_{i=1}^{m_n}\mathbf{x}_i^\top\overline{\mathbf{Q}}(z)\mathbf{A}\mathbf{Q}(z)\mathbf{x}_i\to 0 \quad \text{a.s.}$$

By Lemma 25.4.2, we have

$$\mathbf{Q}(z)\mathbf{x}_i=\frac{\mathbf{Q}_{-i}(z)\mathbf{x}_i}{1+\frac{1}{m_n}\mathbf{x}_i^\top\mathbf{Q}_{-i}(z)\mathbf{x}_i}$$

where

$$\mathbf{Q}_{-i}(z)=\left(\frac{1}{m_n}\sum_{j\neq i}\mathbf{x}_j\mathbf{x}_j'-z\mathbf{I}_n\right)^{-1}$$

is independent of $\mathbf{x}_i$. By Lemma 25.4.3, we have

$$\frac{1}{n}\mathbf{x}_i^\top\overline{\mathbf{Q}}(z)\mathbf{A}\mathbf{Q}(z)\mathbf{x}_i=\frac{\frac{1}{n}\mathbf{x}_i^\top\overline{\mathbf{Q}}(z)\mathbf{A}\mathbf{Q}_{-i}(z)\mathbf{x}_i}{1+\frac{1}{m_n}\mathbf{x}_i^\top\mathbf{Q}_{-i}(z)\mathbf{x}_i}\simeq\frac{\frac{1}{n}\operatorname{tr}\left[\overline{\mathbf{Q}}(z)\mathbf{A}\mathbf{Q}_{-i}(z)\right]}{1+\frac{1}{m_n}\operatorname{tr}\left[\mathbf{Q}_{-i}(z)\right]}$$

Hence, we need to prove the approximation that

$$\frac{1}{n}\operatorname{tr}\left[\left(\mathbf{F}(z)+z\mathbf{I}_n\right)\overline{\mathbf{Q}}(z)\mathbf{A}\mathbf{Q}(z)\right]\simeq\frac{\frac{1}{n}\operatorname{tr}\left[\overline{\mathbf{Q}}(z)\mathbf{A}\mathbf{Q}(z)\right]}{1+\frac{1}{m_n}\operatorname{tr}\left[\mathbf{Q}(z)\right]}$$

If $\mathbf{F}(z)$ exist, for the approximation above to hold, $\mathbf{F}(z)$ must be of the type

$$\mathbf{F}(z)\simeq\left(-z+\frac{1}{1+\frac{1}{m_n}\operatorname{tr}\mathbf{Q}(z)}\right)\mathbf{I}_n$$

By Equation 25.8, we have,

$$m(z) \equiv \frac{1}{n}\operatorname{tr}\left[\overline{\mathbf{Q}}(z)\right] = \frac{1}{n}\operatorname{tr}\left[\mathbf{F}(z)^{-1}\right]$$

taking $\mathbf{A} = \mathbf{I}_n$, we have

$$\frac{1}{n}\operatorname{tr}\left[\mathbf{Q}(z)\right] \simeq \frac{1}{n}\operatorname{tr}\left[\overline{\mathbf{Q}}(z)\right] = m(z) = \frac{1}{-z + \frac{1}{1+\frac{n}{m_n}\frac{1}{n}\operatorname{tr}[\mathbf{Q}(z)]}} \simeq \frac{1}{-z + \frac{1}{1+\rho m(z)}}$$

As $n, m_n \to \infty$, $m(z)$ is solution to

$$m(z) = \frac{1}{-z + \frac{1}{1+\rho m(z)}}$$

or equivalently

$$z\rho m^2(z) - (1-\rho-z)m(z) + 1 = 0$$

This equation has two solutions defined via the two values of the complex square root function. Let

$$z = r\mathrm{e}^{\imath\theta} \text{ where } r \geq 0, \theta \in [0, 2\pi) \Rightarrow \sqrt{z} \in \left\{\pm\sqrt{r}\mathrm{e}^{\imath\theta/2}\right\}$$

and we can conclude that

$$m(z) = \frac{1-\rho-z}{2\rho z} + \frac{\sqrt{\left((1+\sqrt{\rho})^2 - z\right)\left((1-\sqrt{\rho})^2 - z\right)}}{2\rho z}$$

only one of which is such that $\Im[z]\Im[m(z)] > 0$ as imposed by the definition of Stieltjes transforms. By the inverse Stieltjes transform theorem, Theorem 25.2.1, we find that $m(z)$ is the Stieltjes transform of the measure μ with

$$\mu([a,b]) = \frac{1}{\pi}\lim_{\epsilon\downarrow 0}\int_a^b \Im[m(x+\imath\epsilon)]\,\mathrm{d}x$$

for all continuity points $a, b \in \mathbb{R}$ of μ. This term under the square root in $m(z)$ is negative only in the set

$$\left[(1-\sqrt{\rho})^2, (1+\sqrt{\rho})^2\right]$$

(and thus of non-real square root), the latter defines the support of the continuous part of the measure μ with density

$$\frac{\sqrt{\left((1+\sqrt{\rho})^2 - x\right)\left(x - (1-\sqrt{\rho})^2\right)}}{2\rho\pi x}$$

at point x in the set. The case $x = 0$ brings a discontinuity in μ with weight equal to

$$\mu(\{0\}) = -\lim_{y \downarrow 0} \imath y m(\imath y) = \frac{\rho - 1}{2\rho} \pm \frac{\rho - 1}{2\rho}$$

where the sign is established by a second order development of $zm(z)$ in the neighborhood of zero: that is, "+" for $c > 1$ inducing a mass $1 - 1/\rho$ for $p > n$, or "-" for $c < 1$ in which case $\mu(\{0\}) = 0$ and μ has no mass at zero. □

Remark. The asymptotic phenomenon holds not only in the Gaussian case, which also holds

1. if $(x_i)_{1 \le i \le n}$ are iid with finite second moment.
2. if $\mathbf{x}$ is isotropic and log-concave[1] random vector.

27.4 Limits of Extreme Eigenvalues

The weak convergence in Theorem 27.3.1 does not provide much information at the edge on the behavior of the extremal atoms, and what one can extract is that

$$\limsup_{n \to \infty} \lambda_{\min}\left(\widehat{\mathbf{\Sigma}}\right) \le (1 - \sqrt{\rho})^2 \le (1 + \sqrt{\rho})^2 \le \liminf_{n \to \infty} \lambda_{\max}\left(\widehat{\mathbf{\Sigma}}\right) \quad \text{a.s.} \tag{27.10}$$

where the first inequality is considered only in the case where $m_n \ge n$.

The weak convergence above does not provide much information at the edge on the behavior of the extremal atoms. Now, we have more exact result, that if $(X_{n,k})_{n \ge 1, 1 \le k \le n}$ are iid with finite fourth moment then,

$$(1 - \sqrt{\rho})^2 = \lim_{n \to \infty} \lambda_{\min}\left(\widehat{\mathbf{\Sigma}}\right) \le \lim_{n \to \infty} \lambda_{\max}\left(\widehat{\mathbf{\Sigma}}\right) = (1 + \sqrt{\rho})^2 \quad \text{a.s.} \tag{27.11}$$

where the first inequality is considered only in the case where $m_n \ge n$.

Remark. The convergence of the smallest eigenvalue in the left-hand side of (27.11) holds if $(x_i)_{1 \le i \le n}$ are iid with finite second moment.

[1] A probability measure μ on $\mathbb{R}^n$ with density φ is log-concave when $\varphi = e^{-V}$ with V convex.

Theorem 27.4.1

If $\bar{\rho} < 1$ (in particular $m_n > n$ for $n \gg 1$) and if the centered isotropic random vector $\mathbf{x}$ is log-concave or if $(x_i)_{1 \le i \le n}$ are iid then

$$\liminf_{n \to \infty} \frac{E\left(\lambda_{\min}\left(\mathbf{A}_n\right)\right)}{\left(\sqrt{m_n} - \sqrt{n}\right)^2} \ge 1 \tag{27.12}$$

If additionally $\lim_{n \to \infty} \frac{n}{m_n} = \rho$ with $\rho \in (0,1)$, in other words $\underline{\rho} = \bar{\rho} \in (0,1)$, then

$$\lambda_{\min}\left(\widehat{\mathbf{\Sigma}}_n\right) \xrightarrow{p} (1 - \sqrt{\rho})^2 \text{ as } n \to \infty \tag{27.13}$$

Proof. □

Theorem 27.4.2

If the centered isotropic random vector $\mathbf{x}$ is log-concave or if $(x_i)_{1 \le i \le n}$ are iid with finite 4-th moment then

$$\limsup_{n \to \infty} \frac{E\left(\lambda_{\max}\left(\mathbf{A}_n\right)\right)}{\left(\sqrt{m_n} + \sqrt{n}\right)^2} \le 1 \tag{27.14}$$

If additionally $\lim_{n \to \infty} \frac{n}{m_n} = \rho$ with $\rho \in (0,1)$, in other words $\underline{\rho} = \bar{\rho} \in (0,1)$, then

$$\lambda_{\max}\left(\widehat{\mathbf{\Sigma}}_n\right) \xrightarrow{p} (1 + \sqrt{\rho})^2 \text{ as } n \to \infty \tag{27.15}$$

Proof. □

Part X

Statistics Inference

Chapter 28

Statistical Theory

28.1 Populations and Samples

28.2 Statistics

28.2.1 Sufficient Statistics

Definition 28.2.1 (Sufficient Statistics)

A statistic T is said to be sufficient for X, or for the family $\mathcal{P} = \{P_\theta, \theta \in \Omega\}$ of possible distributions of X, or for θ, if the conditional distribution of X given $T = t$ is independent of θ for all t.

Theorem 28.2.1 (Fisher-Neyman Factorization Theorem)

If the probability density function is $p_\theta(x)$, then T is sufficient for θ if and only if nonnegative functions g and h can be found such that

$$p_\theta(x) = h(x)g_\theta[T(x)].$$

Proof. □

28.2.2 Complete Statistics

Definition 28.2.2 (Complete Statistics)

A statistic T is said to be complete, if $Eg(T) = 0$ for all θ and some function g implies that $P(g(T) = 0 \mid \theta) = 1$ for all θ.

28.3 Estimators

Definition 28.3.1 (Estimator)

An estimator is a real-valued function defined over the sample space, that is

$$\delta : \mathbf{X} \to \mathbb{R}. \tag{28.1}$$

It is used to estimate an estimand, θ, a real-valued function of the parameter.

Unbiasedness

Definition 28.3.2 (Unbiasedness)

An estimator $\hat{\theta}$ of θ is unbiased if

$$E\hat{\theta} = \theta, \quad \forall \theta \in \Theta. \tag{28.2}$$

Remark.
- Unbiased estimators of θ may not exist.
-

Example (Nonexistence of Unbiased Estimator)**.**

Consistency

Definition 28.3.3 (Consistency)

An estimator $\hat{\theta}_n$ of θ is consistent if

$$\lim_{n\to\infty} P\left(\left|\hat{\theta}_n - \theta\right| > \varepsilon\right) = 0, \quad \forall \varepsilon > 0, \tag{28.3}$$

that is,

$$\hat{\theta}_n \xrightarrow{p} \theta. \tag{28.4}$$

Example (Consistency of Sample Moments)**.**

Remark.
1. Unbiased But Consistent
2. Biased But Not Consistent

Asymptotic Normality

Definition 28.3.4 (Asymptotic Normality)

An estimator $\hat{\theta}_n$ of θ is asymptotic normality if

$$\sqrt{n}\left(\hat{\theta} - \theta\right) \xrightarrow{d} N\left(0, \sigma_\theta^2\right). \tag{28.5}$$

Efficiency

Definition 28.3.5 (Efficiency)

Robustness

Definition 28.3.6 (Robustness)

Chapter 29

Point Estimation

29.1 Maximum Likelihood Estimator

Suppose that $\mathbf{x}_n = (x_1, \ldots, x_n)$, within a parametric family

$$p\left(X; \theta_0\right) \in \mathcal{P} = \{p(X; \boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta\}$$

The maximum likelihood estimate for observed $\mathbf{x}_n$ is the value $\boldsymbol{\theta} \in \Theta$ which maximizes $L_n\left(\boldsymbol{\theta}\right) := p\left(\mathbf{x}_n; \boldsymbol{\theta}\right)$, i.e.,

$$\hat{\boldsymbol{\theta}} = \max_{\boldsymbol{\theta} \in \Theta} L_n\left(\boldsymbol{\theta}\right). \tag{29.1}$$

In practice, it is often convenient to work with the natural logarithm of the likelihood function, called the log-likelihood:

$$\ell_n\left(\boldsymbol{\theta}\right) := \log L_n\left(\boldsymbol{\theta}\right)$$

Since the logarithm is a monotonic function, the maximum of $\ell_n\left(\boldsymbol{\theta}\right)$ occurs at the same value of $\boldsymbol{\theta}$ as does the maximum of $L_n\left(\boldsymbol{\theta}\right)$

29.1.1 Consistency

To establish consistency, the following conditions are sufficient:

(C1) Identification: $\boldsymbol{\theta}_0$ is identified in the sense that if $\boldsymbol{\theta} \neq \boldsymbol{\theta}_0$ and $\boldsymbol{\theta} \in \Theta$, then $p(X; \boldsymbol{\theta}) \neq p\left(X; \boldsymbol{\theta}_0\right)$ with respect to the dominating measure μ.

(C2) The parameter space Θ of the model is compact.

(C3) The function $\log p(X; \boldsymbol{\theta})$ is continuous in $\boldsymbol{\theta}$ for almost all values of x, i.e.,

$$P\left[\log p(X; \boldsymbol{\theta}) \in C^0(\Theta)\right] = 1 \tag{29.2}$$

(C4) Dominance: there exists $D(x)$ integrable with respect to the distribution $p(X;\boldsymbol{\theta}_0)$ such that $|\log p(X;\boldsymbol{\theta})| < D(x)$ for all $\boldsymbol{\theta} \in \Theta$.

Lemma 29.1.1

If $\boldsymbol{\theta}_0$ is identified and $E_{\boldsymbol{\theta}_0}[|\ln p(X;\boldsymbol{\theta})|] < \infty, \forall \boldsymbol{\theta} \in \Theta$, then $\ell(\boldsymbol{\theta})$ is uniquely maximized at $\boldsymbol{\theta} = \boldsymbol{\theta}_0$.

Proof. By the strict version of Jensen's inequality, with $\theta \neq \theta_0$,

$$\begin{aligned}\ell(\theta_0) - \ell(\theta) =& \mathbb{E}_{\theta_0}\left\{-\ln\left[\frac{p(z \mid \theta)}{p(z \mid \theta_0)}\right]\right\} > -\ln \mathbb{E}_{\theta_0}\left[\frac{p(z \mid \theta)}{p(z \mid \theta_0)}\right] \\ =& -\ln\left[\int f(z \mid \theta)\,\mathrm{d}z\right] = 0\end{aligned}$$

□

Theorem 29.1.1 (Consistency of MLE)

Under the Assumption (1)- (4), we have

$$\hat{\theta} \xrightarrow{p} \theta_0 \tag{29.3}$$

Proof. Suppose

$$\Theta(\epsilon) = \{\theta : \|\theta - \theta_0\| < \epsilon\}, \quad \forall \varepsilon > 0$$

Since $Q_0(\theta)$ is a continuous function, thus

$$\theta^* := \sup_{\theta \in \Theta \cap \Theta(\epsilon)^C} \{\ell(\theta)\}$$

is a achieved for a θ in the compact set $\theta \in \Theta \cap \Theta(\epsilon)^C$ (For open set $\Theta(\epsilon)$, $\Theta \cap \Theta(\epsilon)^C$ is a compact set). And θ_0 is the unique maximized,

$$\exists \delta > 0, \quad \ell(\theta_0) - \ell(\theta^*) = \delta$$

1. For $\theta \in \Theta \cap \Theta(\epsilon)^C$. Suppose

$$A_n = \left\{\sup_{\theta \in \Theta \cap \Theta(\epsilon)^C} \left|\hat{\ell}(\theta; \mathbf{X}_n) - \ell(\theta)\right| < \frac{\delta}{2}\right\}$$

then,

$$A_n \Longrightarrow \hat{\ell}(\theta; \mathbf{X}_n) < \ell(\theta) + \frac{\delta}{2} \leq \ell(\theta^*) + \frac{\delta}{2} = \ell(\theta_0) - \frac{\delta}{2}$$

2. For $\theta \in \Theta(\epsilon)$, suppose

$$B_n = \left\{ \sup_{\theta \in \Theta(\epsilon)} \left| \hat{\ell}\left(\theta\right) - \ell(\theta) \right| < \frac{\delta}{2} \right\}$$

then

$$B_n \implies \forall \theta \in \Theta(\epsilon),\ \hat{\ell}\left(\theta\right) > \ell(\theta) - \frac{\delta}{2}$$

By the uniform law of large numbers, the dominance condition together with continuity establishes the uniform convergence in the probability of the log-likelihood:

$$\sup_{\theta \in \Theta} |\hat{\ell}(\theta) - \ell(\theta)| \xrightarrow{p} 0$$

Thus, we can conclude that

$$P\left(A_n \cap B_n\right) \to 1$$

Within the definition

$$\hat{\theta} = \max_{\theta \in \Theta} \hat{\ell}\left(\theta\right)$$

we have,

$$A_n \cap B_n \implies \hat{\theta} \in \Theta(\epsilon)$$

Hence,

$$\forall \varepsilon > 0,\ P\left[\hat{\theta} \in \Theta(\epsilon)\right] \to 1 \implies \hat{\theta} \xrightarrow{p} \theta_0$$

□

29.1.2 Fisher Information

Definition 29.1.1 (Fisher Information)

The Fisher information of a random variable X with probability density function $p(X;\theta)$ is defined as

$$I(\theta) = \mathbb{E}\left[\left(\frac{\partial}{\partial \theta} \ln p(X;\theta)\right)^2\right]. \tag{29.4}$$

Alternatively, the Fisher information can be expressed as

$$I(\theta) = -\mathbb{E}\left[\frac{\partial^2}{\partial \theta^2} \ln p(X;\theta)\right] \tag{29.5}$$

29.1.3 Asymptotic Normality

(C5) The information matrix $I(\boldsymbol{\theta})$ is positive definite.
(C6) $\left\|\frac{\partial^2 \log p(X;\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top}\right\| \leq M(x)$ for all $\boldsymbol{\theta} \in \Theta$ and $\mathbb{E}_{\boldsymbol{\theta}_0} M(x) < \infty$.

Proof. Since the MLE is the maximizer of the log-likelihood function, the score function evaluated at the MLE is zero, i.e., $\ell_n'(\widehat{\boldsymbol{\theta}}) = \mathbf{0}$. By the Taylor expansion of the score function around $\boldsymbol{\theta}_0$, we have

$$\mathbf{0} = \ell_n'(\widehat{\boldsymbol{\theta}}) = \ell_n'(\boldsymbol{\theta}_0) + \left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right) \ell_n''(\widetilde{\boldsymbol{\theta}})$$

where $\tilde{\boldsymbol{\theta}}$ lies between $\widehat{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}_0$. Define $J_n(\boldsymbol{\theta}) = -\frac{1}{n}\ell_n''(\boldsymbol{\theta})$. Then we have

$$\sqrt{n}\left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right) = J_n^{-1}(\widetilde{\boldsymbol{\theta}}) n^{-1/2} \ell_n'(\boldsymbol{\theta}_0).$$

Then we need to show that:

1. $n^{-1/2}\ell_n'(\boldsymbol{\theta}_0) \xrightarrow{d} \mathcal{N}(\mathbf{0}, I(\boldsymbol{\theta}_0))$;
2. $J_n(\widetilde{\boldsymbol{\theta}}) \xrightarrow{p} I(\boldsymbol{\theta}_0)$.

For the first term, we have

$$n^{-1/2}\ell_n'(\boldsymbol{\theta}_0) = n^{-1/2} \sum_{i=1}^n \ell'(X_i; \boldsymbol{\theta}_0) = \sqrt{n}\frac{1}{n}\sum_{i=1}^n \frac{\partial \log p(X_i; \boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}},$$

thus, by the central limit theorem, we have

$$\sqrt{n}\frac{1}{n}\sum_{i=1}^n \frac{\partial \log p(X_i; \boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}} \xrightarrow{d} \mathcal{N}(\mathbf{0}, I(\boldsymbol{\theta}_0)).$$

For the second term, denote $I_0^*(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{\theta}_0}\left[-\frac{\partial^2 \log p(X;\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top}\right]$, by the triangle inequality, we have

$$\left\|J_n(\widetilde{\boldsymbol{\theta}}) - I(\boldsymbol{\theta}_0)\right\| \leq \left\|J_n(\widetilde{\boldsymbol{\theta}}) - I_0^*(\widetilde{\boldsymbol{\theta}})\right\| + \left\|I_0^*(\widetilde{\boldsymbol{\theta}}) - I(\boldsymbol{\theta}_0)\right\|.$$

According to Lemma, we have

$$\left\|J_n(\widetilde{\boldsymbol{\theta}}) - I_0^*(\widetilde{\boldsymbol{\theta}})\right\| \leq \sup_{\boldsymbol{\theta} \in \Theta} \|J_n(\boldsymbol{\theta}) - I_0^*(\boldsymbol{\theta})\| \xrightarrow{p} 0.$$

Since $I_0^*(\boldsymbol{\theta})$ is continuous in $\boldsymbol{\theta}$, and $\widetilde{\boldsymbol{\theta}} \xrightarrow{p} \boldsymbol{\theta}_0$, by the continuous mapping theorem, we have

$$I_0^*(\widetilde{\boldsymbol{\theta}}) \xrightarrow{p} I(\boldsymbol{\theta}_0).$$

Combining the above results, we have

$$J_n(\widetilde{\boldsymbol{\theta}}) \xrightarrow{p} I(\boldsymbol{\theta}_0).$$

which completes the proof. □

29.1.4 Efficiency

29.2 Modified Likelihood Estimator

Seek a modified likelihood function that depends on as few of the nuisance parameters as possible while sacrificing as little information as possible.

29.2.1 Marginal Likelihood

29.2.2 Conditional Likelihood

Let $\boldsymbol{\theta} = (\boldsymbol{\varphi}, \boldsymbol{\Lambda})$, where $\boldsymbol{\varphi}$ is the parameter vector of interest and $\boldsymbol{\Lambda}$ is a vector of nuisance parameters. The conditional likelihood can be obtained as follows:

1. Find the complete sufficient statistic $S_{\boldsymbol{\Lambda}}$, respectively for $\boldsymbol{\Lambda}$.
2. Construct the conditional log-likelihood

$$\ell_c = \ln\left(f_{Y|S_{\boldsymbol{\Lambda}}}\right) \tag{29.6}$$

where $f_{Y|S_{\boldsymbol{\Lambda}}}$ is the conditional distribution of the response Y given $S_{\boldsymbol{\Lambda}}$.

Remark. Two cases might occur, that, for fixed $\boldsymbol{\varphi}_0$, $S_{\boldsymbol{\Lambda}}(\boldsymbol{\varphi}_0)$ depends on $\boldsymbol{\varphi}_0$; or $S_{\boldsymbol{\Lambda}}(\boldsymbol{\varphi}_0) = S_{\boldsymbol{\Lambda}}$ is independent of $\boldsymbol{\varphi}_0$.

1. Independent:
2. Dependent:

Suppose that the log-likelihood for $\boldsymbol{\theta} = (\boldsymbol{\varphi}, \boldsymbol{\Lambda})$ can be written in the exponential family form

$$\ell(\boldsymbol{\theta}, \mathbf{y}) = \boldsymbol{\theta}'\mathbf{s} - b(\boldsymbol{\theta}) \tag{29.7}$$

Also, suppose $\ell(\boldsymbol{\theta}, \mathbf{y})$ has a decomposition of the form

$$\ell(\boldsymbol{\theta}, \mathbf{y}) = \boldsymbol{\varphi}'\mathbf{s}_1 + \boldsymbol{\Lambda}'\mathbf{s}_2 - b(\boldsymbol{\varphi}, \boldsymbol{\Lambda}) \tag{29.8}$$

Remark. The above decomposition can be achieved only if $\boldsymbol{\varphi}$ is a linear function of θ. The choice of nuisance parameter λ is arbitrary and the inferences regarding $\boldsymbol{\varphi}$ should be unaffected by the parameterization chosen for λ.

The conditional likelihood of the data $\mathbf{Y}$ given $\mathbf{s}_2$ is

$$\ell(\boldsymbol{\varphi} \mid \mathbf{s}_2) = \boldsymbol{\varphi}'\mathbf{s}_1 - b^*(\boldsymbol{\varphi}, \boldsymbol{\Lambda}) \tag{29.9}$$

which is independent of the nuisance parameter and may be used for inferences regarding $\boldsymbol{\varphi}$.

Example. $Y_1 \sim P(\mu_1), Y_2 \sim P(\mu_2)$ are independent. Suppose $\varphi = \ln\left(\frac{\mu_2}{\mu_1}\right) = \ln(\mu_2) - \ln(\mu_1)$ is the parameter of interest and the nuisance parameter is

1. $\lambda_1 = \ln(\mu_1)$.
2.

Then, give the conditional log-likelihood for different nuisance parameters.

Proof. 1. The log-likelihood function in the form of (φ, λ) is

$$\begin{aligned}
\ell(\phi, \lambda_1) \propto & \ln\left[e^{-(\mu_1+\mu_2)}\mu_1^{y_1}\mu_2^{y_2}\right] \\
= & -(\mu_1+\mu_2) + y_1\ln(\mu_1) + y_2\ln(\mu_2) \\
= & -\mu_1\left(1+\frac{\mu_2}{\mu_1}\right) + y_1\ln(\mu_1) + y_2\ln(\mu_1) \\
& - y_2[\ln(\mu_1) - \ln(\mu_2)] \\
= & -\mathrm{e}^{\lambda_1}(1+\mathrm{e}^{\varphi}) + (y_1+y_2)\lambda_1 - y_2\varphi \\
= & s_1\varphi + s_2\lambda_1 - b(\varphi, \lambda_1)
\end{aligned}$$

where $s_1 = -y_2, s_2 = y_1 + y_2, b(\varphi, \lambda_1) = e^{\lambda_1}(1+e^{\varphi})$.
Then, the conditional distribution of Y_1, Y_2 given $S_2 = Y_1+Y_2$ is $b\left(S_2, \frac{\mu_1}{\mu_1+\mu_2}\right)$, thus,

$$\begin{aligned}
\ell(\varphi \mid S_2 = s_2) \propto & y_1\ln\left(\frac{\mu_1}{\mu_1+\mu_2}\right) + y_2\ln\left(\frac{\mu_2}{\mu_1+\mu_2}\right) \\
= & y_1\ln\left(\frac{\mu_1}{\mu_1+\mu_2}\right) + y_2\ln\left(\frac{\mu_1}{\mu_1+\mu_2}\right) \\
& - y_2\left[\ln\left(\frac{\mu_1}{\mu_1+\mu_2}\right) - \ln\left(\frac{\mu_2}{\mu_1+\mu_2}\right)\right] \\
= & (y_1+y_2)\ln\left(\frac{1}{1+e^{\varphi}}\right) - y_2\varphi \\
= & s_1\varphi - b^*(\varphi, s_2)
\end{aligned}$$

where $b^*(\varphi, s_2) = -s_2\ln\left(\frac{1}{1+\varphi^{-1}}\right)$.

□

29.2.3 Profile Likelihood

29.2.4 Quasi Likelihood

29.3 Minimum-Variance Unbiased Estimator

Definition 29.3.1 (UMVU Estimators)

An unbiased estimator $\delta(\mathbf{X})$ of $g(\theta)$ is the uniform minimum variance unbiased (UMVU) estimator of $g(\theta)$ if

$$\mathrm{Var}_\theta \delta(\mathbf{X}) \leq \mathrm{Var}_\theta \delta'(\mathbf{X}), \quad \forall \theta \in \Theta, \tag{29.10}$$

where $\delta'(\mathbf{X})$ is any other unbiased estimator of $g(\theta)$.

Remark. If there exists an unbiased estimator of g, the estimand g will be called U-estimable.

1. If $T(\mathbf{X})$ is a complete sufficient statistic, estimator $\delta(\mathbf{X})$ that only depends on $T(\mathbf{X})$, then for any U-estimable function $g(\theta)$ with

$$E_\theta \delta(T(\mathbf{X})) = g(\theta), \quad \forall \theta \in \Theta, \tag{29.11}$$

   hence, $\delta(T(\mathbf{X}))$ is the unique UMVU estimator of $g(\theta)$.
2. If $T(\mathbf{X})$ is a complete sufficient statistic and $\delta(\mathbf{X})$ is any unbiased estimator of $g(\theta)$, then the UMVU estimator of $g(\theta)$ can be obtained by

$$E\left[\delta(\mathbf{X}) \mid T(\mathbf{X})\right]. \tag{29.12}$$

Example (Estimating Polynomials of a Normal Variance)**.** Let $X_1, \ldots, X_n$ be distributed with joint density

$$\frac{1}{(\sqrt{2\pi}\sigma)^n} \exp\left[-\frac{1}{2\sigma^2} \sum (x_i - \xi)^2\right]. \tag{29.13}$$

Discussing the UMVU estimators of ξ^r, σ^r, ξ/σ.

Proof. 1. σ **is known**:
Since $\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i$ is the complete sufficient statistic of X_i, and

$$E(\bar{X}) = \xi,$$

then the UMVU estimator of ξ is $\bar{X}$.
Therefore, the UMVU estimator of ξ^r is $\bar{X}^r$ and the UMVU estimator of ξ/σ is $\bar{X}/\sigma$.

2. ξ **is known**:
   Since $s^r = \sum (x_i - \xi)^r$ is the complete sufficient statistic of X_i.
   Assume
   $$E\left[\frac{s^r}{\sigma^r}\right] = \frac{1}{K_{n,r}},$$
   where $K_{n,r}$ is a constant depends on n, r.
   Since $s^2/\sigma^2 \sim \mathrm{Ga}(n/2, 1/2) = \chi^2(n)$, then
   $$E\left[\frac{s^r}{\sigma^r}\right] = E\left[\left(\frac{s^2}{\sigma^2}\right)^{\frac{r}{2}}\right] = \int_0^\infty x^{\frac{r}{2}} \frac{1}{2^{\frac{n}{2}}\Gamma(\frac{n}{2})} x^{\frac{n}{2}-1} e^{-\frac{x}{2}}\,\mathrm{d}x = \frac{\Gamma\left(\frac{n+r}{2}\right)}{\Gamma(\frac{n}{2})} \cdot 2^{\frac{r}{2}}.$$
   therefore,
   $$K_{n,r} = \frac{\Gamma(\frac{n}{2})}{2^{\frac{r}{2}} \cdot \Gamma\left(\frac{n+r}{2}\right)}.$$
   Hence,
   $$E\left[s^r K_{n,r}\right] = \sigma^r \text{ and } E[\xi s^{-1} K_{n,-1}] = \xi/\sigma,$$
   which means the UMVU estimator of σ^r is $s^r K_{n,r}$ and the UMVU estimator of ξ/σ is $\xi s^{-1} K_{n,-1}$.
3. **Both** ξ **and** σ **is unknown**:
   Since $(\bar{X}, s_x^r)$ are the complete sufficient statistic of X_i, where $s_x^2 = \sum \left(x_i - \bar{X}\right)^r$.
   Since $s_x^2/\sigma^2 \sim \chi^2(n-1)$, then
   $$E\left[\frac{s_x^r}{\sigma^r}\right] = \frac{1}{K_{n-1,r}}.$$
   Hence,
   $$E\left[s_x^r K_{n-1,r}\right] = \sigma^r,$$
   which means the UMVU estimator of σ^r is $s_x^r K_{n-1,r}$, and
   $$E(\bar{X}^r) = \xi^r,$$
   which means the UMVU estimator of ξ^r is $\bar{X}^r$.
   Since $\bar{X}$ and s_x^r are independent, then
   $$E[\bar{X} s_x^{-1} K_{n-1,-1}] = \xi/\sigma$$
   which means the UMVU estimator of ξ/σ is $\bar{X} s_x^{-1} K_{n-1,-1}$.

□

Example. Let $X_1, \ldots, X_n$ be i.i.d sample from $U(\theta_1 - \theta_2, \theta_1 + \theta_2)$, where $\theta_1 \in \mathbb{R}, \theta_2 \in \mathbb{R}^+$. Discussing the UMVU estimators of θ_1, θ_2.

Proof. Let $X_{(i)}$ be the i-th order statistic of X_i, then $\left(X_{(1)}, X_{(n)}\right)$ is the complete and sufficient statistic for (θ_1, θ_2). Thus it suffices to find a function $\left(X_{(1)}, X_{(n)}\right)$, which is unbiased of (θ_1, θ_2).

Let

$$Y_i = \frac{X_i - (\theta_1 - \theta_2)}{2\theta_2} \sim U(0,1),$$

and

$$Y_{(i)} = \frac{X_{(i)} - (\theta_1 - \theta_2)}{2\theta_2},$$

be the i-th order statistic of Y_i, then we got

$$\begin{aligned} E[X_{(1)}] &= 2\theta_2 E[Y_{(1)}] + (\theta_1 - \theta_2) \\ &= 2\theta_2 \int_0^1 ny(1-y)^{n-1}\,\mathrm{d}y + (\theta_1 - \theta_2) \\ &= \theta_1 - \frac{3n+1}{n+1}\theta_2 \\ E[X_{(n)}] &= 2\theta_2 E[Y_{(n)}] + (\theta_1 - \theta_2) \\ &= 2\theta_2 \int_0^1 ny^n\,\mathrm{d}y + (\theta_1 - \theta_2) \\ &= \theta_1 + \frac{n-1}{n+1}\theta_2 \end{aligned}$$

.

Thus,

$$\begin{aligned} \theta_1 &= E\left[\frac{n-1}{4n}X_{(1)} + \frac{3n+1}{4n}X_{(n)}\right], \\ \theta_2 &= E\left[-\frac{n+1}{4n}X_{(1)} + \frac{n+1}{4n}X_{(n)}\right], \end{aligned}$$

which means the UMVU estimator is

$$\hat{\theta_1} = \frac{n-1}{4n}X_{(1)} + \frac{3n+1}{4n}X_{(n)}, \quad \hat{\theta_2} = -\frac{n+1}{4n}X_{(1)} + \frac{n+1}{4n}X_{(n)}.$$

□

29.4 Accuracy of Estimators

Example (Normal Probability)**.** Let $X_1, \ldots, X_n$ be iid as $\mathcal{N}(\theta, 1)$ and consider the estimation of $p = P(X_i \le u) = \Phi(u - \theta)$. The maximum likelihood estimator of p is $\hat{p} = \Phi(u - \bar{X})$, and we shall attempt to obtain large-sample approximations for the bias and variance of this estimator.

Proof. Since $\bar{X} - \theta$ is likely to be small, it is natural to write

$$\Phi(u - \bar{X}) = \Phi[(u - \theta) - (\bar{X} - \theta)]$$

and to expand the right side about $u - \theta$ by Taylor's theorem as

$$\begin{aligned}\Phi(u - \bar{X}) =& \Phi(u - \theta) - (\bar{X} - \theta)\phi(u - \theta) + \frac{1}{2}(\bar{X} - \theta)^2\phi'(u - \theta) \\ &- \frac{1}{6}(\bar{X} - \theta)^3\phi''(u - \theta) + \frac{1}{24}(\bar{X} - \theta)^4\phi'''(\xi),\end{aligned} \tag{29.14}$$

where ξ is a random quantity that lies between $u - \theta$ and $u - \bar{X}$.

To calculate the bias, we take the expectation of (29.14), which yields

$$\mathbb{E}[\hat{p}] = p + \frac{1}{2n}\phi'(u - \theta) + \frac{1}{24}\mathbb{E}\left[(\bar{X} - \theta)^4\phi'''(\xi)\right].$$

Since the derivatives of $\phi(x)$ all are of the form $P(x)\phi(x)$, where $P(x)$ is a polynomial in x and are therefore all bounded. It follows that

$$\left|\mathbb{E}(\bar{X} - \theta)^4\phi'''(\xi)\right| < M\mathbb{E}(\bar{X} - \theta)^4 = 3M/n^2,$$

for some finite M. Using the fact that $\phi'(x) = -x\phi(x)$, we therefore find that

$$\mathbb{E}(\hat{p}) = p - \frac{1}{2n}(u - \theta)\phi(u - \theta) + O\left(1/n^2\right),$$

where the error term is uniformly $O\left(1/n^2\right)$. The estimator δ therefore has a bias of order $1/n$ which tends to zero as $\theta \to \pm\infty$.

In the same way, by Taylor's theorem, we can expand the square of the estimator as

$$\begin{aligned}\Phi^2(u - \bar{X}) =& \Phi^2(u - \theta) - 2(\bar{X} - \theta)\Phi(u - \theta)\phi(u - \theta) \\ &+ (\bar{X} - \theta)^2\left[\phi^2(\mu - \theta) + \Phi(\mu - \theta)\phi'(\mu - \theta)\right] \\ &- \frac{1}{3}(\bar{X} - \theta)^3\left[2\phi(\xi)\phi'(\xi) + \Phi(\xi)\phi''(\xi) + \phi'^2(\xi)\right]\end{aligned}$$

where ξ lies between $u - \theta$ and $u - \bar{X}$. Taking the expectation of this expansion, we find that

$$\mathbb{E}[\hat{p}^2] = p^2 + \frac{1}{n}\phi^2(u - \theta) - \frac{p}{n}(\mu - \theta)\phi(u - \theta),$$

and

$$[\mathbb{E}(\hat{p})]^2 = p^2 - \frac{p}{n}(\mu - \theta)\phi(u - \theta) + O\left(1/n^2\right).$$

Thus

$$\mathrm{Var}(\hat{p}) = \mathbb{E}[\hat{p}^2] - [\mathbb{E}(\hat{p})]^2 = \frac{1}{n}\phi^2(u-\theta) + O\left(1/n^2\right).$$

and hence that

$$\mathrm{Var}(\sqrt{n}\delta) \to \phi^2(u-\theta)$$

Since $\sqrt{n}(\bar{X}-\theta) \xrightarrow{d} N(0,1)$, and $f(\theta) = \Phi(\mu-\theta)$ is differentiable with $f'(\theta) = -\phi(\mu-\theta)$ not equal to zero when $\mu \neq \theta$, then by the delta method (13.1.4), we have

$$\sqrt{n}(\hat{p}-p) \xrightarrow{d} N\left(0, \phi^2(u-\theta)\right),$$

the limit of the variance in this case is equal to the asymptotic variance.

It is interesting to see what happens if the expansion (29.14) is carried one step less far. Then

$$E[\Phi(u-\bar{X})] = p + \frac{1}{2n}\phi'(u-\theta) + \frac{1}{6}E\left[(\bar{X}-\theta)^3\phi''(\xi)\right].$$

Since the third derivative of ϕ is bounded, the remainder now satisfies

$$\frac{1}{6}E\left|(\bar{X}-\theta)^3\phi'''(\xi)\right| < M'E|\bar{X}-\theta|^3 = O\left(\frac{1}{n^{3/2}}\right).$$

The conclusion is therefore weaker than before. □

Chapter 30

Interval Estimation

30.1 Confidence Interval

30.2 Pivot

30.3 Likelihood Interval

30.4 Prediction Interval

30.5 Tolerance Interval

Chapter 31

Testing Hypotheses

31.1 Testing Hypotheses

31.2 Parametric Tests

31.3 Specific Tests

31.3.1 Goodness of Fit

Likelihood-Ratio Test

31.3.2 Rank statistics

Chapter 32

Bayesian Inference

32.1 Bayes Estimator

We shall look for some estimators that make the risk function $R(\theta, \delta)$ small in some overall sense. There are two ways to solve it: minimize the average risk, and minimize the maximum risk.

This chapter will discuss the first method, also known as, Bayes Estimator.

Definition 32.1.1 (Bayes Estimator)

The Bayes Estimator δ with respect to Λ is minimizing the Bayes Risk of δ

$$r(\Lambda, \delta) = \int R(\theta, \delta) \, \mathrm{d}\Lambda(\theta) \tag{32.1}$$

where Λ is the probability distribution.

In Bayesian arguments, it is important to keep track of which variables are being conditioned. Hence, the notations are as follows:

- The density of X will be denoted by $X \sim f(x \mid \theta)$.
- The prior distribution will be denoted by $\Pi \sim \pi(\theta \mid \lambda)$ or $\Lambda \sim \gamma(\lambda)$, where λ is another parameter (sometimes called a hyperparameter).
- The posterior distribution, which calculates the conditional distributions as that of θ given x and λ, or λ given x, which is denoted by $\Pi \sim \pi(\theta \mid x, \lambda)$ or $\Lambda \sim \gamma(\lambda \mid x)$, that is

$$\pi(\theta \mid x, \lambda) = \frac{f(x \mid \theta) \, \pi(\theta \mid \lambda)}{m(x \mid \lambda)}, \tag{32.2}$$

where marginal distributions $m(x \mid \lambda) = \int f(x \mid \theta) \, \pi(\theta \mid \lambda) \, \mathrm{d}\theta$.

Theorem 32.1.1

Let Θ have distribution Λ, and given $\Theta = \theta$, let X have distribution P_θ. Suppose, the following assumptions hold for the problem of estimating $g(\Theta)$ with non-negative loss function $L(\theta, d)$,

- There exists an estimator δ_0 with finite risk.
- For almost all x, there exists a value $\delta_\Lambda(x)$ minimizing

$$E\{L[\Theta, \delta(x)] \mid X = x\}. \tag{32.3}$$

Then, $\delta_\Lambda(x)$ is a Bayes Estimator.

Remark. Improper prior

Corollary 32.1.1

Suppose the assumptions of Theorem 32.1.1 hold.

1. If $L(\theta, d) = [d - g(\theta)]^2$, then

$$\delta_\Lambda(x) = E[g(\Theta) \mid x]. \tag{32.4}$$

2. If $L(\theta, d) = w(\theta)[d - g(\theta)]^2$, then

$$\delta_\Lambda(x) = \frac{E[w(\theta) g(\Theta) \mid x]}{E[w(\theta) \mid x]}. \tag{32.5}$$

3. If $L(\theta, d) = |d - g(\theta)|$, then $\delta_\Lambda(x)$ is any median of the conditional distribution of Θ given x.
4. If

$$L(\theta, d) = \begin{cases} 0 \text{ when } |d - \theta| \leq c \\ 1 \text{ when } |d - \theta| > c \end{cases},$$

then $\delta_\Lambda(x)$ is the midpoint of the interval I of length $2c$ which maxmizes $P(\Theta \in I \mid x)$.

Proof. □

Theorem 32.1.2

Necessary condition for Bayes Estimator

Methodologies have been developed to deal with the difficulty which sometimes incorporates frequentist measures to assess the choice of Λ.

- Empirical Bayes.
- Hierarchical Bayes.
- Robust Bayes.

- Objective Bayes.

32.1.1 Single-Prior Bayes

The Single-Prior Bayes model in a general form as

$$\begin{aligned} X \mid \theta &\sim f(x \mid \theta), \\ \Theta \mid \gamma &\sim \pi(\theta \mid \lambda), \end{aligned} \tag{32.6}$$

where we assume that the functional form of the prior and the value of λ is known (we will write it as $\gamma = \gamma_0$).

Given a loss function $L(\theta, d)$, we would then determine the estimator that minimizes

$$\int L(\theta, d(x))\, \pi(\theta \mid x)\, \mathrm{d}\theta, \tag{32.7}$$

where $\pi(\theta \mid x)$ is posterior distribution given by

$$\pi(\theta \mid x) = \frac{f(x \mid \theta)\, \pi(\theta \mid \gamma_0)}{\int f(x \mid \theta)\, \pi(\theta \mid \gamma_0)\, \mathrm{d}\theta}.$$

In general, this Bayes estimator under squared error loss is given by

$$E(\Theta \mid x) = \frac{\int \theta f(x \mid \theta)\, \pi(\theta \mid \gamma_0)\, \mathrm{d}\theta}{\int f(x \mid \theta)\, \pi(\theta \mid \gamma_0)\, \mathrm{d}\theta}. \tag{32.8}$$

Example. Consider

$$\begin{aligned} X_i &\overset{\text{i.i.d}}{\sim} N(\mu, \Gamma^{-1}), \quad i = 1, 2, \ldots, n \\ \mu &\sim N(0, 1), \\ \Gamma &\sim \text{Gamma}(2, 1), \end{aligned}$$

calculate the Single-Prior Bayes estimator under squared error loss.

Proof.

$$\begin{aligned} p(\mathbf{X} \mid \mu, \Gamma) &= \Gamma^n (2\pi)^{-\frac{n}{2}} \exp\left[-2\Gamma^2 \sum_{i=1}^{n} (x_i - \mu)^2\right], \\ p(\mu) &= \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{\mu^2}{2}\right), \\ p(\Gamma) &= \frac{1}{\Gamma(2)} \Gamma \exp(-\Gamma). \end{aligned}$$

Therefore,

$$h(\mathbf{X}, \mu, \Gamma) = C\Gamma^n \exp\left[-2\Gamma^2 \sum_{i=1}^{n} (x_i - \mu)^2\right] \exp\left(-\frac{\mu^2}{2}\right) \Gamma \exp(-\Gamma),$$

where $C = \frac{(2\pi)^{-\frac{n+1}{2}}}{\Gamma(2)}$.

For μ, we have

$$\pi\left(\mu \mid \mathbf{X}, \Gamma\right) = \frac{h\left(\mathbf{X}, \mu, \Gamma\right)}{p(\mu \mid \mathbf{X})}$$

□

For exponential families

Theorem 32.1.3

32.1.2 Hierarchical Bayes

In a Hierarchical Bayes model, rather than specifying the prior distribution as a single function, we specify it in a **hierarchy**. Thus, the Hierarchical Bayes model in a general form as

$$\begin{aligned} X \mid \theta &\sim f\left(x \mid \theta\right), \\ \Theta \mid \gamma &\sim \pi\left(\theta \mid \lambda\right), \\ \Gamma &\sim \psi\left(\gamma\right), \end{aligned} \tag{32.9}$$

where we assume that $\psi\left(\cdot\right)$ is known and not dependent on any other unknown hyperparameters.

Remark. We can continue this hierarchical modeling and add more stages to the model, but this is not then done in practice.

Given a loss function $L\left(\theta, d\right)$, we would then determine the estimator that minimizes

$$\int L\left(\theta, d\left(x\right)\right) \pi\left(\theta \mid x\right) \mathrm{d}\theta, \tag{32.10}$$

where $\pi\left(\theta \mid x\right)$ is posterior distribution given by

$$\pi\left(\theta \mid x\right) = \frac{\int f\left(x \mid \theta\right) \pi\left(\theta \mid \gamma\right) \psi\left(\gamma\right) \mathrm{d}\gamma}{\int\int f\left(x \mid \theta\right) \pi\left(\theta \mid \gamma\right) \psi\left(\gamma\right) \mathrm{d}\theta \, \mathrm{d}\gamma}.$$

Remark. The posterior distribution can also be written as

$$\pi\left(\theta \mid x\right) = \int \pi\left(\theta \mid x, \gamma\right) \pi\left(\gamma \mid x\right) \mathrm{d}\gamma,$$

where $\pi\left(\gamma \mid x\right)$ is the posterior distribution of Γ, unconditional on θ. The equation 32.10 can be written as

$$\int L\left(\theta, d\left(x\right)\right) \pi\left(\theta \mid x\right) \mathrm{d}\theta = \int \left[\int L\left(\theta, d\left(x\right)\right) \pi\left(\theta \mid x, \gamma\right) \mathrm{d}\theta\right] \pi\left(\gamma \mid x\right) \mathrm{d}\gamma.$$

which shows that **the Hierarchical Bayes estimator can be thought of as a mixture of Single-Prior estimators**.

Example (Poisson Hierarchy)**.** Consider

$$\begin{aligned} X_i \mid \lambda &\overset{\text{i.i.d}}{\sim} \text{Poisson}(\lambda), \quad i = 1, 2 \ldots, n \\ \lambda \mid b &\sim \text{Gamma}(a, b), \text{a known}, \\ \frac{1}{b} &\sim \text{Gamma}(k, \tau), \end{aligned} \tag{32.11}$$

calculate the Hierarchical Bayes estimator under squared error loss.

Theorem 32.1.4

For the Hierarchical Bayes model (32.9),

$$K[\pi(\lambda \mid x), \psi(\lambda)] < K[\pi(\theta \mid x), \pi(\theta)], \tag{32.12}$$

where K is the Kullback-Leibler information for discrimination between two densities.

Proof. □

Remark.

32.1.3 Empirical Bayes

32.1.4 Bayes Prediction

Chapter 33

Nonparametric Statistics

33.1 Probability Distribution

33.1.1 Cumulative Distribution Function

Let $X_1, \ldots, X_n \sim F$ where $F(x) = \mathbb{P}(X \leq x)$ is a distribution function on the real line.

Definition 33.1.1 (Empirical Cumulative Distribution Function)

The empirical cumulative distribution function $\widehat{F}_n$ is the CDF that puts mass $1/n$ at each data point X_i, that,

$$\widehat{F}_n(x) = \frac{1}{n} \sum_{i=1}^{n} I\left(X_i \leq x\right) \tag{33.1}$$

Denote $Y = \sum_{i=1}^{n} I\left(X_i \leq x\right)$, then Y is a binomial random variable with parameters n and $F(x)$, that is, $Y \sim b(n, F(x))$. Therefore, we have

$$E\left[\hat{F}_n(x)\right] = F(x), \quad \text{Var}\left[\hat{F}_n(x)\right] = \frac{1}{n} F(x)[1 - F(x)]$$

so that $\hat{F}_n(a)$ is unbiased and its variance is of order $1/n$. In addition, it follows from CLT (15.1.2) that

$$\sqrt{n}\left[\hat{F}_n(x) - F(x)\right] \xrightarrow{d} \mathcal{N}(0, F(x)[1 - F(x)]).$$

Thus, $\hat{F}_n(x)$ is a consistent estimator of $F(x)$ for each fixed x.

However, a much stronger consistency property can be asserted if the difference between $\hat{F}_n(x)$ and $F(x)$ is considered not only for a fixed x but simultaneously

for all x, namely

$$D_n = \sup_x \left|\hat{F}_n(x) - F(x)\right| \xrightarrow{P} 0, \quad n \to \infty.$$

For a still stronger result, see, for example, Serfling (1980, Section 2.1.4) or Billingsley (1986, Theorem 20.6).

33.1.2 Probability Density Function

Histogram Estimator Since

$$f(x) = \lim_{h \to 0} \frac{F(x+h) - F(x-h)}{2h},$$

one might consider the estimator

$$\hat{f}_n(x) = \frac{\hat{F}_n(x+h) - \hat{F}_n(x-h)}{2h}.$$

and hope that with $h = h_n$ tending to 0 as $n \to \infty$, the estimator $\hat{f}_n(x)$, the so-called Rosenblatt estimator, will be consistent for $f(x)$.

Proof. The basic properties of $\hat{p}_n$ are obtained from the fact that

$$n\left[\hat{F}_n(x+h_n) - \hat{F}_n(x-h_n)\right] \sim b(n,p),$$

where $p = F(x+h_n) - F(x-h_n)$. It follows that

$$\mathbb{E}\left[\hat{f}_n(x)\right] = \frac{F(x+h_n) - F(x-h_n)}{2h}.$$

The bias is therefore

$$b(x) = \frac{F(x+h_n) - F(x-h_n)}{2h} - \lim_{h_n \to 0} \frac{F(x+h_n) - F(x-h_n)}{2h}$$

which tends to be zero, provided $h_n \to 0$, as $n \to \infty$.

Similarly, the variance of $\hat{f}_n(x)$ is

$$\mathrm{Var}\left[\hat{f}_n(x)\right] = \frac{p(1-p)}{4nh^2}.$$

As $h_n \to 0$, the value of p

$$p_n = F\left(x+h_n\right) - F\left(x-h_n\right) \to 0.$$

thus,

$$\operatorname{Var}\left[\hat{f}_n(x)\right] \sim \frac{p_n}{2h_n} \cdot \frac{1}{2nh_n}.$$

Since the first factor on the right side tends to $f(x) > 0$, $\operatorname{Var}\left[\hat{f}_n(x)\right] \to 0$ as $h_n \to 0$ if in addition $nh_n \to \infty$, that is, if h_n tends to 0 more slowly than $1/n$ or, equivalently, if $\frac{1}{n} = o(h_n)$. From these results, we immediately obtain sufficient conditions for the consistency of $\hat{f}_n(x)$. □

It is interesting to note that $\hat{f}_n(x)$ is itself a probability density. Since it is non-negative, one only needs to show that $\int_{-\infty}^{\infty} \hat{f}_n(x)\,\mathrm{d}x = 1$. This is easily seen by writing

$$\hat{f}_n(x) = \frac{1}{2nh} \sum_{i=1}^{n} I(x - h_n \le X_i \le x + h_n)$$

Then

$$\int \hat{f}_n(x)\,\mathrm{d}x = \frac{1}{2nh} \sum_{i=1}^{n} \int_{-\infty}^{\infty} I_i(x)\,\mathrm{d}x = \frac{1}{2nh} \sum_{i=1}^{n} \int_{x_j-h}^{x_j+h} \mathrm{d}x = 1.$$

Remark. Although $\hat{f}_n(x)$ is consistent for estimating $f(x)$ when the h_n's satisfy $h_n \to 0$ and $nh_n \to \infty$, note that $\hat{f}_n(x)$ as an estimator of $f(x)$ involves two approximations: $f(x)$ by $[F(x + h_n) - F(x - h_n)]/2h_n$ and the latter by $\hat{f}_n(x)$. As a result, the estimator turns out to be less accurate than one might have hoped. Besides, the estimator $\hat{f}_n(x)$, although a density, is a step function with discontinuities at every point $x_i \pm h_n, i = 1, \ldots, n$. If we assume the true $f(x)$ to be a smooth density, we may prefer its estimator also to be smooth.

Kernel Density Estimation We can rewrite it in the form:

$$\hat{f}_n(x) = \frac{1}{2nh_n} \sum_{i=1}^{n} I\left(x - h_n < X_i \le x + h_n\right) = \frac{1}{nh_n} \sum_{i=1}^{n} K_0\left(\frac{X_i - x}{h}\right),$$

where $K_0(u) = \frac{1}{2} I(-1 < u \le 1)$. A simple generalization of the Rosenblatt estimator is given by

$$\hat{f}_n(x) = \frac{1}{nh_n} \sum_{i=1}^{n} K\left(\frac{X_i - x}{h_n}\right), \tag{33.2}$$

where $K : \mathbb{R} \to \mathbb{R}$ is an integrable function satisfying $\int K(u)\,\mathrm{d}\mu = 1$. Such a function K is called a kernel and the parameter h_n is called a bandwidth of the estimator (33.2). The function $x \mapsto \hat{f}_n(x)$ is called the kernel density estimator or the Parzen-Rosenblatt estimator. In addition, we shall restrict attention to kernel K that are symmetric about 0.

In generalization of (33.2), note that $\hat{f}_n$ is a probability density since it is non-negative and since

$$\int \hat{f}_n(x)\,\mathrm{d}x = \frac{1}{n}\sum_{i=1}^{n}\int K\left(\mu - x_i\right)\mathrm{d}\mu = \frac{1}{n}\sum_{i=1}^{n}\int K(\mu)\,\mathrm{d}\mu = 1.$$

Theorem 33.1.1

Let the density f satisfies $f(x) \le f_{\max} < \infty$ for all x and let K be a kernel that $intK^2(\mu)\,\mathrm{d}\mu < \infty$. Then for any sequence $h_n, n = 1, 2, \ldots,$

$$\mathrm{Var}\left[\hat{f}_n(x)\right] = \frac{f_{\max}}{nh_n}\int K^2(\mu)\,\mathrm{d}\mu + o\left(\frac{1}{nh_n}\right).$$

Proof. Denote

$$\eta_i(x) = K\left(\frac{x - X_i}{h}\right) - \mathbb{E}\left[K\left(\frac{x - X_i}{h}\right)\right].$$

The random variables $\eta_i(x)$ are i.i.d with zero mean and variance

$$\begin{aligned}\mathbb{E}\left[\eta_i^2(x)\right] \le& \mathbb{E}\left[K^2\left(\frac{x - X_i}{h}\right)\right] \\ =& \int K^2(\frac{x - z}{h})f(z)\,\mathrm{d}z \le f_{\max}h\int K^2(\mu)\,\mathrm{d}\mu.\end{aligned}$$

Then, we have

$$\mathrm{Var}\left[\hat{f}_n(x)\right] = \frac{1}{n^2h^2}\sum_{i=1}^{n}\mathbb{E}\left[\eta_i^2(x)\right] = \frac{1}{nh^2}\mathbb{E}\left[\eta_1^2(x)\right] \le \frac{f_{\max}}{nh}\int K^2(\mu)\,\mathrm{d}\mu.$$

□

Theorem 33.1.2 (Bias of Kernel Density Estimator)

Let f be three times differentiable with a bounded third derivative in a neighborhood of y and let K be a kernel symmetric about 0, with

$$\int K^2(\mu)\,\mathrm{d}\mu < \infty, \quad \int \mu^2 K(\mu)\,\mathrm{d}\mu = \tau^2 < \infty, \quad \int |\mu|^3 K(\mu)\,\mathrm{d}\mu < \infty.$$

Then for any sequence $h_n, n = 1, 2, \ldots,$

$$\mathrm{Bias}\left[\hat{f}_n(x)\right] = \frac{1}{2}h_n^2 f''(x)\tau^2 + o\left(h_n^2\right).$$

Proof. Suppressing the subscript n, we find for the bias of $\hat{f}_n(x)$,

$$\begin{aligned}\text{Bias}\left[\hat{f}_n(x)\right] &= E\left[\frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}K\left(\frac{x-X_i}{h}\right)\right] - f(x)\\ &= \int \frac{1}{h}K\left(\frac{x-X}{h}\right) f(X)\,\mathrm{d}X - f(x)\\ &= \int K(\mu)[f(x-h\mu) - f(x)]\,\mathrm{d}\mu.\end{aligned}$$

By the Taylor expansion of $f(x-h\mu)$ about x, we have

$$f(x-h\mu) = f(x) + h\mu f'(x) + \frac{1}{2}(h\mu)^2 f''(x) + \frac{1}{6}(h\mu)^3 f'''(\xi),$$

where ξ lies between x and $x-h\mu$. Using the fact that $\int \mu K(\mu)\,\mathrm{d}\mu = 0$, we therefore have

$$\text{Bias}\left[\hat{f}_n(x)\right] = \frac{1}{2}h^2 f''(x)\int \mu^2 K(\mu)\,\mathrm{d}\mu + R_n,$$

where, with $|f'''(z)| \leq M$,

$$|R_n| \leq \frac{Mh^3}{6}\int |z|^3 K(z)dz = o\left(h^2\right),$$

which proves the first part of the theorem. □

33.2 Kernel Methods

33.2.1 Positive Definite Kernels

Definition 33.2.1 (Positive Definite Kernel)

Let $\mathcal{X}$ be a set, a function $K : \mathcal{X}\times\mathcal{X} \to \mathbb{R}$ is called a positive definite kernel on $\mathcal{X}$ if and only if it is

1. symmetric, that is,

$$K(\mathbf{x},\mathbf{x}') = K(\mathbf{x}',\mathbf{x}), \quad \forall \mathbf{x},\mathbf{x}' \in \mathcal{X} \tag{33.3}$$

2. positive definite, that is,

$$\sum_{i=1}^{n}\sum_{j=1}^{n} c_i c_j K(\mathbf{x}_i,\mathbf{x}_j) \geq 0, \tag{33.4}$$

holds for any $x_1,\ldots,x_n \in \mathcal{X}$, given $n \in \mathbb{N}, c_1,\ldots,c_n \in \mathbb{R}$.

Construction of the Reproducing Kernel Hilbert Space

Theorem 33.2.1 (Morse-Aronszajn's Theorem)

For any set $\mathcal{X}$, suppose $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is positive definite, then there is a unique RKHS $\mathcal{H} \subset \mathbb{R}^{\mathcal{X}}$ with reproducing kernel K.

Proof. 1. How to build a valid pre-RKHS $\mathcal{H}_0$?

Consider the vector space $\mathcal{H}_0 \subset \mathcal{R}^{\mathcal{X}}$ spanned by the functions $\{K(\cdot, \mathbf{x})\}_{\mathbf{x}\in\mathcal{X}}$. For any $f, g \in \mathcal{H}_0$, suppose

$$f = \sum_{i=1}^{m} a_i K(\cdot, \mathbf{x}_i), \quad g = \sum_{j=1}^{n} b_j K(\cdot, \mathbf{y}_j)$$

and let the inner product of $\mathcal{H}_0$ be

$$\langle f, g \rangle = \sum_{i=1}^{m}\sum_{j=1}^{n} a_i b_j K(\mathbf{x}_i, \mathbf{y}_j) \tag{33.5}$$

Let $\mathbf{x} \in \mathcal{X}$,

$$\langle f, K(\cdot, \mathbf{x}) \rangle_{\mathcal{H}_0} = \sum_{i=1}^{m} a_i K(\mathbf{x}, \mathbf{x}_i) = f(\mathbf{x})$$

And, we also have

$$\langle f, g \rangle_{\mathcal{H}_0} = \sum_{i=1}^{m} a_i g(\mathbf{x}_i) = \sum_{j=1}^{n} b_j f(\mathbf{y}_j)$$

Suppose

$$f = \sum_{i=1}^{m} a_i K(\cdot, \mathbf{x}_i), \quad g = \sum_{j=1}^{n} b_j K(\cdot, \mathbf{y}_j), \quad h = \sum_{k=1}^{p} c_k K(\cdot, \mathbf{z}_k)$$

(a) Linearity: For any $\alpha, \beta \in \mathbb{R}$, $\langle \alpha f + \beta g, h \rangle_{\mathcal{H}_0} = \alpha \langle f, h \rangle_{\mathcal{H}_0} + \beta \langle g, h \rangle_{\mathcal{H}_0}$.

$$\begin{aligned}
\langle \alpha f + \beta g, h \rangle_{\mathcal{H}_0} =& \left[\alpha \sum_{i=1}^{m} a_i K(\cdot, \mathbf{x}_i) + \beta \sum_{j=1}^{n} b_j K(\cdot, \mathbf{y}_j) \right] \cdot \sum_{k=1}^{p} c_k K(\cdot, \mathbf{z}_k) \\
=& \alpha \sum_{i=1}^{m}\sum_{k=1}^{p} a_i c_k K(\mathbf{x}_i, \mathbf{z}_k) + \beta \sum_{j=1}^{n}\sum_{k=1}^{p} b_j c_k K(\mathbf{y}_j, \mathbf{z}_k) \\
=& \alpha \langle f, h \rangle_{\mathcal{H}_0} + \beta \langle g, h \rangle_{\mathcal{H}_0}
\end{aligned}$$

(b) Conjugate Symmetry: $\langle f, g \rangle_{\mathcal{H}_0} = \langle g, f \rangle_{\mathcal{H}_0}$.

$$\begin{aligned}
\langle f, g \rangle_{\mathcal{H}_0} =& \sum_{i=1}^{m}\sum_{j=1}^{n} a_i b_j K(\mathbf{x}_i, \mathbf{y}_j) = \sum_{j=1}^{n}\sum_{i=1}^{m} b_j a_i K(\mathbf{y}_j, \mathbf{x}_i) \\
=& \langle g, f \rangle_{\mathcal{H}_0}
\end{aligned}$$

(c) Positive Definiteness: $\langle f, f\rangle_{\mathcal{H}_0} \geq 0$ and $\langle f, f\rangle_{\mathcal{H}_0} = 0$ if and only if $f = 0$.
By positive definiteness of K, we have:

$$\langle f, f\rangle_{\mathcal{H}_0} = \|f\|^2_{\mathcal{H}_0} = \sum_{i=1}^{m}\sum_{j=1}^{m} a_i a_j K(\mathbf{x}_i, \mathbf{x}_j) \geq 0$$

As for, $\langle f, f\rangle_{\mathcal{H}_0} = 0$ if and only if $f = 0$, we have,
$\Rightarrow$ If $f = 0$, that is $f = \sum_{i=1}^{m} a_i K(\cdot, \mathbf{x}_i) = 0$, we have

$$\langle f, f\rangle_{\mathcal{H}_0} = \sum_{i=1}^{m} a_i f = 0$$

$\Leftarrow$ For $\forall \mathbf{x} \in \mathcal{X}$, by Cauchy-Schwarz Inequality, we have,

$$|f(\mathbf{x})| = \left|\langle f, K(,\cdot\mathbf{x})\rangle_{\mathcal{H}_0}\right| \leq \|f\|_{\mathcal{H}_0} \cdot K(\mathbf{x}, \mathbf{x})^{\frac{1}{2}}$$

therefore, if $\|f\|_{\mathcal{H}_0} = 0$, then $f = 0$

Hence, the definition in equation (33.5) is a valid inner product, which is a valid pre-RKHS $\mathcal{H}_0$.

□

Examples of Kernels

Example (Gaussian Kernel)**.**

$$K(\mathbf{x}, \mathbf{y}) = \exp\left(-\frac{\|\mathbf{x} - \mathbf{y}\|^2}{2\sigma^2}\right), \quad \mathbf{x}, \mathbf{y} \in \mathbb{R}^d \tag{33.6}$$

Proof. 1. It is obvious that $K(\mathbf{x}, \mathbf{y})$ is symmetric, we only need to show $K(\mathbf{x}, \mathbf{y})$ is positive definite.

$$\begin{aligned} K(\mathbf{x}, \mathbf{y}) &= \exp\left(-\frac{\|\mathbf{x} - \mathbf{y}\|^2}{2\sigma^2}\right) \\ &= \exp\left(-\frac{1}{2\sigma^2}\|\mathbf{x}\|^2\right) \cdot \exp\left(\frac{1}{\sigma^2}\langle \mathbf{x}, \mathbf{y}\rangle\right) \cdot \exp\left(-\frac{1}{2\sigma^2}\|\mathbf{y}\|^2\right) \end{aligned}$$

By the Taylor expansion of the exponential function, that

$$\exp\left(\frac{x}{\sigma^2}\right) = \sum_{n=0}^{+\infty}\left\{\frac{x^n}{\sigma^{2n} \cdot n!}\right\}$$

Hence,

$$\exp\left(\frac{1}{\sigma^2}\langle \mathbf{x}, \mathbf{y}\rangle\right) = \sum_{n=0}^{+\infty}\left\{\frac{\langle \mathbf{x}, \mathbf{y}\rangle^n}{\sigma^{2n} \cdot n!}\right\}$$

By the Multinomial Theorem, we have

$$\begin{aligned}\langle \mathbf{x}, \mathbf{y} \rangle^n &= \left(\sum_{i=1}^{d} x_i y_i \right)^n = \sum_{k_1+k_2+\cdots+k_d=n} \left[\binom{n}{k_1, k_2, \ldots, k_d} \prod_{i=1}^{d} (x_i y_i)^{k_i} \right] \\ &= \sum_{k_1+k_2+\cdots+k_d=n} \left[\binom{n}{k_1, k_2, \ldots, k_d}^{\frac{1}{2}} \prod_{i=1}^{d} x_i^{k_i} \cdot \binom{n}{k_1, k_2, \ldots, k_d}^{\frac{1}{2}} \prod_{i=1}^{d} y_i^{k_i} \right]\end{aligned}$$

Therefore,

$$\begin{aligned}K(\mathbf{x}, \mathbf{y}) &= \exp\left(-\frac{\|\mathbf{x}-\mathbf{y}\|^2}{2\sigma^2} \right) = \exp\left(-\frac{\|\mathbf{x}\|^2}{2\sigma^2} \right) \cdot \exp\left(-\frac{\|\mathbf{y}\|^2}{2\sigma^2} \right) \cdot \sum_{n=0}^{+\infty} \left\{ \frac{\langle \mathbf{x}, \mathbf{y} \rangle^n}{\sigma^{2n} \cdot n!} \right\} \\ &= \sum_{n=0}^{+\infty} \frac{\exp\left(-\frac{\|\mathbf{x}\|^2}{2\sigma^2} \right)}{\sigma^n \cdot \sqrt{n!}} \cdot \frac{\exp\left(-\frac{\|\mathbf{y}\|^2}{2\sigma^2} \right)}{\sigma^n \cdot \sqrt{n!}} \cdot \langle \mathbf{x}, \mathbf{y} \rangle^n\end{aligned}$$

Let

$$c_{\sigma,n}(\mathbf{x}) = \frac{\exp\left(-\frac{\|\mathbf{x}\|^2}{2\sigma^2} \right)}{\sigma^n \cdot \sqrt{n!}}, \quad f_{n,\mathbf{k}}(\mathbf{x}) = \binom{n}{k_1, k_2, \ldots, k_d}^{\frac{1}{2}} \prod_{i=1}^{d} x_i^{k_i}$$

then,

$$\begin{aligned}K(\mathbf{x}, \mathbf{y}) &= \sum_{n=0}^{+\infty} \sum_{k_1+k_2+\cdots+k_d=n} c_{\sigma,n}(\mathbf{x}) f_{n,\mathbf{k}}(\mathbf{x}) \cdot c_{\sigma,n}(\mathbf{y}) f_{n,\mathbf{k}}(\mathbf{y}) \\ &= \langle \Phi(\mathbf{x}), \Phi(\mathbf{y}) \rangle\end{aligned}$$

where $\Phi(\mathbf{x})_{\sigma,n,\mathbf{k}} = c_{\sigma,n}(\mathbf{x}) f_{n,\mathbf{k}}(\mathbf{x})$.

$$\begin{aligned}\sum_{i=1}^{n} \sum_{j=1}^{n} c_i c_j K(\mathbf{x}_i, \mathbf{x}_j) &= \sum_{i=1}^{n} \sum_{j=1}^{n} c_i c_j \langle \Phi(\mathbf{x}_i), \Phi(\mathbf{x}_j) \rangle \\ &= \left\langle \sum_{i=1}^{n} c_i \Phi(\mathbf{x}_i), \sum_{i=1}^{n} c_i \Phi(\mathbf{x}_i) \right\rangle \geq 0\end{aligned}$$

for any $x_1, \ldots, x_n \in \mathcal{X}$, given $n \in \mathbb{N}, c_1, \ldots, c_n \in \mathbb{R}$, i.e., $K(\mathbf{x}, \mathbf{y})$ is positive definite.

□

Chapter 34

Minimax Theory

34.1 Fano's Inequality

Let $X \sim P_\theta$, $\theta \in \Theta_0 \subset \Theta$, in which Θ_0 are assumed to be finithe, e.g, $\{\theta_1, \ldots, \theta_M\}$, and θ uniformly distributed on Θ_0. Let $\hat{\theta}$ be an estimator of θ based on X, Then

$$P\left(\theta \neq \hat{\theta}\right) = \frac{1}{M}\sum_1^M P_{\theta_i}\left(\hat{\theta} \neq \theta_i\right) \geq 1 - \frac{I(\theta, X) + \log 2}{\log M}. \tag{34.1}$$

Question: How to upper bound $I(\theta, X)$?

There are various ways to do that, one earlier bond is

$$I(\theta, X) \leq \frac{1}{M^2}\sum_{i=1}^M \sum_{j=1}^M D_{KL}\left(P_{\theta_i} \| \theta_j\right). \tag{34.2}$$

To use such a bound needs to be very careful with the construction of Θ_0. Alternatively, insight from the information theory may provide another method to do it in a way that takes advantage of known metric entropy.

Typically, Θ_0 is a subset of Θ.

Original problem $X \sim P_\theta$, $\theta \in \Theta$. Here θ can be a finite dimension or infinite dimension (e.g., $\theta = f(x)$, pdf of X).

Let $\pi(\theta)$ be a prior distribution on Θ. To apply Fano's Inequality, we need to choose Θ finite. Regardlessly, we want a general bound on $I(\theta; X)$, in which $\theta \sim \pi$.

Recall

$$H(X) = \begin{cases} -\sum_i p_i \log p_i & \text{discrete} \\ -\int p(x) \log p(x)\, \mathrm{d}x & \text{contious} \end{cases} \tag{34.3}$$

Given the Prof P. Shannon achieves (within one bit) the lower bound on the expected code length of any prefix code, with code length $\log \frac{1}{p_i}$ (ignoring rounding up), which is $H(X)$.

If we mistakenly use q to code, then the expected extra bits are (also called redundancy)

$$\sum_i p_i \log \frac{1}{q_i} - \sum_i p_i \log \frac{1}{q_i} = \sum_i p_i \log \frac{p_i}{q_i} \geq 0. \tag{34.4}$$

Bayes misfu?? mis?? the Bayes Redundancy. Let q be any pdf, redundancy at θ is

$$\int f(x,\theta) \log \frac{f(x,\theta)}{q(x)} \,\mathrm{d}x, \tag{34.5}$$

where $f(x,\theta)$ is pdf of X. Bayes redundancy is

$$\begin{aligned} &\int_\Theta \left(\int f(x,\theta) \log \frac{f(x,\theta)}{q(x)} \,\mathrm{d}x \right) \cdot \pi(\theta) \,\mathrm{d}\theta \\ =&\int \int_\Theta f(x,\theta) \log \frac{\pi(\theta) f(x,\theta)}{\pi(\theta) q(x)} \cdot \pi(\theta) \,\mathrm{d}\theta \,\mathrm{d}x, \end{aligned} \tag{34.6}$$

Let $q^*(x) = \int_\Theta f(x,\theta)\pi(\theta) \,\mathrm{d}\theta$,

$$\begin{aligned} =&\int \int_\Theta f(x,\theta) \log \frac{\pi(\theta) f(x,\theta)}{\pi(\theta) q^*(x)} \cdot \pi(\theta) \,\mathrm{d}\theta \,\mathrm{d}x \\ &+ \int \int_\Theta f(x,\theta) \log \frac{q^*(x)}{q(x)} \pi(\theta) \,\mathrm{d}\theta \,\mathrm{d}x, \end{aligned} \tag{34.7}$$

in which, the first part is the Bayes redundancy of q^*, and the second part

$$\begin{aligned} =&\int \log \frac{q^*(x)}{q(x)} \left(\int_\Theta f(x,\theta)\pi(\theta) \,\mathrm{d}\theta \right) \mathrm{d}x \\ =&\int \log \frac{q^*(x)}{q(x)} q^*(x) \,\mathrm{d}x \geq 0. \end{aligned} \tag{34.8}$$

Thus Bayes redundancy of q^* is $I(\theta; X)$.

Our approach is to provide a sensible upper bound on $I(\theta; X)$, that is not specific to the choice of Θ_0. Rather, it reflects the native of the $\{p_\theta, \theta \in \Theta\}$ (or a subset of it) more.

Suppose we have i.i.d observations $X_1, X_2, \dots X_n \sim p_\theta$, let

$$d_k(p,q) = \sqrt{D(p\|q)} = \sqrt{\int p(x) \log \frac{p(x)}{q(x)} \,\mathrm{d}x}, \tag{34.9}$$

which is not a metric. Let G_ε be an ε-cover of the family $\{p_\theta, \theta \in \Theta\}$, i.e.,

$$\forall \theta \in \Theta, \exists \theta \in G_\varepsilon, \text{ s.t. } D(p_\theta \| p_{\theta'}) \leq \varepsilon^2. \tag{34.10}$$

Let $M = |G_\varepsilon|$, $q(x_1, \ldots, x_n) = \frac{1}{M}\sum_{i=1}^M p_{\theta_i}^n$ (centroid) and $p_{\theta_i}^n = p_{\theta_i}(x_1)\ldots p_{\theta_i}(x_n)$. Then,

$$\begin{aligned} D_{\mathrm{KL}}\left(p_\theta^n \| q\right) &= \int p_\theta^n \log \frac{p_\theta^n}{\frac{1}{M}\sum_{i=1}^M p_{\theta_i}^n}\,\mathrm{d}x \\ &\le \int p_\theta^n \log \frac{p_\theta^n}{\frac{1}{M} p_{\theta_i}^n}\,\mathrm{d}x \\ &= \log M + \inf_i \int p_\theta^n \log \frac{p_\theta^n}{p_{\theta_i}^n}\,\mathrm{d}x \\ &= \log M + \inf_i D_{\mathrm{KL}}\left(p_\theta \| p_{\theta_i}\right) \\ &\le \log M + n\varepsilon^2. \end{aligned} \tag{34.11}$$

This holds for all $\theta \in \Theta$. If we have a subset Θ and for any prior π, we have

$$I\left(\theta; X^n\right) \le \log |G_\varepsilon| + n\varepsilon^2. \tag{34.12}$$

$$\begin{aligned} D_{\mathrm{KL}}\left(p_{\theta_i}^n \| q\right) &\le \log M, \quad 1 \le i \le M, \\ D_{\mathrm{KL}}\left(p_\theta^n \| q\right) &\le \log M + n\varepsilon^2, \quad \forall \theta \in \Theta. \end{aligned}$$

If $D_{\mathrm{KL}}\left(p_{\theta_i} \| p_{\theta_j}\right) \ge \eta^2 > 0$, then $D_{\mathrm{KL}}\left(p_{\theta_i}^n \| p_{\theta_j}^n\right) \ge n\eta^2$. But with the centroid density q, we have

$$D_{\mathrm{KL}}\left(p_{\theta_i}^n \| q\right) \le \log M.$$

Then, we have the interesting situation that $D_{\mathrm{KL}}\left(p_{\theta_i}^n \| p_{\theta_j}^n\right)$ are very large for $i \ne j$, yet $D_{\mathrm{KL}}\left(p_{\theta_i}^n \| q\right)$ are small. Relaxing speaking!

34.2 Minimax Rate

Consider a loss function $l(\theta, \theta')$

Theorem 34.2.1

Suppose on a finite set $\Theta_0 \subset \Theta$, we have

$$\min_{\theta_i \neq \theta_j} l(\theta_i, \theta_j) \geq \Delta > 0, \tag{34.13}$$

for any $\theta_i \neq \theta_j \in \Theta_0$ and $\theta \in \Theta$, we have

$$l(\theta_i, \theta) + l(\theta_j, \theta) \geq c\Delta, \tag{34.14}$$

for some constant $c > 0$. Then

$$\inf_{\hat{\theta}} \sup_{\theta \in \Theta} \mathbb{E}_\theta l\left(\theta, \hat{\theta}\right) \geq \frac{c\Delta}{2}\left(1 - \frac{V_k(\varepsilon) + n\varepsilon^2 + \log 2}{\log |\Theta_0|}\right), \tag{34.15}$$

where $V_k(\varepsilon)$ is the covering entropy of $\{p_\theta, \theta \in \Theta\}$ under d_{KL}.

For f, g with different supports, we possibly have $D(f\|g) = \infty$, suppose density are w.r.t a probability measure μ. Given original observations $X_1, \ldots, X_n \sim f \in \mathcal{F}$, let $Y_1, \ldots, Y_n$ be i.i.d uniform w.r.t μ, and $V_1, \ldots, V_n$ be coin flips. Suppose

$$Z_i = \begin{cases} X_i, & \text{if } V_i = 1, \\ Y_i, & \text{if } V_i = 0. \end{cases}$$

Then, $Z_1, \ldots, Z_n \sim \frac{f}{2} + \frac{1}{2}$.

Proof. Let

$$\tilde{\theta} = \arg\min_{\theta_i \in \Theta_0} l(\theta_i, \hat{\theta}).$$

Then, we have $\theta \neq \tilde{\theta}$, we know

$$l(\theta, \hat{\theta}) \geq l(\tilde{\theta}, \hat{\theta}),$$

and

$$l(\theta, \hat{\theta}) + l(\tilde{\theta}, \hat{\theta}) \geq c\Delta.$$

Consequently, $l(\theta, \hat{\theta}) \geq \frac{c\Delta}{2}$. Thus,

$$\inf_{\hat{\theta}} \sup_{\theta \in \Theta} \mathbb{E}_\theta l(\theta, \hat{\theta}) \geq \inf_{\hat{\theta}} \sup_{\theta \in \Theta_0} \mathbb{E}_\theta l(\theta, \hat{\theta}).$$

□

34.3 Applications

1. Consider a fixed design regression with independent observations:

$$Y_i = u(X_i) + e_i, \quad 1 \le i \le n,$$

where $x_i = i/n$, $e_i \sim N(0,1)$ and $\mathcal{U}$ consists of all functions g on $[0,1]$ that are uniformly bounded between $-A$ and A for some positive constant A and $|g(x) - g(y)| \le L|x-y|$ for some constant $L > 0$. The loss function of interest is $\ell(u,\hat{u}) = \int_0^1 (u(x) - \hat{u}(x))^2 \, \mathrm{d}x$. Show that

$$\inf_{\hat{u}} \sup_{u \in \mathcal{U}} E_u \ell(u,\hat{u}) \asymp n^{-2/3}.$$

Note that some results given in the lectures on regression may not be directly applicable because here we deal with a fixed design. You may consider a piecewise constant estimator for upper bounding and you may use the fact that $\mathcal{U}$ has L_2 metric entropy order $1/\epsilon$.

Proof. Since X are fixed design, then $y_i \sim N(u(x_i), 1), i = 1, 2, \ldots, n$, which are independent but no longer iid Let P_u denotes the distribution of Y with regression function u. Thus,

$$D_{\mathrm{KL}}\left(p_u^n \| p_v^n\right) = \frac{1}{2} \sum_{i=1}^{n} \left[u(x_i) - v(x_i)\right]^2,$$

where $p_u^n(\mathbf{y}) = \prod_{i=1}^n p_u(y_i)$, and $p_v^n(\mathbf{y}) = \prod_{i=1}^n p_v(y_i)$.
Lower Bound: □

2. Consider the collection $\mathcal{A} = \{(a,b) : -\infty < a < b < \infty\}$ comprising sets in the real number line. Show its VC dimension is 2.

Proof. Since we have $S_{\mathcal{A}}(n) = \frac{n(n+1)}{2} + 1$, it follows that the VC dimension of $\mathcal{A}$ is 2. □

Chapter 35

Multivariate Extensions

35.1 Applications

35.1.1 Mean vector

Let $\mathbf{X}_i$, $i = 1, \ldots, n$ be drawn from a p-dimensional distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. The sample mean vector is given by

$$\bar{\mathbf{X}} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{X}_i.$$

By the multivariate central limit theorem, we have

$$\sqrt{n}(\bar{\mathbf{X}} - \boldsymbol{\mu}) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}).$$

Applying the continuous mapping theorem and Theorem 16.1.1, we obtain

$$n(\bar{\mathbf{X}} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\bar{\mathbf{X}} - \boldsymbol{\mu}) \sim \chi^2_p.$$

If $\boldsymbol{\Sigma}$ is known, then the confidence region for $\boldsymbol{\mu}$ with confidence level $1 - \alpha$ is given by

$$n(\bar{\mathbf{X}} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\bar{\mathbf{X}} - \boldsymbol{\mu}) \leq C_p,$$

where

$$\int_0^{C_p} \chi^2_p(t)\, \mathrm{d}t = 1 - \alpha.$$

In applications where $\boldsymbol{\Sigma}$ is unknown, we can use the sample covariance matrix $\widehat{\boldsymbol{\Sigma}}$ to replace $\boldsymbol{\Sigma}$, and $\widehat{\boldsymbol{\Sigma}}^{-1}$ is a consistent estimator of $\boldsymbol{\Sigma}^{-1}$. This leads to the following confidence region for $\boldsymbol{\mu}$:

$$n(\bar{\mathbf{X}} - \boldsymbol{\mu})^\top \widehat{\boldsymbol{\Sigma}}^{-1} (\bar{\mathbf{X}} - \boldsymbol{\mu}) \leq C_p,$$

which provides asymptotically valid confidence regions for $\boldsymbol{\mu}$ at level $1 - \alpha$ for any non-singular $\boldsymbol{\Sigma}$ and fixed shape of distribution F.

Proof. To show that

$$\Pr\left(n(\bar{\mathbf{X}}-\boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}(\bar{\mathbf{X}}-\boldsymbol{\mu})\le C_p\right)\to 1-\alpha,\ \text{as } n\to\infty.$$

□

Remark. If $\mathbf{X}_i$, $i=1,\dots,n$ are drawn from a p-dimensional multivariate normal distribution, the exact distribution of $n(\bar{\mathbf{X}}-\boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}(\bar{\mathbf{X}}-\boldsymbol{\mu})$ is a Hotelling's T^2 distribution with parameters p and $n-1$.

We next consider the power of

35.1.2 Difference of two mean vectors

$$T_{mn}=\left[(\boldsymbol{\eta}-\boldsymbol{\xi})-(\bar{\mathbf{Y}}-\bar{\mathbf{X}})\right]^\top\left(\frac{1}{m}\hat{\boldsymbol{\Sigma}}+\frac{1}{n}\hat{\boldsymbol{\Gamma}}\right)^{-1}\left[(\boldsymbol{\eta}-\boldsymbol{\xi})-(\bar{\mathbf{Y}}-\bar{\mathbf{X}})\right]\le C_p,$$

Univariate case:

$$T_{mn}=\frac{(\bar{Y}-\bar{X})/\sqrt{\frac{1}{m}+\frac{1}{n}}}{\sqrt{\left[\sum_i\left(X_i-\bar{X}\right)^2+\sum_j\left(Y_j-\bar{Y}\right)^2\right]/(m+n-2)}}$$

Since

$$\frac{(\bar{Y}-\bar{X})-(\eta-\xi)}{\sqrt{\frac{\sigma^2}{m}+\frac{\gamma^2}{n}}}\sim\mathcal{N}(0,1)$$

and

$$\frac{\sum_i\left(X_i-\bar{X}\right)^2}{\sigma^2}+\frac{\sum_j\left(Y_j-\bar{Y}\right)^2}{\gamma^2}\sim\chi^2_{m+n-2},$$

then, we have

$$T'_{mn}=\frac{\frac{(\bar{Y}-\bar{X})-(\eta-\xi)}{\sqrt{\frac{\sigma^2}{m}+\frac{\gamma^2}{n}}}}{\sqrt{\left(\frac{\sum_i\left(X_i-\bar{X}\right)^2}{\sigma^2}+\frac{\sum_j\left(Y_j-\bar{Y}\right)^2}{\gamma^2}\right)/(m+n-2)}}\sim t_{m+n-2}.$$

Under the null hypothesis, we have

$$T'_{mn}=T_{mn}\cdot\sqrt{\frac{\sum_i\left(X_i-\bar{X}\right)^2+\sum_j\left(Y_j-\bar{Y}\right)^2}{\gamma^2\sum_i\left(X_i-\bar{X}\right)^2+\sigma^2\sum_j\left(Y_j-\bar{Y}\right)^2}}\cdot\sqrt{\frac{(m+n)\sigma^2\gamma^2}{n\sigma^2+m\gamma^2}}.$$

It is easy to see that

$$\sum_i \left(X_i - \bar{X}\right)^2 + \sum_j \left(Y_j - \bar{Y}\right)^2 \to_p m\sigma^2 + n\gamma^2,$$
$$\gamma^2 \sum_i \left(X_i - \bar{X}\right)^2 + \sigma^2 \sum_j \left(Y_j - \bar{Y}\right)^2 \to_p (m+n)\sigma^2\gamma^2 > 0,$$

and by the continuous mapping theorem with the fact that $m/(m+n) \to \rho$, we have

$$\sqrt{\frac{\sum_i \left(X_i - \bar{X}\right)^2 + \sum_j \left(Y_j - \bar{Y}\right)^2}{\gamma^2 \sum_i \left(X_i - \bar{X}\right)^2 + \sigma^2 \sum_j \left(Y_j - \bar{Y}\right)^2}} \cdot \sqrt{\frac{(m+n)\sigma^2\gamma^2}{n\sigma^2 + m\gamma^2}} \to_p \sqrt{\frac{\rho\sigma^2 + (1-\rho)\gamma^2}{(1-\rho)\sigma^2 + \rho\gamma^2}}.$$

Proof.

$$\Pr\left(T_{mn} \le C_k\right) \to \gamma, \text{ as } m, n \to \infty$$

For convenience, suppose $m < n$, then we define

$$\mathbf{Z}_i = \mathbf{X}_i - \sqrt{\frac{m}{n}} Y_i + \frac{m}{\sqrt{mn}}$$

□

35.1.3 Simple Linear Regression

Suppose

$$\mathbf{X}_i = \boldsymbol{\alpha} + \mathbf{v}_i\boldsymbol{\beta} + \boldsymbol{\varepsilon}_i, \quad i = 1, \ldots, n$$

where $\mathbf{v}_i$ is a p-dimensional vector of known constants, and $\boldsymbol{\varepsilon}_i$ is a p-dimensional random vector with mean $\mathbf{0}$ and covariance matrix $\boldsymbol{\Sigma}$. The least squares estimator of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are given by

$$\hat{\beta}_j = \frac{\sum_i (X_{i,j} - \bar{X}_j)(v_{i,j} - \bar{v}_j)}{\sum_i (v_{i,j} - \bar{v}_j)^2}, \quad \hat{\alpha}_j = \bar{X}_j - \hat{\beta}_j \bar{v}_j.$$

where $\bar{\mathbf{X}} = \frac{1}{n}\sum_i \mathbf{X}_i$ and $\bar{\mathbf{v}} = \frac{1}{n}\sum_i \mathbf{v}_i$.

Denote

$$d_{n,j}^{(i)} = \frac{v_{i,j} - \bar{v}_j}{\sqrt{\sum_i (v_{i,j} - \bar{v}_j)^2}}.$$

We can rewrite the least squares estimator as

$$\widehat{\beta}_j - \beta_j = \frac{\sum_i d_{n,j}^{(i)} \left[(X_{i,j} - \bar{X}_j) - \mathbb{E}(X_{i,j} - \bar{X}_j)\right]}{\sqrt{\sum_i (v_{i,j} - \bar{v}_j)^2}},$$

By Theorem, we have

$$\left(\sqrt{\sum_i (v_{i,1}-\bar{v}_1)^2}(\hat{\beta}_1-\beta_1),\ldots,\sqrt{\sum_i (v_{i,p}-\bar{v}_p)^2}(\hat{\beta}_p-\beta_p)\right)^\top \xrightarrow{d} \mathcal{N}(\mathbf{0},\mathbf{\Sigma}).$$

35.1.4 Multinomial One-Sample Test

Consider a sequence of n independent multinomial trials with $k+1$ categories, where the probability of the i-th category is p_i, $i=1,\ldots,k+1$.

Let us now consider testing the hypothesis

$$H_0 : p_i = p_i^{(0)}, \quad i=1,\ldots,k+1$$

against the alternatives that $p_i \neq p_i^{(0)}$ for at least some i. The standard test for this problem is Pearson's χ^2-test, which rejects H_0 when

$$Q = n\sum_{i=1}^{k+1}\left(\frac{Y_i}{n}-p_i^{(0)}\right)^2/p_i^{(0)} \geq C_k,$$

The asymptotic distribution of Q is χ^2_k under H_0.

Proof. It follows from (5.4.14) and Theorem 16.1.1 that

$$n\sum_{i=1}^{k}\sum_{j=1}^{k} a_{ij}\left(\frac{Y_i}{n}-p_i^{(0)}\right)\left(\frac{Y_j}{n}-p_j^{(0)}\right) \xrightarrow{d} \chi^2_k, \tag{35.1}$$

where

$$a_{ij} = \begin{cases} \frac{1}{p_i^{(0)}}+\frac{1}{p_{k+1}^{(0)}} & \text{if } i=j, \\ \frac{1}{p_{k+1}^{(0)}} & \text{if } i\neq j. \end{cases}$$

The left side of (35.1) can be written as

$$n\sum_{i=1}^{k}\frac{1}{p_i^{(0)}}\left(\frac{Y_i}{n}-p_i^{(0)}\right)^2 + \frac{n}{p_{k+1}^{(0)}}\sum_{i=1}^{k}\sum_{j=1}^{k}\left(\frac{Y_i}{n}-p_i^{(0)}\right)\left(\frac{Y_j}{n}-p_j^{(0)}\right),$$

The last term is equal to

$$n\left[\sum_{i=1}^{k}\left(\frac{Y_i}{n}-p_i^{(0)}\right)\right]^2/p_{k+1}^{(0)} = n\left(\frac{Y_{k+1}}{n}-p_{k+1}^{(0)}\right)^2/p_{k+1}^{(0)},$$

and completes the proof. □

35.1.5 Contingency Table

Part XI

Computational Statistics

Chapter 36

Random Generator

Random number generator is a key component in computational statistics. It is used to generate random numbers from a given probability distribution. In this chapter, we will introduce some basic concepts and algorithms of random number generation.

36.1 Uniform Random Number Generation

36.2 Non-uniform Random Number Generation

For non-uniform random number generation

36.2.1 Inversion Method

Theorem 36.2.1

Let X be a random variable with cumulative distribution function $F(x)$, then $F(x)$ is a non-decreasing function and $F(x) \in [0,1]$. Let $U \sim \mathcal{U}(0,1)$ be a random variable with uniform distribution on $(0,1)$, then

$$F^{-1}(U) \sim F. \tag{36.1}$$

Proof. Since $F(x)$ is a non-decreasing function, it is invertible. Let $Y = F^{-1}(U)$, then

$$\Pr(Y \leq y) = \Pr(F^{-1}(U) \leq y) = \Pr(U \leq F(y)) = F(y).$$

□

Example (Normal Distribution)**.** Let $X \sim \mathcal{N}(0,1)$ be a random variable with standard normal distribution, then the cumulative distribution function of X is

$$F(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} \exp\left(-\frac{t^2}{2}\right) \mathrm{d}t. \tag{36.2}$$

Since there is no closed form of the inverse of $F(x)$, we can use the approximation form:

$$F^{-1}(u) \approx t - \frac{a_0 - a_1 t}{1 + b_1 t + b_2 t^2}, \tag{36.3}$$

where $t = \sqrt{-2 \log u}$.

36.2.2 Rejection Sampling Method

The inversion method is a general method to generate random numbers from a given probability distribution. However, it is not always easy to find the inverse of the cumulative distribution function. In this case, we can use the rejection sampling method.

Suppose we want to generate random numbers from a probability distribution $f(x)$, which is not easy to sample from. We can find a proposal distribution $g(x)$, which is easy to sample from and satisfies

$$\exists M > 0, \quad f(x) \leq M g(x). \tag{36.4}$$

Then the rejection sampling method is as follows:

Algorithm 2: Rejection Sampling Method

Input: Proposal distribution $g(x)$, constant M

1 Draw a sample $x \sim g(x)$ and $u \sim \mathcal{U}(0,1)$;

2 **if** $u \leq \frac{f(x)}{Mg(x)}$ **then**

3 Accept x;

4 **else**

5 Reject x and go to step 1;

Output: Sample x

Theorem 36.2.2

The rejection sampling method generates a sample x from the probability distribution $f(x)$.

Proof. Let $I = 1$ if x is accepted and $I = 0$ if x is rejected. Then the probability of accepting x given x is

$$\Pr(I = 1 \mid x) = \Pr(u \leq \frac{f(x)}{Mg(x)}) = \frac{f(x)}{Mg(x)}.$$

Thus, the probability of accepting x is

$$\Pr(x \mid I = 1) = \frac{\Pr(x, I = 1)}{\Pr(I = 1)} = \frac{\Pr(x)\Pr(I = 1 \mid x)}{\int \Pr(x)\Pr(I = 1 \mid x)\,\mathrm{d}x} = \frac{f(x)/M}{\int f(x)/M\,\mathrm{d}x} = f(x).$$

□

36.3 Markov Chain Monte Carlo

Markov Chain Monte Carlo (MCMC) is a class of algorithms for sampling from a probability distribution based on constructing a Markov chain that has the desired distribution as its equilibrium distribution. The state of the chain after several steps is then used as a sample of the desired distribution. The quality of the sample improves as a function of the number of steps.

36.3.1 Metropolis-Hastings Sampling

We want to sample from a distribution $\pi(x)$, where $x \in \mathcal{X}$.

Algorithm 3: Random Walk Metropolis-Hastings Sampling

Input: Initial state $x^{(0)}$, number of iterations N

for $i = 1, \ldots, N$ **do**
- Sample $y \sim \mathcal{N}(x^{(i-1)}, \Sigma)$;
- Compute $\alpha = \min\left\{1, \frac{\pi(y)}{\pi(x^{(i-1)})}\right\}$;
- Sample $u \sim \mathcal{U}(0, 1)$;
- **if** $u < \alpha$ **then**
  - $x^{(i)} = y$;
- **else**
  - $x^{(i)} = x^{(i-1)}$;

Output: Samples $x^{(1)}, \ldots, x^{(N)}$

Random Walk Metropolis-Hastings Sampling

Remark. It is usually difficult to sample from a high-dimensional distribution due to the rejection progressively increasing with the dimensionality.

36.3.2 Gibbs Sampling

Gibbs sampling is a special case of Metropolis-Hastings sampling. It is applicable when the joint distribution is not known explicitly or is difficult to sample from directly, but the conditional distribution of each variable is known and easier to sample from. Gibbs sampling only attempts transitions in the coordinate axis direction, using the conditional distribution of the current point to determine the next step's proposal distribution. All proposal samples are accepted without rejection, resulting in potentially higher efficiency.

Let $\mathbf{x} = (x_1, \ldots, x_d)^\top$ be a random vector with joint distribution $\pi(\mathbf{x})$, and let $\pi(x_i|\mathbf{x}_{-i})$ be the conditional distribution of x_i given $\mathbf{x}_{-i} = (x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_d)^\top$. The Gibbs sampling algorithm is as follows:

Algorithm 4: Gibbs Sampling

Input: Initial state $\mathbf{x}^{(0)}$, number of iterations N, burn-in period B

for $i = 1, \ldots, N$ **do**
  for $j = 1, \ldots, d$ **do**
    Sample $x_j^{(i)} \sim \pi(x_j|x_1^{(i)}, \ldots, x_{j-1}^{(i)}, x_{j+1}^{(i-1)}, \ldots, x_d^{(i-1)})$;
  $\mathbf{x}^{(i)} = (x_1^{(i)}, \ldots, x_d^{(i)})^\top$;

Output: Samples $\mathbf{x}^{(B+1)}, \ldots, \mathbf{x}^{(N)}$

Chapter 37

Monte Carlo Integration

Suppose we want to estimate the expectation of a function $h(x)$ for a probability distribution $\pi(x)$, i.e., we want to estimate

$$\mu = \mathbb{E}_\pi[h(x)] = \int h(x)\pi(x)\,\mathrm{d}x. \tag{37.1}$$

37.1 Monte Carlo Integration

If we can sample from $\pi(x)$, then we can use the Monte Carlo integration method as follows:

$$\hat{\mu} = \frac{1}{N}\sum_{i=1}^{N} h(x_i), \tag{37.2}$$

where $x_i \sim \pi(x), \quad i = 1, \ldots, N$.

37.2 Importance Sampling

Importance sampling is a variance reduction technique that can be used when sampling from a distribution $\pi(x)$ is difficult, but sampling from a distribution $g(x)$ is easy. The idea is to sample from $g(x)$ and then reweight the samples so that they are distributed according to $\pi(x)$.

If the probability distribution $\pi(x)$ is difficult to sample from, we can find a proposal distribution $g(x)$. Then the Importance sampling method is as follows:

Algorithm 5: Importance Sampling Method

Input: Proposal distribution $g(x)$, number of samples N

for $i = 1, \ldots, N$ **do**

  Draw a sample $x_i \sim g(x)$;

  Compute $w_i = \frac{\pi(x_i)}{g(x_i)}$;

Calculate $\hat{\mu} = \frac{1}{N} \sum_{i=1}^{N} w_i h(x_i)$;

Output: Estimate $\hat{\mu}$

Normalized Importance Sampling If we do not know the normalization constant of $\pi(x)$, we can use the normalized importance sampling method as follows:

$$\hat{\mu} = \frac{\sum_{i=1}^{N} w_i h(x_i)}{\sum_{i=1}^{N} w_i}. \tag{37.3}$$

Chapter 38

Bootstrap

Bootstrap is a statistical method for estimating the sampling distribution of an estimator by sampling with replacement from the original sample, most often to derive robust estimates of standard errors and confidence intervals of a population parameter like a mean, median, proportion, odds ratio, correlation coefficient, or regression coefficient.

38.1 Bootstrap Principle

Suppose the i.i.d samples $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$ from an unknown probability distribution F on some probability space $\mathcal{X}$. Let $\hat{\theta}_n$ be an estimator of θ based on the sample $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$, that is,

$$\hat{\theta}_n = s\left(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\right),$$

where $s(\cdot)$ is some algorithm.

The bootstrap principle is to use the sample $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$ to estimate the sampling distribution of $\hat{\theta}_n$.

Nonparametric Bootstrap In the b-th bootstrap replicate, we sample with replacement from the original sample $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$ to get $\left\{\mathbf{x}_1^{*b}, \mathbf{x}_2^{*b}, \ldots, \mathbf{x}_n^{*b}\right\}$, and then compute the bootstrap estimate of $\hat{\theta}_n$ as

$$\hat{\theta}_n^{*b} = s\left(\mathbf{x}_1^{*b}, \mathbf{x}_2^{*b}, \ldots, \mathbf{x}_n^{*b}\right). \tag{38.1}$$

Parametric Bootstrap

Bayesian Bootstrap

Smooth Bootstrap

Block Bootstrap

38.2 Standard Error Estimation

The bootstrap estimate of the standard error of $\hat{\theta}_n$ is

$$\widehat{\text{se}}_{\text{boot}} = \sqrt{\frac{1}{B-1}\sum_{b=1}^{B}\left(\hat{\theta}_n^{*b} - \hat{\theta}_n^{*}\right)^2}, \tag{38.2}$$

where $\hat{\theta}_n^{*b}$ is the b-th bootstrap replicate of $\hat{\theta}_n$ and $\hat{\theta}_n^{*}$ is the bootstrap estimate of $\hat{\theta}_n$, that is,

$$\hat{\theta}_n^{*} = \frac{1}{B}\sum_{b=1}^{B}\hat{\theta}_n^{*b}. \tag{38.3}$$

38.3 Bias Estimation

The bootstrap estimate of the bias of $\hat{\theta}_n$ is

$$\widehat{\text{Bias}}_{\text{boot}} = \hat{\theta}_n^{*} - \hat{\theta}_n. \tag{38.4}$$

Remark. The bias-corrected bootstrap estimate of $\hat{\theta}_n$ is

$$\hat{\theta}_n^{*} = \hat{\theta}_n - \widehat{\text{Bias}}_{\text{boot}} = 2\hat{\theta}_n - \hat{\theta}_n^{*}. \tag{38.5}$$

38.4 Confidence Interval Estimation

Percentile Confidence Interval The $1-\alpha$ percentile confidence interval of $\hat{\theta}_n$ is

$$\left[\hat{\theta}_n^{*}(\alpha/2), \hat{\theta}_n^{*}(1-\alpha/2)\right], \tag{38.6}$$

where $\hat{\theta}_n^{*}(\alpha/2)$ is the $\alpha/2$-th percentile of the bootstrap distribution of $\hat{\theta}_n$ and $\hat{\theta}_n^{*}(1-\alpha/2)$ is the $(1-\alpha/2)$-th percentile of the bootstrap distribution of $\hat{\theta}_n$.

Bootstrap-t Confidence Interval The $1-\alpha$ bootstrap-t confidence interval of $\hat{\theta}_n$ is

$$\left[\hat{\theta}_n - t_{n-1}^{*}(1-\alpha/2)\,\widehat{\text{se}}_{\text{boot}}, \hat{\theta}_n - t_{n-1}^{*}(\alpha/2)\,\widehat{\text{se}}_{\text{boot}}\right], \tag{38.7}$$

where $t_{n-1}^{*}(1-\alpha/2)$ is the $(1-\alpha/2)$-th percentile of the bootstrap distribution of t_{n-1} and $t_{n-1}^{*}(\alpha/2)$ is the $\alpha/2$-th percentile of the bootstrap distribution of t_{n-1}.

Bias-corrected and accelerated (BCa) Confidence Interval The $1-\alpha$ bias-corrected and accelerated (BCa) confidence interval of $\hat{\theta}_n$ is

$$\left[\hat{\theta}_n^*\left(\alpha_{\mathrm{BCa}}/2\right), \hat{\theta}_n^*\left(1-\alpha_{\mathrm{BCa}}/2\right)\right], \tag{38.8}$$

where α_{BCa} is the percentile of the bootstrap distribution of $\hat{\theta}_n$ that satisfies

$$\alpha_{\mathrm{BCa}} = \Phi\left(z_0 + \frac{z_0 + z_\alpha - \widehat{\mathrm{Bias}}_{\mathrm{boot}}}{1-\hat{\mathrm{a}}}\right), \tag{38.9}$$

where $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution, z_0 is the $1-\alpha/2$ percentile of the standard normal distribution, z_α is the $\alpha/2$ percentile of the standard normal distribution, and $\hat{\mathrm{a}}$ is the percentile of the bootstrap distribution of $\hat{\theta}_n$ that satisfies

$$\hat{\mathrm{a}} = \frac{\sum_{b=1}^{B}\left(\hat{\theta}_n^{*b} - \hat{\theta}_n^*\right)^3}{6\left[\sum_{b=1}^{B}\left(\hat{\theta}_n^{*b} - \hat{\theta}_n^*\right)^2\right]^{3/2}}. \tag{38.10}$$

38.5 Hypothesis Testing

One-sample Hypothesis Testing

Two-sample Hypothesis Testing

38.6 Jackknife

Let $\mathbf{x}_{-i}$ be the sample with x_i removed, $\mathbf{x}_{-i} = (x_1, x_2, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n)^\top$, and denote the corresponding value of the statistic of interest as

$$\hat{\theta}_{-i} = s\left(\mathbf{x}_{-i}\right) \tag{38.11}$$

Bias of $\hat{\theta}$ For almost all reasonable and practical estimates, we have

$$\mathrm{Bias}(\hat{\theta}_n) = \mathbb{E}(\hat{\theta}_n) - \theta \to 0, \quad n \to \infty$$

Then, it is reasonable to assume a power series of the type

$$\mathbb{E}(\hat{\theta}_n) = \theta + \frac{a_1}{n} + \frac{a_2}{n^2} + \frac{a_3}{n^3} + \ldots$$

with some coefficients $\{a_k\}$. And we have

$$\mathbb{E}(\hat{\theta}_{-i}) = \mathbb{E}(\hat{\theta}_{n-1}) = \theta + \frac{a_1}{n-1} + \frac{a_2}{(n-1)^2} + \frac{a_3}{(n-1)^3} + \dots$$

For the sake of a smaller variance, we average all such estimates and let

$$\hat{\theta}_{(\cdot)} = \frac{1}{n}\sum_{i=1}^{n} \hat{\theta}_{-i}$$

thus,

$$\mathbb{E}(\hat{\theta}_{(\cdot)}) = \theta + \frac{a_1}{n-1} + \frac{a_2}{(n-1)^2} + \frac{a_3}{(n-1)^3} + \dots$$

Thus, we have

$$(n-1)\mathbb{E}\left[\hat{\theta}_{(\cdot)} - \hat{\theta}_n\right] = \frac{a_1}{n} + \frac{a_2}{n^2} + \frac{a_3}{n^3} + \cdots = \text{Bias}(\hat{\theta})$$

Hence, we can get the jackknife estimate bias for $\hat{\theta}$ be

$$\widehat{\text{Bias}}_{\text{jack}} = (n-1)\left(\hat{\theta}_{(\cdot)} - \hat{\theta}_n\right) \tag{38.12}$$

Remark. It is easy to combine the averaged Jackknife estimator $\hat{\theta}_{-i}$ with the original $\hat{\theta}$, to kill the main term in the bias of $\hat{\theta}$, thus,

$$\begin{aligned}\mathbb{E}\left[n\hat{\theta}_n - (n-1)\hat{\theta}_{(\cdot)}\right] &= [n\theta - (n-1)\theta] + [a_1 - a_1] + \left[\frac{a_2}{n} - \frac{a_2}{n-1}\right] + \dots \\ &= \theta + \frac{a_2}{n(n-1)} + \cdots = \theta + \frac{a_2}{n^2} + O\left(n^{-3}\right)\end{aligned}$$

This removes the bias in the special case that the bias is $O\left(n^{-1}\right)$ and removes it to $O\left(n^{-2}\right)$ in other cases.

Variance of $\hat{\theta}$ The jackknife estimate of variance for $\hat{\theta}$ is

$$\widehat{\text{Var}}_{\text{jack}} = \frac{n-1}{n}\sum_{i=1}^{n}\left(\hat{\theta}_{-i} - \hat{\theta}_{(\cdot)}\right)^2, \quad \text{where } \hat{\theta}_{(\cdot)} = \frac{1}{n}\sum_{i=1}^{n}\hat{\theta}_{-i} \tag{38.13}$$

The jackknife method of estimation can fail if the statistic $\hat{\theta}_{\text{jack}}$ is not smooth. Smoothness implies that relatively small changes to data values will cause only a small change in the statistic.

Example (Sample Mean)**.**

Example (Sample Correlation Coefficient)**.**

Part XII

Regression Analysis

Chapter 39

Linear Regression

Chapter 40

Generalized Linear Model

40.1 Introduction

Suppose the response Y has a distribution in the exponential family

$$f\left(y \mid \theta, \phi\right) = \exp\left[\frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi)\right]$$

with link function g, such that,

$$E\left(Y \mid \mathbf{x}\right) = \mu = g^{-1}(\eta), \quad \eta = \mathbf{x}^{\top}\boldsymbol{\beta} \tag{40.1}$$

where the link function provides the relationship between the linear predictor and the mean of the distribution function. If $\eta = \theta$, the link function is called **canonical link function**.

Remark. A generalized linear model (GLM) is a flexible generalization of ordinary linear regression that allows for the response variable to have an error distribution other than the normal distribution.

Table 40.1: Commonly Used Link Functions

Distribution	Support of Distribution	Link Function $g(\mu)$	Mean Function $g^{-1}(\eta)$
Normal	real:$(-\infty, +\infty)$	μ	η
Bernoulli	integer: $\{0, 1\}$	$\log\left(\frac{\mu}{1-\mu}\right)$	$\frac{1}{1+\exp(-\eta)}$
Poisson	integer: $0, 1, 2, \ldots$	$\log\left(\mu\right)$	$\exp\left(\eta\right)$

Maximum Likelihood Suppose the log-likelihood function be

$$\ell(\boldsymbol{\beta} \mid \mathbf{x}, y) = \log[f(y \mid \theta, \phi)] = \log\left[f\left(y \mid g^{-1}(\eta), \phi\right)\right] \tag{40.2}$$

where g is the canonical link function and $\eta = \mathbf{x}^\top \boldsymbol{\beta}$.

Let

$$U(\boldsymbol{\beta}) = \frac{\partial \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}, \quad A(\boldsymbol{\beta}) = -\frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}' \partial \boldsymbol{\beta}}$$

be the score function and observed information matrix.

If $\hat{\boldsymbol{\beta}}$ is the maximum likelihood estimate, then

$$U\left(\hat{\boldsymbol{\beta}}\right) = \mathbf{0}$$

By the mean value theorem,

$$\begin{aligned} U\left(\hat{\boldsymbol{\beta}}\right) - U(\boldsymbol{\beta}_0) &= \frac{\partial U(\boldsymbol{\beta}^*)}{\partial \boldsymbol{\beta}}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\right) \\ \Rightarrow -U(\boldsymbol{\beta}_0) &= -A(\boldsymbol{\beta}^*)\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\right) \end{aligned}$$

where $\boldsymbol{\beta}^* \in \left[\boldsymbol{\beta}_0, \hat{\boldsymbol{\beta}}\right]$. Thus,

$$\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}_0 + A^{-1}(\boldsymbol{\beta}^*) U(\boldsymbol{\beta}_0)$$

Suppose $\hat{\boldsymbol{\beta}}_t, \hat{\boldsymbol{\beta}}_{t+1}$ be the maximum likelihood estimate at the t-th and (t+1)-th iterations, respectively. Two algorithms can be used to obtain the maximum likelihood estimate $\hat{\boldsymbol{\beta}}$.

1. Newton-Raphson Method:

$$\hat{\boldsymbol{\beta}}_{t+1} = \hat{\boldsymbol{\beta}}_t + A^{-1}\left(\hat{\boldsymbol{\beta}}_t\right) U\left(\hat{\boldsymbol{\beta}}_t\right) \Leftrightarrow A\left(\hat{\boldsymbol{\beta}}_t\right) \hat{\boldsymbol{\beta}}_{t+1} = A\left(\hat{\boldsymbol{\beta}}_t\right) \hat{\boldsymbol{\beta}}_t + U\left(\hat{\boldsymbol{\beta}}_t\right) \tag{40.3}$$

   where

$$U(\boldsymbol{\beta}) = \frac{\partial \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \tag{40.4}$$

   is the score function and

$$A(\boldsymbol{\beta}) = -\frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}' \partial \boldsymbol{\beta}} \tag{40.5}$$

   is the observed information matrix.

2. Fisher Scoring Method:

$$\hat{\boldsymbol{\beta}}_{t+1} = \hat{\boldsymbol{\beta}}_t + I^{-1}\left(\hat{\boldsymbol{\beta}}_t\right) U\left(\hat{\boldsymbol{\beta}}_t\right) \Leftrightarrow I\left(\hat{\boldsymbol{\beta}}_t\right) \hat{\boldsymbol{\beta}}_{t+1} = I\left(\hat{\boldsymbol{\beta}}_t\right) \hat{\boldsymbol{\beta}}_t + U\left(\hat{\boldsymbol{\beta}}_t\right) \tag{40.6}$$

   where $U(\boldsymbol{\beta})$ is the score function and

$$I(\boldsymbol{\beta}) = E[A(\boldsymbol{\beta})] = -E\left[\frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}' \partial \boldsymbol{\beta}}\right] \tag{40.7}$$

   is the Fisher information matrix.

Bayesian Methods

40.2 Binary Data

Suppose

$$Y \sim b(m, \pi), \quad i = 1, 2, \ldots, n \tag{40.8}$$

with link function

$$\eta = g(\pi) = \log\left(\frac{\pi}{1-\pi}\right) = \mathbf{x}^\top \boldsymbol{\beta} \tag{40.9}$$

Remark.

The likelihood function is

$$f(\boldsymbol{\pi} \mid \mathbf{x}, \mathbf{y}) = \prod_{i=1}^{n} \binom{m_i}{y_i} \pi_i^{y_i} (1-\pi_i)^{m_i - y_i} \tag{40.10}$$

and the log-likelihood function is

$$\begin{aligned} \ell(\boldsymbol{\beta}) &= \log\left[f(\boldsymbol{\pi} \mid \mathbf{x}, \mathbf{y})\right] = \sum_{i=1}^{n} \ell_i(\boldsymbol{\beta}) \\ &= \sum_{i=1}^{n} \left\{ \log\left[\binom{m_i}{y_i}\right] + y_i \log(\pi_i) + (m_i - y_i)\log(1-\pi_i) \right\} \\ &= \sum_{i=1}^{n} \left[y_i \log\left(\frac{\pi_i}{1-\pi_i}\right) + m_i \log(1-\pi_i) \right] + \sum_{i=1}^{n} \log\left[\binom{m_i}{y_i}\right] \end{aligned} \tag{40.11}$$

where

$$\pi_i = \frac{\exp\left(\mathbf{x}_i^\top \boldsymbol{\beta}\right)}{1 + \exp\left(\mathbf{x}_i^\top \boldsymbol{\beta}\right)} \tag{40.12}$$

Thus,

$$\begin{aligned} U_r(\boldsymbol{\beta}) &= \sum_{i=1}^{n} (y_i - m_i \pi_i) x_{ir} \\ I_{sr}(\boldsymbol{\beta}) &= \sum_{i=1}^{n} m_i \pi_i (1-\pi_i) x_{is} x_{ir} \end{aligned}$$

40.3 Polytomous Data

Definition 40.3.1 (Polytomous Data)

A response is polytomous if the response of an individual or item in a study is **restricted to one of a fixed set of possible values**.

Remark. There are two types of scales, pure scales and compound scales[1]. For pure scales, there are several types:

1. **Nominal Scale**: a scale used for labeling variables into distinct classifications and does not involve a quantitative value or order.
2. **Ordinal Scale**: a variable measurement scale used to simply depict the order of variables and not the difference between each of the variables.
3. **Interval Scale**: a numerical scale where the order of the variables is known as well as the difference between these variables.

Let the category probabilities given $\mathbf{x}_i$ be

$$\pi_j(\mathbf{x}_i) = P(Y = y_j \mid \mathbf{x} = \mathbf{x}_i) \tag{40.13}$$

and the cumulative probabilities given $\mathbf{x}_i$ be

$$r_j(\mathbf{x}_i) = P\left(Y \leq \sum_{r \leq j} y_r \mid \mathbf{x} = \mathbf{x}_i\right) \tag{40.14}$$

where $i = 1, 2, \ldots, n, \quad j = 1, 2, \ldots, k$.

Here, multinomial distribution is in many ways the most natural distribution to consider in the context of a polytomous response variable. The density function of the multinomial distribution is,

$$P(Y_1 = y_1, \ldots, Y_k = y_k) = \begin{cases} \frac{m!}{y_1! \cdots y_k!} \pi_1^{y_1} \cdot \cdots \cdot \pi_k^{y_k}, & \sum_{i=1}^k y_i = m \\ 0 & \text{otherwise} \end{cases}$$

for non-negative integers $y_1, \ldots, y_k$.

As for the link function, we have

Nominal Scale

$$\pi_j(\mathbf{x}_i) = \frac{\exp[\eta_j(\mathbf{x}_i)]}{\sum_{j=1}^k \exp[\eta_j(\mathbf{x}_i)]} \tag{40.15}$$

where $\eta_j(\mathbf{x}_i) = \eta_j(\mathbf{x}_0) + (\mathbf{x}_i - \mathbf{x}_0)' \boldsymbol{\beta}_j + \alpha_i$.

Ordinal Scale

1. Logistic Scale:

$$\log\left[\frac{r_j(\mathbf{x}_i)}{1 - r_j(\mathbf{x}_i)}\right] = \theta_j - \mathbf{x}_i^\top \boldsymbol{\beta} \tag{40.16}$$

2. Complementary Log-Log Scale:

$$\log\{-\log[1 - r_j(\mathbf{x}_i)]\} = \theta_j - \mathbf{x}_i^\top \boldsymbol{\beta} \tag{40.17}$$

[1] A bivariate response with one response ordinal and the other continuous is an example of compound scales.

Interval Scale Suppose the j-th category exits a cardinal number or score, s_j, where the difference between scores is a measure of distance between or separation of categories.

1.
$$\log\left[\frac{r_j(\mathbf{x}_i)}{1-r_j(\mathbf{x}_i)}\right] = \varsigma_0 + \varsigma_1\left(\frac{s_j+s_{j+1}}{2}\right) - \mathbf{x}_i^\top\boldsymbol{\beta} - \mathbf{x}_i^\top\boldsymbol{\xi}(c_j - \bar{c}) \tag{40.18}$$
where $c_j = \frac{s_j+s_{j+1}}{2}$ or $c_j = \text{logit}\left(\frac{s_j+s_{j+1}}{2}\right)$.

2.
$$\pi_j(\mathbf{x}_i) = \frac{\exp[\eta_j(\mathbf{x}_i)]}{\sum_{j=1}^k \exp[\eta_j(\mathbf{x}_i)]} \tag{40.19}$$
where $\eta_j(\mathbf{x}_i) = \eta_j + \left(\mathbf{x}_i^\top\boldsymbol{\beta}\right)s_j + \alpha_i$.

3.
$$\sum_{j=1}^k \pi_j(\mathbf{x}_i)\, s_j = \mathbf{x}_i\boldsymbol{\beta} \tag{40.20}$$

40.4 Count Data

Departures from the idealized Poisson model are to be expected. Therefore, we avoid the assumption of Poisson variation and assume only that

$$\text{Var}(Y) = \sigma^2 E(Y) \tag{40.21}$$

with link function

$$\log(\mu) = \eta = \mathbf{x}^\top\boldsymbol{\beta} \tag{40.22}$$

where $\mu = E(Y \mid \mathbf{x})$.

For the response in the Poisson distribution, i.e.

$$P(Y = y \mid \mu) = \frac{e^{-\mu}\mu^y}{y!}$$

and the log-likelihood function is

$$\ell(\boldsymbol{\beta}) \propto \sum_{i=1}^n (y_i \log(\mu_i) - \mu_i) \tag{40.23}$$

where $\mu_i = E(Y \mid \mathbf{x} = \mathbf{x}_i)$.

Chapter 41

Quantile Regression

Definition 41.0.1 (Smooth Quantile Loss)

Consider the following smooth quantile loss with a convolution operation:

$$\rho_{\tau,h}(\mu) := \int \rho_\tau(v)\kappa_h(\nu - \mu)\,\mathrm{d}\nu, \tag{41.1}$$

where ρ_τ is the quantile loss function and κ_h is a kernel function with bandwidth h.

Chapter 42

Survival Analysis

42.1 General Formulation

Definition 42.1.1 (Survival Function)

The survival function[a] is defined to be

$$S(t) = P(T > t) = \int_t^\infty f(u)\,\mathrm{d}u = 1 - F(t). \tag{42.1}$$

where t is some specified time, T is a random variable denoting the time of death.

[a] The survival function is the probability that the time of death is later than some specified time t.

Definition 42.1.2 (Lifetime Distribution Function)

The lifetime distribution function is defined to be

$$F(t) = P(T \leq t) \tag{42.2}$$

If F is differentiable then the derivative, which is the density function of the lifetime distribution[a], is defined to be

$$f(t) = F'(t) = \frac{\mathrm{d}}{\mathrm{d}t}F(t) \tag{42.3}$$

[a] The function f is sometimes called the event density; it is the rate of death or failure events per unit time.

Definition 42.1.3 (Hazard Function)

The Hazard function[a] is defined to be

$$\lambda(t) = \lim_{\varepsilon \to 0^+} \left[\frac{P(t \leq T < t + \varepsilon \mid T \geq t)}{\varepsilon} \right] = \frac{f(t)}{S(t)} \tag{42.4}$$

[a]The Hazard function is the event rate at time t conditional on survival until time t or later (that is, $T \geq t$).

Property. The relationship among $\lambda(t)$, $f(t)$, $S(t)$,

1.
$$\lambda(t) = -\frac{\mathrm{d} \log[S(t)]}{\mathrm{d}t} \tag{42.5}$$

2.
$$S(t) = \exp\left[-\int_0^t \lambda(x)\, \mathrm{d}x \right] \tag{42.6}$$

3.
$$f(t) = \lambda(t) \exp\left[-\int_0^t \lambda(x)\, \mathrm{d}x \right] \tag{42.7}$$

Proof. □

Example (Constant Hazards)**.** Suppose

$$\lambda(t) = \lambda \tag{42.8}$$

then

$$S(t) = \exp\left[-\int_0^t \lambda(x)\, \mathrm{d}x \right] = \exp\left[-\int_0^t \lambda\, \mathrm{d}x \right] = \exp(-\lambda t)$$
$$f(t) = \lambda(t) \exp\left[-\int_0^t \lambda(x)\, \mathrm{d}x \right] = \lambda \exp\left[-\int_0^t \lambda\, \mathrm{d}x \right] = \lambda \exp(-\lambda t)$$

which is the exponential distribution.

Example (Bathtub Hazards)**.**

$$\lambda(t) = \alpha t + \frac{\beta}{1 + \gamma t} \tag{42.9}$$

42.2 Estimation of Survival Function

Parametric Approach Suppose $t_1, t_2, \ldots, t_n$ are failure times corresponding to censor indicators $\delta_1, \delta_2, \ldots, \delta_n$. The likelihood function is

$$\begin{aligned} f\left(\boldsymbol{\theta} \mid \mathbf{t}, \boldsymbol{\delta}\right) &= \prod_{i=1}^{n} \left[f\left(t_i\right)\right]^{\delta_i} \left[S\left(t_i\right)\right]^{1-\delta_i} \\ &= \prod_{i=1}^{n} \left(\frac{f\left(t_i\right)}{S\left(t_i\right)}\right)^{\delta_i} S\left(t_i\right) \\ &= \prod_{i=1}^{n} \left[\lambda\left(t_i\right)\right]^{\delta_i} S\left(t_i\right) \end{aligned} \tag{42.10}$$

where $\lambda(t), S(t)$ depends on some parameter θ.

Example. Suppose $\boldsymbol{T}$ have exponential density, that,

$$f(t) = \lambda \mathrm{e}^{-\lambda t}, \quad S(t) = \mathrm{e}^{-\lambda t}$$

Thus,

$$\begin{aligned} \ell(\lambda) &= \log[\ell(\theta)] = \sum_{i=1}^{n} \left[\delta_i \log(\lambda) - \lambda t_i\right] \\ &= \left(\sum_{i=1}^{n} \delta_i\right) \log(\lambda) - \lambda \left(\sum_{i=1}^{n} t_i\right) \end{aligned}$$

Hence,

$$\frac{\partial \ell(\lambda)}{\partial \lambda} = \frac{\sum_{i=1}^{n} \delta_i}{\lambda} - \sum_{i=1}^{n} t_i = 0 \Rightarrow \hat{\lambda} = \frac{\sum_{i=1}^{n} \delta_i}{\sum_{i=1}^{n} t_i}$$

Nonparametric Approach Then, for $t_{(k)} \leq t < t_{(k+1)}$,

$$\begin{aligned} \hat{S}(t) &= \prod_{j=1}^{k} \left(\frac{n_j - d_j}{n_j}\right) \\ &= \left(1 - \frac{d_1}{n_1}\right)\left(1 - \frac{d_2}{n_2}\right) \cdots \left(1 - \frac{d_k}{n_k}\right) \\ &\approx \left[1 - \hat{\lambda}\left(t_1\right)\right]\left[1 - \hat{\lambda}\left(t_2\right)\right] \ldots \left[1 - \hat{\lambda}\left(t_k\right)\right] \end{aligned} \tag{42.11}$$

where $\hat{S}(t)$ is referred to as Kaplan-Meier estimate.

42.3 Proportional Hazards Model

Let $t_1, t_2, \ldots, t_n$ be the failure times associated with censor indicator $\delta_1, \delta_2, \ldots, \delta_n$ and the covariate vectors $\mathbf{x}_i$.

Further, let $t_{(1)} \leq t_{(2)} \leq \cdots \leq t_{(m)}$ be the ordered uncensored failure times corresponding to $\delta_{(j)} = 1, j = 1, 2, \ldots, m$, and $x_{(1)}, x_{(2)}, \ldots, x_{(m)}$ are the associated covariate vectors. Note (j) represents the label for the individual who dies at $t_{(j)}$.

The proportional hazards model specifying the hazard at time t for an individual whose covariate vector is $\mathbf{x}$ is given by

$$\lambda(t) = \lambda_0(t)\mathrm{e}^{\mathbf{x}^\top \boldsymbol{\beta}} \tag{42.12}$$

where $\lambda_0(t)$ is referred to as the baseline hazard function.

The exact likelihood function is

$$\ell\left[\beta, \lambda_0(t)\right] = \prod_{i=1}^{n} \left[\lambda_i\left(t_i\right)\right]^{\delta_i} S\left(t_i\right) \tag{42.13}$$

depends on both the nonparametric function $\lambda_0(t)$ and the parameter $\boldsymbol{\beta}$. Thus, it might be difficult to estimate $\lambda_0(t)$ and $\boldsymbol{\beta}$ simultaneously.

The partial likelihood function is

$$\ell_p(\boldsymbol{\beta}) = \prod_{j=1}^{m} \frac{\mathrm{e}^{\mathbf{x}'_{(j)}\boldsymbol{\beta}}}{\sum_{l \in R\left(t_{(j)}\right)} \mathrm{e}^{\mathbf{x}'_l \boldsymbol{\beta}}} = \prod_{i=1}^{n} \left[\frac{\mathrm{e}^{\mathbf{x}_i^\top \boldsymbol{\beta}}}{\sum_{l \in R(t_i)} \mathrm{e}^{\mathbf{x}'_l \boldsymbol{\beta}}}\right]^{\delta_i} \tag{42.14}$$

where $R(t)$ is the set of individuals who are alive and uncensored at a time just before t_i, which is called the risk set.

Another set of work converts the survival prediction problem into a classification problem by dividing the continuous time-to-event into nonoverlapping intervals

Chapter 43

Nonparametric Regression

43.1 Uniform Stability of Regularized Kernel Model

43.1.1 Introduction

Suppose $\mathcal{X}$ be the input space, $\mathcal{Y}$ be the output space, $\mathcal{D}$ be some (almost) completely unknown probability distribution on $\mathcal{X} \times \mathcal{Y}$. Given the n i.i.d observed data, which sampled from an unknown distribution $\mathcal{D}$, that,

$$S := \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\} \subseteq \mathcal{D} \tag{43.1}$$

and the goal of us is to estimate the functional relationship between $\mathcal{X}$ and $\mathcal{Y}$.

To formalize the problem, we now aim at finding a predictor function f^* among the function space $\mathcal{F} := \{f : \mathcal{X} \to \mathcal{Y}\}$ based on the observed data S, which minimizes the true risk

$$R[f] := \mathbb{E}_{\mathcal{D}}\left[L\left(y, f(\mathbf{x})\right)\right] \tag{43.2}$$

where $L : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ is an arbitrary convex loss function, typically assumed that the smaller $L\left(y, f(\mathbf{x})\right)$ is, the better the approximation of y is. Thus, we are trying to find a predictor f^* with risk close to the optimal risk

$$R^* := \inf\left\{R[f] \mid f : \mathcal{X} \to \mathcal{Y}\right\} \tag{43.3}$$

Finding the predictor function f^* which minimizing the empirical risk

$$R_{\text{emp}}(f) = \frac{1}{n}\sum_{i=1}^{n} L\left(y_i, f(\mathbf{x}_i)\right) \tag{43.4}$$

is a natural thing for us to be trying to do. However, as it is known to all, just minimizing the empirical risk is suicidal, which almost certainly leads to overfitting. Minimizing R_{emp} only makes sense if we simultaneously somehow restrict ourselves

to the $\mathcal{F}$, which are of just the right level of complexity. One way to do this is by explicitly restricting the function space $\mathcal{F}$ to a simple space, as in structural risk minimization, which is to introduce a penalty functional $\Omega[f]$ that somehow measures the complexity of each function $f \in \mathcal{F}$, and to minimize the regularized risk

$$R_{\text{reg}}[f] = R_{\text{emp}}[f] + \Omega[f] \tag{43.5}$$

In this report, we restrict the predictor function $f \in \mathcal{F}$ among the reproducing kernel Hilbert space $\mathcal{H}$, and the regularized risk has the form

$$R_{\text{reg}}[f] = \frac{1}{n}\sum_{i=1}^{n} L\left(y_i, f(\mathbf{x}_i)\right) + \frac{\lambda}{2}\|f\|_{\mathcal{H}}^2 \tag{43.6}$$

thus, we can estimate f^* by soving the following optimization problem

$$\hat{f} = \arg\min_{f\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^{n} L\left(y_i, f(\mathbf{x}_i)\right) + \frac{\lambda}{2}\|f\|_{\mathcal{H}}^2 \tag{43.7}$$

where $\lambda > 0$ is a regularization parameter to reduce the danger of overfitting. Since $L\left(y, f(\mathbf{x})\right)$ is convex in f, the minimizer $\hat{f}$ is uniquely determined and a simple gradient descent algorithm can be used to find $\hat{f}$. So the main focus of this report is to answer a remaining question we want to know, whether the risk $R[\hat{f}]$ is close to the optimal risk R^*, which will influence the stability of our algorithm.

43.1.2 Some Notations and Concepts

Before getting into the formal discussion, we will introduce some notations and concepts.

- $S^{\setminus i} := S \setminus \{(\mathbf{x}_i, y_i)\}$ be the sample where the i-th observation is removed.
- $S^i := S^{\setminus i} \cup \{(\mathbf{x}, y)\}$ be the sample where the i-th observation is replaced by $(\mathbf{x}, y)$.

and let $\hat{f}_{\setminus i}$ be the estimated result based on sample $S^{\setminus i}$, $\hat{f}_i$ based on sample S^i and $\hat{f}$ based on sample S.

To quantify the stability of our algorithm, we will introduce one important concept — **Uniform Stability**.

Definition 43.1.1 (Uniform Stability)

The algorithm is uniformly β-stable with respect to the loss function $L\left(y, f(\mathbf{x})\right)$, if for all samples $S := \{\mathbf{x}_i, y_i\}_{i=1}^n \subset \mathcal{D}$ and $i \in [n]$,

$$\sup_{(x,y)\in\mathcal{D}} \left|L\left(y, \hat{f}(\mathbf{x})\right) - L\left(y, \hat{f}_{\setminus i}(\mathbf{x})\right)\right| \leq \beta \tag{43.8}$$

i.e. the algorithm is stable to remove a single sample at all points.

43.1.3 Uniform Stability of Regularized Kernel Model

Firstly, we will provide an auxiliary lemma.

Lemma 43.1.1 (Convex Functions and Derivatives)

For any differentiable convex function $f : \mathbb{R} \to \mathbb{R}$ and any $a, b \in \mathbb{R}$, we have

$$[f'(a) - f'(b)](a - b) \geq 0 \tag{43.9}$$

Proof. Due to the convexity of f we know that $f(a) + (b - a)f'(a) \leq f(b)$ and, likewise, $f(b) + (a - b)f'(b) \leq f(a)$. Summing up both inequalities and subtracting the terms in $f(a)$ and $f(b)$ proves (43.9). □

Then, we will show that the algorithm we studied in this paper satisfied the definition 43.1.1, and the corresponding value of β can be calculated.

Theorem 43.1.1 (Algorithmic Stability of Risk Minimizers (bousquet2002stability;

The algorithm that minimizes the regularized empirical risk in (43.6) has stability

$$\beta = \frac{2C^2\kappa^2}{n\lambda} \tag{43.10}$$

where κ is bound on $\|k(x, \cdot)\| = \sqrt{k(x, x)}$, $\|\cdot\|$ is the RKHS norm induced by k, and C is a bound on the Lipschitz constant of the loss function $L(y, f(\mathbf{x}))$, which can be viewed as a function of f.

Remark. We can see that the stability of the algorithm depends on the regularization constant via $\frac{1}{\lambda n}$, hence we may be able to afford to choose weaker regularization if the sample size n increases.

Proof. To distinguish between different training sets, we use $R_{\text{reg}}[f, S]$ and $R_{\text{reg}}[f, S^{\setminus i}]$ (and likewise $R_{\text{emp}}[f, S]$) during the remainder of the proof.

Since $\hat{f}$ minimizes $R_{\text{reg}}[f, S]$, that is, the **functional derivative stephane2014lecture** of $R_{\text{reg}}[f, S]$ at $\hat{f}$ vanishes, and so does $R_{\text{reg}}[f, S^{\setminus i}]$ at $\hat{f}_{\setminus i}$,

$$\begin{aligned} \partial_f R_{\text{reg}}\left[\hat{f}, S\right] &= \partial_f R_{\text{emp}}\left[\hat{f}, S\right] + \lambda\hat{f} = 0 \\ \partial_f R_{\text{reg}}\left[\hat{f}_{\setminus i}, S^{\setminus i}\right] &= \partial_f R_{\text{emp}}\left[\hat{f}_{\setminus i}, S^{\setminus i}\right] + \lambda\hat{f}_{\setminus i} = 0 \end{aligned} \tag{43.11}$$

Next, we construct an auxiliary risk function $\tilde{R}[f]$ by

$$\tilde{R}[f] := \left\langle \partial_f R_{\text{emp}}\left[\hat{f}, S\right] - \partial_f R_{\text{emp}}\left[\hat{f}_{\setminus i}, S^{\setminus i}\right], f - \hat{f}_{\setminus i} \right\rangle + \frac{\lambda}{2}\left\|f - \hat{f}_{\setminus i}\right\|_{\mathcal{H}}^2 \tag{43.12}$$

$\tilde{R}[f]$ is a convex function in f (the first term is linear, the second quadratic).

Additionally, by construction, we have

$$\tilde{R}[\hat{f}_{\setminus i}] = 0 \tag{43.13}$$

Furthermore, taking the functional derivative of $\tilde{R}[f]$, that,

$$\partial_f \tilde{R}[f] = \partial_f R_{\text{emp}}\left[\hat{f}, S\right] - \partial_f R_{\text{emp}}\left[\hat{f}_{\setminus i}, S^{\setminus i}\right] + \lambda\left(f - \hat{f}_{\setminus i}\right) = \partial_f R_{\text{emp}}\left[\hat{f}, S\right] + \lambda f \tag{43.14}$$

the functional derivative of $\tilde{R}[f]$ (43.14) vanishes at $f = \hat{f}$ due to (43.11), thus the minimum of $\tilde{R}[f]$ is obtained for $f = \hat{f}$. Therefore, combined with $\tilde{R}[\hat{f}_{\setminus i}] = 0$, we can conclude that $\tilde{R}[\hat{f}] \leq 0$.

To obtain bounds on $\left\|\hat{f} - \hat{f}_{\setminus i}\right\|$, we have to get rid of some of the first terms in $\tilde{R}[f]$, since

$$\begin{aligned}
& n\left\langle \partial_f R_{\text{emp}}\left[\hat{f}, S\right] - \partial_f R_{\text{emp}}\left[\hat{f}_{\setminus i}, S^{\setminus i}\right], \hat{f} - \hat{f}_{\setminus i}\right\rangle \\
=& \sum_{j \neq i}\left[L'\left(y_j, \hat{f}(\mathbf{x}_j)\right) - L'\left(y_j, \hat{f}_{\setminus i}(\mathbf{x}_j)\right)\right]\left[\hat{f}(\mathbf{x}_j) - \hat{f}_{\setminus i}(\mathbf{x}_j)\right] \\
& + L'\left(y_i, \hat{f}(\mathbf{x}_i)\right)\left[\hat{f}(\mathbf{x}_i) - \hat{f}_{\setminus i}(\mathbf{x}_i)\right] \\
\geq & L'\left(y_i, \hat{f}(\mathbf{x}_i)\right)\left[\hat{f}(\mathbf{x}_i) - \hat{f}_{\setminus i}(\mathbf{x}_i)\right]
\end{aligned} \tag{43.15}$$

The first equation is since the functional derivative $\partial_f(f) = k(\mathbf{x}, .)$ and then collecting the common terms between $R_{\text{emp}}\left[\hat{f}, S\right]$ and $R_{\text{emp}}\left[\hat{f}_{\setminus i}, S^{\setminus i}\right]$. And, as for the last inequation, we use Lemma 43.1.1 applied to the loss function $L\left(y, f(\mathbf{x})\right)$ which is a convex function of $f(\mathbf{x})$.

Combine the above result with the fact $\tilde{R}[\hat{f}] \leq 0$, we have

$$\left\langle \partial_f R_{\text{emp}}\left[\hat{f}, S\right] - \partial_f R_{\text{emp}}\left[\hat{f}_{\setminus i}, S^{\setminus i}\right], \hat{f} - \hat{f}_{\setminus i}\right\rangle + \frac{\lambda}{2}\left\|\hat{f} - \hat{f}_{\setminus i}\right\|_{\mathcal{H}}^2 \leq 0 \tag{43.16}$$

thus,

$$L'\left(y_i, \hat{f}(\mathbf{x}_i)\right)\left[\hat{f}(\mathbf{x}_i) - \hat{f}_{\setminus i}(\mathbf{x}_i)\right] + \frac{n\lambda}{2}\left\|\hat{f} - \hat{f}_{\setminus i}\right\|_{\mathcal{H}}^2 \leq 0 \tag{43.17}$$

and by the convexity of loss function $L\left(y, f(\mathbf{x})\right)$,

$$\begin{aligned}
\frac{n\lambda}{2}\left\|\hat{f} - \hat{f}_{\setminus i}\right\|_{\mathcal{H}}^2 \leq & L'\left(y_i, \hat{f}(\mathbf{x}_i)\right)\left[\hat{f}_{\setminus i}(\mathbf{x}_i) - \hat{f}(\mathbf{x}_i)\right] \\
\leq & L\left(y_i, \hat{f}(\mathbf{x}_i)\right) - L\left(y_i, \hat{f}_{\setminus i}(\mathbf{x}_i)\right) \\
\leq & \left|L\left(y_i, \hat{f}(\mathbf{x}_i)\right) - L\left(y_i, \hat{f}_{\setminus i}(\mathbf{x}_i)\right)\right|
\end{aligned} \tag{43.18}$$

By the Cauchy-Schwarz inequality, we can see that, for any $f, f' \in \mathcal{H}$ and any $\mathbf{x} \in \mathcal{X}$,

$$|f(\mathbf{x}) - f'(\mathbf{x})| = |\langle f - f', k(\mathbf{x}, \cdot)\rangle| \leq \|f - f'\|_{\mathcal{H}} \|k(\mathbf{x}, \cdot)\|_{\mathcal{H}} \leq \kappa \|f - f'\|_{\mathcal{H}} \tag{43.19}$$

and since $L(y, f(\mathbf{x}))$ is Lipschitz continuous at $\mathbf{x}_i$, we have

$$\left|L\left(y, \hat{f}(\mathbf{x}_i)\right) - L\left(y, \hat{f}_{\setminus i}(\mathbf{x}_i)\right)\right| \leq C\left|\hat{f}(\mathbf{x}_i) - \hat{f}_{\setminus i}(\mathbf{x}_i)\right| \leq C\kappa\left\|\hat{f} - \hat{f}_{\setminus i}\right\|_{\mathcal{H}} \tag{43.20}$$

Combine equation (43.18) and (43.20), we get

$$\frac{n\lambda}{2}\left\|\hat{f} - \hat{f}_{\setminus i}\right\|_{\mathcal{H}}^2 \leq C\kappa\left\|\hat{f} - \hat{f}_{\setminus i}\right\|_{\mathcal{H}} \tag{43.21}$$

thus,

$$\left\|\hat{f} - \hat{f}_{\setminus i}\right\|_{\mathcal{H}} \leq \frac{2C\kappa}{n\lambda} \tag{43.22}$$

Therefore, by the equation (43.20) for every $\mathbf{x}$, we have

$$\left|L\left(y, \hat{f}(\mathbf{x})\right) - L\left(y, \hat{f}_{\setminus i}(\mathbf{x})\right)\right| \leq C\kappa\left\|\hat{f} - \hat{f}_{\setminus i}\right\|_{\mathcal{H}} \leq \frac{2C^2\kappa^2}{n\lambda} \tag{43.23}$$

□

Within the uniform stability of our algorithm, we will also prove that the β-stable algorithm exhibits uniform convergence of the empirical risk $R_{\text{emp}}[f]$ to the true risk $R[f]$.

Theorem 43.1.2 (McDiarmid's Bound (mcdiarmid1989method))

Suppose $\xi_1, \ldots, \xi_n$ be i.i.d real value random variables and assume that there exists a function $g : \mathbb{R}^n \to \mathbb{R}$ with the property that for all $i \in [n]$ and $c_i > 0$,

$$\sup_{\xi_1, \ldots, \xi_n, \xi_i' \in \xi} |g(\xi_1, \ldots, \xi_n) - g(\xi_1, \ldots, \xi_{i-1}, \xi_i', \xi_{i+1}, \ldots, \xi_n)| \leq c_i \tag{43.24}$$

where ξ_i' is drawn from the same distribution as ξ_i. Then

$$\mathbb{P}\left\{|g(\xi_1, \ldots, \xi_n) - \mathbb{E}[g(\xi_1, \ldots, \xi_n)]| > \varepsilon\right\} \leq 2\exp\left(-\frac{2\varepsilon^2}{\sum_{i=1}^n c_i^2}\right) \tag{43.25}$$

Theorem 43.1.3 (Bousquet and Elisseeff (bousquet2001algorithmic; ofer2011algorit

Assume that we have a β-stable algorithm with the additional requirement that the loss function $L\left(y, f(\mathbf{x})\right) \leq M$ for all $(\mathbf{x}, y) \in \mathcal{D}$ and for all samples $S := \{\mathbf{x}_i, y_i\}_{i=1}^n \subset \mathcal{D}$. Then, for any $n \geq 1$

$$\mathbb{P}\left\{\left|R_{\text{emp}}[\hat{f}, S] - R[\hat{f}]\right| > \varepsilon + 2\beta\right\} \leq 2\exp\left(-\frac{n\varepsilon^2}{2(n\beta + M)^2}\right) \tag{43.26}$$

Proof. Within the i.i.d assumption, we have

$$\mathbb{E}_{S\sim\mathcal{D}}\left[R_{\text{emp}}[\hat{f}]\right] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{S\sim\mathcal{D}}\left[L\left(y_i, \hat{f}(\mathbf{x}_i)\right)\right] = \mathbb{E}_{S\sim\mathcal{D}}\left[L\left(y_i, \hat{f}(\mathbf{x}_i)\right)\right] \tag{43.27}$$

If we replace $(\mathbf{x}_i, y_i)$ by $(\mathbf{x}, y)$, we can get

$$\mathbb{E}_{S\sim\mathcal{D}}\left[R_{\text{emp}}[\hat{f}]\right] = \mathbb{E}_{S,(\mathbf{x},y)\sim\mathcal{D}}\left[L\left(y, \hat{f}_i(\mathbf{x})\right)\right] \tag{43.28}$$

and with the observation that

$$\mathbb{E}_{\mathcal{D}}\left[R[\hat{f}]\right] = \mathbb{E}_{S,(\mathbf{x},y)\sim\mathcal{D}}\left[L\left(y, \hat{f}(\mathbf{x})\right)\right] \tag{43.29}$$

In order to bound on the expected difference between $R_{\text{emp}}[\hat{f}, S]$ and $R[\hat{f}]$, which leads to

$$\begin{aligned}\mathbb{E}_{\mathcal{D}}\left[R_{\text{emp}}[\hat{f}, S] - R[\hat{f}]\right] =& \mathbb{E}_{S,(\mathbf{x},y)\sim\mathcal{D}}\left[L\left(y, \hat{f}_i(\mathbf{x})\right)\right] - \mathbb{E}_{S,(\mathbf{x},y)\sim\mathcal{D}}\left[L\left(y, \hat{f}(\mathbf{x})\right)\right] \\ =& \mathbb{E}_{S,(\mathbf{x},y)\sim\mathcal{D}}\left[L\left(y, \hat{f}_i(\mathbf{x})\right) - L\left(y, \hat{f}(\mathbf{x})\right)\right] \leq 2\beta\end{aligned} \tag{43.30}$$

By the triangle inequality, we have

$$\left|R[\hat{f}] - R[\hat{f}_i]\right| \leq \left|R[\hat{f}] - R[\hat{f}_{\setminus i}]\right| + \left|R[\hat{f}_{\setminus i}] - R[\hat{f}_i]\right| \leq 2\beta \tag{43.31}$$

Also, we have

$$\begin{aligned}\left|R_{\text{emp}}[\hat{f}, S] - R_{\text{emp}}\left(\hat{f}_i, S^i\right)\right| \leq& \frac{1}{n}\sum_{j\neq i}\left|L\left(y_j, \hat{f}(\mathbf{x}_j)\right) - L\left(y_j, \hat{f}_i(\mathbf{x}_j)\right)\right| \\ &+ \frac{1}{n}\left|L\left(y_i, \hat{f}(\mathbf{x}_i)\right) - L\left(y_i, \hat{f}_i(\mathbf{x}_i)\right)\right| \\ \leq& \frac{n-1}{n}2\beta + \frac{2M}{n} \leq 2\beta + \frac{2M}{n}\end{aligned} \tag{43.32}$$

and,

$$\begin{aligned}\left|\left[R_{\text{emp}}[\hat{f}, S] - R[\hat{f}]\right] - \left[R_{\text{emp}}\left(\hat{f}_i, S^i\right) - R[\hat{f}_i]\right]\right| \leq & \left|R_{\text{emp}}[\hat{f}, S] - R_{\text{emp}}\left(\hat{f}_i, S^i\right)\right| \\ & + \left|R[\hat{f}] - R[\hat{f}_i]\right| \\ \leq & 4\beta + \frac{2M}{n}\end{aligned} \tag{43.33}$$

Thus, by the Theorem 43.1.2, we have $c_i = 4\beta + \frac{2M}{n}$, that,

$$\begin{aligned}\mathbb{P}\left\{|R_{\text{emp}}[\hat{f}, S] - R[\hat{f}] - 2\beta| > \varepsilon\right\} \leq & \mathbb{P}\left\{|R_{\text{emp}}[\hat{f}, S] - R[\hat{f}]| > \varepsilon + 2\beta\right\} \\ \leq & 2\exp\left(-\frac{n\varepsilon^2}{2(2n\beta + M)^2}\right)\end{aligned} \tag{43.34}$$

□

Within the above two theorems, we can directly get the following practical consequence.

Corollary 43.1.1 (Uniform Convergence Bounds for RKHS)

The algorithm minimizing the regularized risk $R_{\text{reg}}[f]$, as in (43.6), and with the assumptions of Theorem 43.1.1 and 43.1.3, we obtain

$$\mathbb{P}\left\{\left|R_{\text{emp}}[\hat{f}] - R[\hat{f}]\right| > \varepsilon + 2\beta\right\} \leq 2\exp\left(-\frac{n}{2}\left(\frac{\varepsilon}{M}\right)^2\left(1 + \frac{4}{\lambda M}(C\kappa)^2\right)^{-2}\right) \tag{43.35}$$

where

$$\beta = \frac{2C^2\kappa^2}{n\lambda}$$

Remark. For practical considerations, (43.35) may be very useful, even if the rates are not optimal, since the bound is predictive even for small sample sizes and moderate regularization strength. Still, we expect that the constants

Comments

The idea of the discussion content was inspired by **hofmann2008kernel** review, and the report is organized following **scholkopf2002learning** structure, and so are the main proof ideas.

Chapter 44

High Dimensional Regression Analysis

44.1 Lasso for Linear Regression

If $p > n$, then the least squares estimator is not unique. In this case, we can use the lasso to select a unique estimator. The lasso estimator is defined as

$$\hat{\boldsymbol{\beta}}_{\text{lasso}} := \underset{\boldsymbol{\beta} \in \mathbb{R}^p}{\arg\min} \left\{ \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 / n + \lambda \|\boldsymbol{\beta}\|_1 \right\}, \tag{44.1}$$

where $\lambda > 0$ is a tuning parameter. In addition, the optimization problem (44.1) can be rewritten as

$$\hat{\boldsymbol{\beta}}_{\text{lasso}} := \underset{\boldsymbol{\beta} \in \mathbb{R}^p : \|\boldsymbol{\beta}\|_1 \leq R}{\arg\min} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 / n,$$

with a one-to-one correspondence between λ and R.

44.1.1 Numerical Algorithms

The lasso estimator can be computed by the following numerical algorithms.

Cyclic Coordinate Descent

Cyclic coordinate descent is an iterative algorithm. At each iteration, we update one coordinate of $\boldsymbol{\beta}$ while fixing all other coordinates. The cyclic coordinate descent algorithm for the lasso is given in Algorithm 6.

Algorithm 6: Cyclic Coordinate Descent for the Lasso Estimator

Input: Data $\{(\mathbf{X}_i, y_i)\}_{i=1}^n$, tuning parameter $\lambda > 0$, initial value $\boldsymbol{\beta}^{(0)} \in \mathbb{R}^p$, and tolerance $\epsilon > 0$.

$\mathbf{r} \leftarrow \mathbf{y} - \mathbf{X}\boldsymbol{\beta}^{(0)}$;

while $\left\|\boldsymbol{\beta}^{(t)} - \boldsymbol{\beta}^{(t-1)}\right\|_\infty > \epsilon$ **do**

  for $j = 1, \ldots, p$ **do**

    $\beta_j^{(t+1)} \leftarrow \frac{S_\lambda\left(\mathbf{X}_j^\top \mathbf{r}/n\right)}{\mathbf{X}_j^\top \mathbf{X}_j/n}$;

    $\mathbf{r} \leftarrow \mathbf{r} + \left(\beta_j^{(t+1)} - \beta_j^{(t)}\right)\mathbf{X}_j$;

Output: Estimate $\hat{\boldsymbol{\beta}}_{\text{lasso}}$.

44.1.2 Selection of the Tuning Parameter

K-Fold Cross-Validation

44.2 Theory for the Lasso

Let $\boldsymbol{\beta}^0$ be the true parameter vector, and let $\hat{\boldsymbol{\beta}}$ be the lasso estimator. We assume that the design matrix $\mathbf{X}$ is orthogonal, i.e., $\mathbf{X}^\top\mathbf{X} = \mathbf{I}_p$, and that the noise vector $\boldsymbol{\varepsilon}$ is a random vector with mean zero and covariance matrix $\sigma^2\mathbf{I}_n$. We further denote $S_0 := \left\{j : \beta_j^0 \neq 0\right\}$ as the active set, and $s_0 := |S_0|$ as the cardinality of the active set. We also denote $\phi_0 := \min_{j\in S_0}\left|\beta_j^0\right|$ as the minimum absolute value of the nonzero coefficients.

In this section, we will discuss two important properties of the lasso estimator:

1. Prediction Error:

$$\left\|\mathbf{X}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0\right)\right\|_2^2/n \lesssim \frac{\log p}{n}s_0, \quad \text{with } s_0 = o(n/\log p).$$

2. l_1-Error:

$$\left\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0\right\|_1 \lesssim \sqrt{\frac{\log p}{n}}s_0, \quad \text{with } s_0 = o(\sqrt{n/\log p}).$$

Lemma 44.2.1 (Basic Inequality)

$$\left\|\mathbf{X}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0\right)\right\|_2^2/n + \lambda\left\|\hat{\boldsymbol{\beta}}\right\|_1 \leq 2\boldsymbol{\varepsilon}^\top\mathbf{X}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0\right)/n + \lambda\left\|\boldsymbol{\beta}^0\right\|_1.$$

Proof. Since $\hat{\boldsymbol{\beta}}$ is the minimizer of the objective function in (44.1), we have

$$\left\|\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\right\|_2^2/n + \lambda\left\|\hat{\boldsymbol{\beta}}\right\|_1 \leq \left\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}^0\right\|_2^2/n + \lambda\left\|\boldsymbol{\beta}^0\right\|_1.$$

By rearranging the terms, we have

$$\left\|\mathbf{X}\left(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^0\right)\right\|_2^2/n+\lambda\left\|\hat{\boldsymbol{\beta}}\right\|_1 \leq 2\boldsymbol{\varepsilon}^\top\mathbf{X}\left(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^0\right)/n+\lambda\left\|\boldsymbol{\beta}^0\right\|_1.$$

□

Corollary 44.2.1

Assume that $\hat{\sigma}_j^2 = 1$ for all j and that the compatibility condition holds for S_0, with $\hat{\boldsymbol{\Sigma}}$ normalized in this way. For some $t > 0$, let the regularization parameter by

$$\lambda := 4\hat{\sigma}\sqrt{\frac{t^2+2\log p}{n}},$$

where $\hat{\sigma}^2$ is an estimator of the noise variance σ^2. Then with probability at least $1-\alpha$, where

$$\alpha := 2\exp\left[-t^2/2\right] + \Pr(\hat{\sigma} \leq \sigma),$$

we have

$$\left\|\mathbf{X}\left(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^0\right)\right\|_2^2/n+\lambda\left\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^0\right\|_1 \leq 4\lambda^2 s_0/\phi_0^2.$$

Proof. By Lemma 6.2, for

$$\mathcal{F} = \left\{\max_{1\leq j\leq p} 2|\boldsymbol{\varepsilon}^\top\mathbf{X}^{(j)}|/n \leq \lambda_0\right\},$$

we have for all $t > 0$,

$$\Pr(\mathcal{F}) = 1-2\exp\left[-t^2/2\right], \quad \text{where} \quad \lambda_0 = 2\sigma\sqrt{\frac{t^2+2\log p}{n}}.$$

As in the proof of corollary 6.2, if $\hat{\sigma} > \sigma$, then we have $\lambda \geq 2\lambda_0$, and

$$\Pr(\mathcal{F}\cap\{\hat{\sigma}>\sigma\}) = 1-\Pr(\mathcal{F}^c\cup\{\hat{\sigma}\leq\sigma\}) \geq 1-\Pr(\mathcal{F}^c)-\Pr(\hat{\sigma}\leq\sigma) = 1-\alpha,$$

where $\alpha = 2\exp\left[-t^2/2\right]+\Pr(\hat{\sigma}\leq\sigma)$. And since the compatibility condition holds, according to Theorem 6.1, we have

$$\left\|\mathbf{X}\left(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^0\right)\right\|_2^2/n+\lambda\left\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^0\right\|_1 \leq 4\lambda^2 s_0/\phi_0^2,$$

with probability at least $1-\alpha$. □

44.3 Other Lasso-Type Estimators

44.3.1 Adaptive Lasso

The adaptive lasso estimator is defined as

$$\hat{\boldsymbol{\beta}}_{\text{adaptive}} := \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\arg\min} \left\{ \frac{1}{2n} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda \sum_{j=1}^{p} \frac{|\beta_j|}{\hat{\sigma}_j} \right\},$$

where $\lambda > 0$ is a tuning parameter, and $\hat{\sigma}_j$ is an estimator of σ_j.

44.3.2 Elastic Net

The elastic net estimator is defined as

$$\hat{\boldsymbol{\beta}}_{\text{elastic}} := \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\arg\min} \left\{ \frac{1}{2n} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda_1 \|\boldsymbol{\beta}\|_1 + \lambda_2 \|\boldsymbol{\beta}\|_2^2 \right\},$$

where $\lambda_1, \lambda_2 > 0$ are tuning parameters.

44.3.3 Group Lasso

The group lasso estimator is defined as

$$\hat{\boldsymbol{\beta}}_{\text{group}} := \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\arg\min} \left\{ \frac{1}{2n} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda \sum_{g=1}^{G} \sqrt{p_g} \|\boldsymbol{\beta}_g\|_2 \right\},$$

where $\lambda > 0$ is a tuning parameter, and $\boldsymbol{\beta}_g$ is the subvector of $\boldsymbol{\beta}$ corresponding to the gth group.

44.3.4 Fused Lasso

The fused lasso estimator is defined as

$$\hat{\boldsymbol{\beta}}_{\text{fused}} := \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\arg\min} \left\{ \frac{1}{2n} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda \sum_{j=1}^{p-1} |\beta_{j+1} - \beta_j| \right\},$$

44.4 Nonconvex Penalties

The main drawback of the lasso estimator is that it is a biased estimator. To reduce the bias, we can use nonconvex penalties, such as the smoothly clipped absolute deviation (SCAD) penalty and the minimax concave penalty (MCP).

44.4.1 SCAD

The SCAD penalty is defined as

$$\psi_{\text{SCAD}}(\beta) := \begin{cases} \lambda\beta, & \text{if} \quad |\beta| \leq \lambda, \\ \frac{(a\lambda - |\beta|)_+}{(a-1)\lambda}, & \text{if} \quad \lambda < |\beta| \leq a\lambda, \\ \frac{1}{(a-1)\lambda^2}\left[(a+1)\lambda^2 - 2a\lambda\,|\beta| + \beta^2\right], & \text{if} \quad a\lambda < |\beta|\,, \end{cases}$$

where $a > 2$ is a constant.

44.4.2 MCP

The MCP penalty is defined as

$$\psi_{\text{MCP}}(\beta) := \begin{cases} \lambda\beta, & \text{if} \quad |\beta| \leq \lambda, \\ \frac{|\beta|^2}{2(a-1)}, & \text{if} \quad \lambda < |\beta| \leq a\lambda, \\ \frac{a\lambda|\beta| - \lambda^2/2}{a-1}, & \text{if} \quad a\lambda < |\beta|\,, \end{cases}$$

where $a > 1$ is a constant.

Part XIII

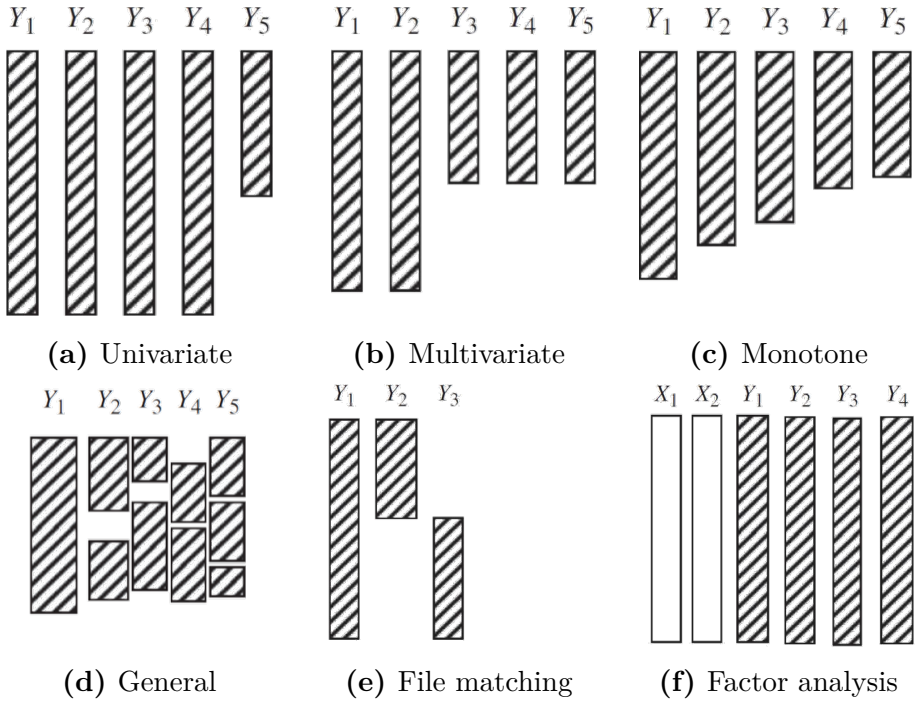
Statistics Applications

Chapter 45

Missingness Data

45.1 The Problem of Missing Data

We are concerned with the problem of the analysis of such a data matrix when some of the entries in the matrix are not observed (Figure 45.1).



(a) Univariate **(b)** Multivariate **(c)** Monotone

(d) General **(e)** File matching **(f)** Factor analysis

Figure 45.1: Examples of missingness patterns

Notations for missing data are as follows

- $Y = (y_{ij})$ denote the $(n \times p)$ rectangular data matrix, where only a portion of Y are observed and $y_{ij} = \star$ indicates this entry is missing;

- $M = (m_{ij})$ denote the *missingness indicator matrix* for y_{ij}, taking $m_{ij} = 0$ for y_{ij} is observed, and $m_{ij} = 1$ for y_{ij} is missing.
- In order to simplify, let $y_i = (y_{i1}, y_{i2}, \ldots, y_{ip})$, $m_i = (m_{i1}, m_{i2}, \ldots, m_{ip})$ and $y_{(0)i}$ be the components of y_i that are observed for unit i, $y_{(1)i}$ be the components of y_i that are missing for unit i.

45.1.1 Missingness Mechanisms

The missingness mechanism is characterized by the conditional distribution of m_i given y_i, say

$$f_{M|Y}\left(m_i \mid y_i, \phi\right), \tag{45.1}$$

where ϕ denotes unknown parameters.

Definition 45.1.1 (Missing Completely at Random, MCAR)

If missingness does not depend on the value of the data, missing or observed, i.e., if for all y_i and any distinct values y^* in the sample space of Y,

$$f_{M|Y}\left(m_i \mid y_i, \phi\right) = f_{M|Y}\left(m_i \mid y^*, \phi\right), \tag{45.2}$$

then the data are called missing completely at random (MCAR).

Definition 45.1.2 (Missing at Random, MAR)

If missingness depends on y_i only through the observed components $y_{(0)}$, i.e., if for all y_i and any distinct values $y^*_{(1)}$ in the sample space of $y_{(1)}$,

$$f_{M|Y}\left(m_i \mid y_{(0)i}, y_{(1)i}, \phi\right) = f_{M|Y}\left(m_i \mid y_{(0)i}, y^*_{(1)}, \phi\right), \tag{45.3}$$

then the data are called missing at random, (MAR).

Definition 45.1.3 (Missing Not at Random, MNAR)

If missingness depends on y_i the missing components $y_{(1)}$, i.e., if some y_i and some values $y^*_{(1)}$ in the sample space of $y_{(1)}$,

$$f_{M|Y}\left(m_i \mid y_{(0)i}, y_{(1)i}, \phi\right) \neq f_{M|Y}\left(m_i \mid y_{(0)i}, y^*_{(1)}, \phi\right), \tag{45.4}$$

then the data are called missing not at random (MNAR).

45.1.2 Commonly Used Methods for Missing Data

1. Complete-case Analysis: discard incompletely recorded units, only use the units with the complete data.
2. Weighting Procedures: randomization inferences from sample survey data without nonresponse commonly weight sampled units by their design weights.
3. Imputation Methods: impute the missing values, and the resultant completed data are analyzed by standard methods.
4. **Model-based Methods**: A broad class of procedures is generated by defining a model for the complete data and basing inferences on the likelihood or posterior distribution under that model, with parameters estimated by procedures such as maximum likelihood.
5. Hybrid Approaches: approaches based on estimating equations have been proposed that combine the aspects of modeling and weighting.

45.2 Likelihood-Based Inference with Missing Data

We can model the density of the joint distribution of Y and M using the "selection model" factorization

$$p(Y = y, M = m \mid \theta, \psi) = f_Y(y \mid \theta) f_{M|Y}(m \mid y, \psi),$$

where θ is the parameter vector governing the data model, and ψ is the parameter vector governing the model for the missingness mechanism.

The full likelihood based on the observed values $\left(y_{(0)}, m\right)$ and the assumed joint distribution model above is defined to be

$$L_{\text{full}}\left(\theta, \psi \mid y_{(0)}, m\right) = \int f_Y\left(y_{(0)}, y_{(1)} \mid \theta\right) f_{M|Y}\left(m \mid y_{(0)}, y_{(1)}, \psi\right) \mathrm{d}y_{(1)} \tag{45.5}$$

The likelihood of θ ignoring the missingness mechanism is defined to be

$$L_{\text{ign}}\left(\theta \mid y_{(0)}\right) = \int f_Y\left(y_{(0)}, y_{(1)} \mid \theta\right) \mathrm{d}y_{(1)} \tag{45.6}$$

45.2.1 Ignorable Missingness Mechanism

Definition 45.2.1 (Ignorable missingness mechanism)

The missingness mechanism is called ignorable if for any given $\tilde{m}$ and $\tilde{y}_{(0)}$ the inferences for θ based on the ignorable likelihood equation evaluated at $m = \tilde{m}$ and $\tilde{y}_0$ are the same as the full likelihood equation.

Remark (Another definition of ignorable missingness mechanism).

$$\frac{L_{\text{full}}\left(\theta, \psi \mid \tilde{y}_{(0)}, \tilde{m}\right)}{L_{\text{full}}\left(\theta^*, \psi \mid \tilde{y}_{(0)}, \tilde{m}\right)} = \frac{L_{\text{ign}}\left(\theta \mid \tilde{y}_{(0)}\right)}{L_{\text{ign}}\left(\theta^* \mid \tilde{y}_{(0)}\right)} \quad \forall \theta, \theta^*, \psi. \tag{45.7}$$

Theorem 45.2.1

The missingness mechanism is ignorable for direct likelihood inference on $\left(\tilde{m}, \tilde{y}_{(0)}\right)$ if

1. Parameter distinctness: The parameters θ and ψ are distinct, that is, $\Omega_{\theta,\psi} = \Omega_\theta \times \Omega_\psi$.
2. Factorization of the full likelihood: The full likelihood, with $(y_0, m) = (\tilde{y}_0, \tilde{m})$ factors as

$$L_{\text{full}}\left(\theta, \psi \mid \tilde{y}_{(0)}, \tilde{m}\right) = L_{\text{ign}}\left(\theta \mid \tilde{y}_{(0)}\right) \times L_{\text{rest}}\left(\psi \mid \tilde{y}_{(0)}, \tilde{m}\right) \quad \forall \theta, \psi \in \Omega_{\theta,\psi} \tag{45.8}$$

Corollary 45.2.1

If the missing data are MAR at $\left(\tilde{m}, \tilde{y}_{(0)}\right)$, and θ and ψ are distinct, the missingness mechanism is ignorable for likelihood inference.

Proof. Since,

$$f_{M|Y}\left(\tilde{m} \mid \tilde{y}_{(0)}, y_{(1)}, \psi\right) = f_{M|Y}\left(\tilde{m} \mid \tilde{y}_{(0)}, y^*_{(1)}, \psi\right) \quad \forall y_{(1)}, y^*_{(1)}, \psi \tag{45.9}$$

therefore,

$$\begin{aligned} f\left(\tilde{y}_{(0)}, \tilde{m} \mid \theta, \psi\right) &= f_{M|Y}\left(\tilde{m} \mid \tilde{y}_{(0)}, \psi\right) \times \int f_Y\left(\tilde{y}_{(0)}, y_{(1)} \mid \theta\right) \mathrm{d}y_{(1)} \\ &= f_{M|Y}\left(\tilde{m} \mid \tilde{y}_{(0)}, \psi\right) \times f_Y\left(\tilde{y}_{(0)} \mid \theta\right) \end{aligned} \tag{45.10}$$

yields the factored likelihood equation 45.8. □

Ignorable Missingness Mechanism v.s. Nonignorable Missingness Mechanism

Example (Exponential Sample)**.** The joint density of n independent and identically distributed scalar units from the exponential distribution with mean $\theta > 0$ is

$$f_Y(y \mid \theta) = \theta^{-n} \exp\left\{-\sum_{i=1}^{n} \frac{y_i}{\theta}\right\}. \tag{45.11}$$

The log-likelihood function is

$$\ell_Y(\theta \mid y) = \ln\left\{\theta^{-n}\exp\left(-\sum_{i=1}^{n}\frac{y_i}{\theta}\right)\right\} = -n\ln\theta - \sum_{i=1}^{n}\frac{y_i}{\theta}. \tag{45.12}$$

Differentiating to θ gives the likelihood equation

$$-\frac{n}{\theta} + \sum_{i=1}^{n}\frac{y_i}{\theta^2} = 0. \tag{45.13}$$

Thus, we obtain the ML estimates

$$\hat{\theta} = \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i. \tag{45.14}$$

Example (Incomplete Exponential Sample)**.** Suppose we have an incomplete univariate exponential sample with $y_{(0)} = (y_1, \ldots, y_r)^\top$ observed and $y_{(1)} = (y_{r+1}, \ldots, y_n)^\top$ missing. Thus, $m = (m_1, \ldots, m_n)^\top$, where $m_i = 0, i = 1, \ldots, r$ and $m_i = 1$, $i = r+1, \ldots, n$.

The likelihood of the ignorable missingness mechanism is

$$L_{\text{ign}}\left(\theta \mid y_{(0)}\right) = \theta^{-r}\exp\left(-\sum_{i=1}^{r}\frac{y_i}{\theta}\right). \tag{45.15}$$

If each unit is observed with probability ψ that does not depend on Y, that is,

$$f_{M|Y}(m \mid y, \psi) = \frac{n!}{r!(n-r)!}\psi^r(1-\psi)^{n-r} \tag{45.16}$$

then,

$$f\left(y_{(0)}, m \mid \theta, \psi\right) = \frac{n!}{r!(n-r)!}\psi^r(1-\psi)^{n-r}\theta^{-r}\exp\left(-\sum_{i=1}^{r}\frac{y_i}{\theta}\right) \tag{45.17}$$

Because the missing data are MAR, if ψ and θ are distinct, then likelihood-based inferences about θ can be based on the ignorable likelihood, the ML estimate of θ is

$$\hat{\theta} = \frac{1}{r}\sum_{i=1}^{r} y_i. \tag{45.18}$$

If each unit is observed only if values less than c, that is

$$f_{M|Y}(m \mid y, \psi) = \prod_{i=1}^{n} f\left(m_i \mid y_i, \psi\right), \tag{45.19}$$

where

$$f\left(m_i \mid y_i, \psi\right) = \begin{cases} 1, & y_i \geq c \\ 0, & \text{otherwise} \end{cases} \tag{45.20}$$

Hence,

$$\begin{aligned} L_{\text{full}}\left(\theta \mid y_{(0)}, m\right) &= \prod_{i=1}^{r} f_Y\left(y_i \mid \theta\right) \Pr\left(y_i < c \mid y_i, \theta\right) \times \prod_{i=r+1}^{n} \Pr\left(y_i \geq c \mid \theta\right) \\ &= \theta^{-r} \exp\left(-\sum_{1}^{r} \frac{y_i}{\theta}\right) \times \exp\left(-\frac{(n-r)c}{\theta}\right) \end{aligned} \tag{45.21}$$

Maximizing the above equation for θ gives the ML estimate

$$\hat{\theta} = \frac{1}{r}\left[\sum_{i=1}^{r} y_i + (n-r)c\right]. \tag{45.22}$$

The inflation of the sample mean in this expression reflects the censoring of the missing values.

45.2.2 Expectation-Maximization Algorithm

Let $\theta^{(i)}$ be the current estimate of the parameter θ. The E step of EM finds the expected complete-data loglikelihood if θ were $\theta^{(t)}$:

$$Q\left(\theta \mid \theta^{(t)}\right) = \int \ell\left(\theta \mid Y_{(0)}, Y_{(1)}\right) f\left(Y_{(1)} \mid Y_{(0)}, \theta^{(t)}\right) \mathrm{d}Y_{(1)}. \tag{45.23}$$

The M step of EM determines $\theta^{(t+1)}$ by maximizing this expected completedata loglikelihood:

$$Q\left(\theta^{(t+1)} \mid \theta^{(t)}\right) \geq Q\left(\theta \mid \theta^{(t)}\right), \quad \forall\theta. \tag{45.24}$$

Hence, the EM algorithm for likelihood-based inference with missing data is

1. replace missing values by estimated values
2. estimate parameters
3. re-estimate the missing values assuming the new parameter estimates are correct
4. re-estimate parameters, and so forth, iterating until apparent convergence

Convergence Properties of EM Algorithm with Missing Data

Theorem 45.2.2

Every GEM algorithm increases $\ell\left(\theta \mid Y_{(0)}\right)$ at each iteration, that is,

$$\ell\left(\theta^{(t+1)} \mid Y_{(0)}\right) \geq \ell\left(\theta^{(t)} \mid Y_{(0)}\right), \tag{45.25}$$

with equality if and only if

$$Q\left(\theta^{(t+1)} \mid \theta^{(t)}\right) = Q\left(\theta^{(t)} \mid \theta^{(t)}\right). \tag{45.26}$$

Proof. The distribution of the complete data Y can be factored as follows:

$$f(Y \mid \theta) = f\left(Y_{(0)}, Y_{(1)} \mid \theta\right) = f\left(Y_{(0)} \mid \theta\right) f\left(Y_{(1)} \mid Y_{(0)}, \theta\right) \tag{45.27}$$

The corresponding decomposition of the log-likelihood is

$$\ell(\theta \mid Y) = \ell\left(\theta \mid Y_{(0)}, Y_{(1)}\right) = \ell\left(\theta \mid Y_{(0)}\right) + \ln f\left(Y_{(1)} \mid Y_{(0)}, \theta\right) \tag{45.28}$$

Let,

$$\ell\left(\theta \mid Y_{(0)}\right) = \ell(\theta \mid Y) - \ln f\left(Y_{(1)} \mid Y_{(0)}, \theta\right) \tag{45.29}$$

The expectation of both sides of the above equation over the distribution of the missing data $Y_{(1)}$, given the observed data $Y_{(0)}$ and a current estimate of θ, say $\theta_{(t)}$, is

$$\ell\left(\theta \mid Y_{(0)}\right) = Q\left(\theta \mid \theta^{(t)}\right) - H\left(\theta \mid \theta^{(t)}\right), \tag{45.30}$$

where

$$\begin{aligned} Q\left(\theta \mid \theta^{(t)}\right) &= \int \ell\left(\theta \mid Y_{(0)}, Y_{(1)}\right) f\left(Y_{(1)} \mid Y_{(0)}, \theta^{(t)}\right) \mathrm{d}Y_{(1)} \\ H\left(\theta \mid \theta^{(t)}\right) &= \int \ln f\left(Y_{(1)} \mid Y_{(0)}, \theta\right) f\left(Y_{(1)} \mid Y_{(0)}, \theta^{(t)}\right) \mathrm{d}Y_{(1)} \end{aligned} \tag{45.31}$$

Since,

$$\begin{aligned}
&H\left(\theta^{(t)}, \theta^{(t)}\right) - H\left(\theta, \theta^{(t)}\right) \\
&= \int \ln f\left(Y_{(1)} \mid Y_{(0)}, \theta^{(t)}\right) f\left(Y_{(1)} \mid Y_{(0)}, \theta^{(t)}\right) \mathrm{d}Y_{(1)} \\
&\quad - \int \ln f\left(Y_{(1)} \mid Y_{(0)}, \theta\right) f\left(Y_{(1)} \mid Y_{(0)}, \theta^{(t)}\right) \mathrm{d}Y_{(1)} \\
&= \int \ln \left[\frac{f\left(Y_{(1)} \mid Y_{(0)}, \theta^{(t)}\right)}{f\left(Y_{(1)} \mid Y_{(0)}, \theta\right)}\right] f\left(Y_{(1)} \mid Y_{(0)}, \theta^{(t)}\right) \mathrm{d}Y_{(1)} \\
&= \int -\ln \left[\frac{f\left(Y_{(1)} \mid Y_{(0)}, \theta\right)}{f\left(Y_{(1)} \mid Y_{(0)}, \theta^{(t)}\right)}\right] f\left(Y_{(1)} \mid Y_{(0)}, \theta^{(t)}\right) \mathrm{d}Y_{(1)} \\
&\geq -\ln \int \frac{f\left(Y_{(1)} \mid Y_{(0)}, \theta\right)}{f\left(Y_{(1)} \mid Y_{(0)}, \theta^{(t)}\right)} f\left(Y_{(1)} \mid Y_{(0)}, \theta^{(t)}\right) \mathrm{d}Y_{(1)} = 0
\end{aligned} \tag{45.32}$$

Therefore,

$$H\left(\theta \mid \theta^{(t)}\right) \leq H\left(\theta^{(t)} \mid \theta^{(t)}\right) \tag{45.33}$$

Hence, the difference in values of $\ell\left(\theta \mid Y_{(0)}\right)$ at successive iterates is given by

$$\begin{aligned}
\ell\left(\theta^{(t+1)} \mid Y_{(0)}\right) - \ell\left(\theta^{(t)} \mid Y_{(0)}\right) &= \left[Q\left(\theta^{(t+1)} \mid \theta^{(t)}\right) - Q\left(\theta^{(t)} \mid \theta^{(t)}\right)\right] \\
&\quad - \left[H\left(\theta^{(t+1)} \mid \theta^{(t)}\right) - H\left(\theta^{(t)} \mid \theta^{(t)}\right)\right] \\
&\geq 0
\end{aligned} \tag{45.34}$$

□

Theorem 45.2.3

Suppose a sequence of EM iterates is such that

1. $D^{10}Q\left(\theta^{(t+1)} \mid \theta^{(t)}\right) = 0$, where D here denotes derivative, and D^{10} means the derivative for the first argument, that is, define

$$D^{10}Q\left(\theta^{(t+1)} \mid \theta^{(t)}\right) = \frac{\partial}{\partial \theta} Q\left(\theta \mid \theta^{(t)}\right)\Big|_{\theta=\theta^{(t+1)}} = 0. \tag{45.35}$$

2. $\theta^{(t)}$ converges to θ^*.
3. $f\left(Y_{(1)} \mid Y_{(0)}, \theta\right)$ is smooth in θ, where smooth is defined in the proof.

Then

$$D\ell\left(\theta^* \mid Y_{(0)}\right) \equiv \frac{\partial}{\partial \theta} \ell\left(\theta \mid Y_{(0)}\right)\Big|_{\theta=\theta^*} = 0, \tag{45.36}$$

so that if the $\theta^{(t)}$ converge, they converge to a stationary point.

Proof.

$$\begin{aligned} D\ell\left(\theta^{(t+1)} \mid Y_{(0)}\right) &= D^{10}Q\left(\theta^{(t+1)} \mid \theta^{(t)}\right) - D^{10}H\left(\theta^{(t+1)} \mid \theta^{(t)}\right) \\ &= -D^{10}H\left(\theta^{(t+1)} \mid \theta^{(t)}\right) \\ &= -\frac{\partial}{\partial\theta}\int\left[\ln f\left(Y_{(1)} \mid Y_{(0)}, \theta\right)\right] f\left(Y_{(1)} \mid Y_{(0)}, \theta^{(t)}\right) dY_{(1)}\Big|_{\theta=\theta^{(t+1)}} \end{aligned} \tag{45.37}$$

which assumes sufficient smoothness to interchange the order of differentiation and integration,

$$\begin{aligned} &= -\int \frac{\partial}{\partial\theta} f\left(Y_{(1)} \mid Y_{(0)}, \theta\right) dY_{(1)}\Big|_{\theta=\theta^{(t+1)}} \\ &= -\int \frac{\partial}{\partial\theta} f\left(Y_{(1)} \mid Y_{(0)}, \theta\right) dY_{(1)}\Big|_{\theta=\theta^{(t+1)}} = 0 \end{aligned} \tag{45.38}$$

□

Examples of EM Algorithm with Missing Data

Example (Multivariate Normal Sample)**.** Let $y = (y_{ij})$, where $i = 1, \ldots, n$, $j = 1, \ldots, p$, be a matrix representing an independent and identically distributed sample of n units from the multivariate normal distribution with mean vector $\mu = (\mu_1, \ldots, \mu_p)$ and covariance matrix $\Sigma = (\sigma_{jk})$, $j = 1, \ldots, p$; $k = 1, \ldots, p$. Thus, y_{ij} represents the value of the jth variable for the ith unit in the sample. The density of y is

$$f_Y(y \mid \mu, \Sigma) = (2\pi)^{-np/2}|\Sigma|^{-n/2} \exp\left\{-\frac{1}{2}\sum_{i=1}^{n}(y_i - \mu)\,\Sigma^{-1}\,(y_i - \mu)^\top\right\}. \tag{45.39}$$

The loglikelihood of $\theta = (\mu, \Sigma)$ is then

$$\ell_Y(\mu, \Sigma \mid y) = -\frac{n}{2}\ln|\Sigma| - \frac{1}{2}\sum_{i=1}^{n}(y_i - \mu)\,\Sigma^{-1}\,(y_i - \mu)^\top \tag{45.40}$$

Maximizing above equation with respect to θ and Σ gives the ML estimate

$$\hat{\mu} = \bar{y}, \quad \hat{\Sigma} = \frac{n-1}{n}S, \tag{45.41}$$

where $\bar{y} = (\bar{y}_1, \ldots, \bar{y}_p)$ is the row vector of sample means, and $S = (s_{jk})$ is the $(p \times p)$ sample covariance matrix with (j, k) th element $s_{jk} = \frac{1}{n-1}\sum_{i=1}^{n}(y_{ij} - \bar{y}_i)\,(y_{ik} - \bar{y}_k)$

Example (Incomplete Multivariate Normal Sample)**.** Suppose $Y = \left(Y_{(0)}, Y_{(1)}\right)$, where Y represents a random sample of size n on $(Y_1, \ldots, Y_p), Y_{(0)}$ the set of observed values, and $Y_{(1)}$ the missing data. Also, let $y_{(0),i}$ represent the set of variables with values observed for unit i, $\quad i = 1, \ldots, n$.

The log-likelihood based on the observed data is then

$$\ell\left(\mu, \Sigma \mid Y_{(0)}\right) = -\frac{1}{2}\sum_{i=1}^{n} \ln\left|\Sigma_{(0),i}\right| - \frac{1}{2}\sum_{i=1}^{n}\left(y_{(0),i} - \mu_{(0),i}\right)^{\top} \Sigma_{(0),i}^{-1}\left(y_{(0),i} - \mu_{(0),i}\right), \tag{45.42}$$

where $\mu_{(0),i}$ and $\Sigma_{(0),i}$ are the mean and covariance matrix of the observed components of Y for unit i.

The exponential family form of multivariate normal distribution with (μ, Σ) is

$$f_Y(y \mid \mu, \Sigma) = (2\pi)^{-np/2}|\Lambda|^{n/2} \exp\left[\eta^T \sum_{i=1}^{n} y_i - \frac{1}{2}\sum_{i=1}^{n} \operatorname{tr}\left(\Lambda y_i y_i^T\right) - \frac{n}{2}\eta^T \Lambda \eta\right], \tag{45.43}$$

where $\Lambda = \Sigma^{-1}$ and $\eta = \Sigma^{-1}\mu$. And

$$\ln f_Y(y \mid \mu, \Sigma) = -\frac{np}{2}\ln(2\pi) + \frac{n}{2}\ln|\Lambda| - \frac{n}{2}\eta^T\Lambda\eta + \eta^T\sum_{i=1}^{n} y_i - \frac{1}{2}\sum_{i=1}^{n}\operatorname{tr}\left(\Lambda y_i y_i^T\right) \tag{45.44}$$

Hence,

$$\begin{aligned} Q\left(\theta \mid \theta^{(t)}\right) &= E_{Y_{(0)},\theta^{(t)}}\left[\ell\left(\theta \mid Y_{(0)}, Y_{(1)}\right)\right] \\ &= -\frac{np}{2}\ln(2\pi) + \frac{n}{2}\ln|\Lambda| - \frac{n}{2}\eta^T\Lambda\eta \\ &\quad + \eta^T E\left(\sum_{i=1}^{n} y_i\right) - \frac{1}{2}\sum_{i=1}^{n}\operatorname{tr}\left(\Lambda E\left(y_i y_i^T\right)\right) \end{aligned} \tag{45.45}$$

Therefore, the EM algorithm for the incomplete multivariate normal sample is,

- E-step:

$$\begin{aligned} E\left(\sum_{i=1}^{n} y_{ij} \mid Y_{(0)}, \theta^{(t)}\right) &= \sum_{i=1}^{n} y_{ij}^{(t+1)}, \quad j = 1, \ldots, p \\ E\left(\sum_{i=1}^{n} y_{ij} y_{ik} \mid Y_{(0)}, \theta^{(t)}\right) &= \sum_{i=1}^{n}\left(y_{ij}^{(t+1)} y_{ik}^{(t+1)} + c_{jki}^{(t+1)}\right), \quad j, k = 1, \ldots, p \end{aligned} \tag{45.46}$$

where

$$\begin{aligned} y_{ij}^{(t+1)} &= \begin{cases} y_{ij}, & \text{if } y_{ij} \text{ is observed} \\ E\left(y_{ij} \mid y_{(0),i}, \theta^{(t)}\right), & \text{if } y_{ij} \text{ is missing} \end{cases} \\ c_{jki}^{(t+1)} &= \begin{cases} 0, & \text{if } y_{ij} \text{ or } y_{ik} \text{ is observed} \\ \operatorname{Cov}\left(y_{ij}, y_{ik} \mid y_{(0),i}, \theta^{(t)}\right), & \text{if } y_{ij} \text{ and } y_{ik} \text{ are missing} \end{cases} \end{aligned} \tag{45.47}$$

- M-step:

$$\begin{aligned}
\mu_j^{(t+1)} &= n^{-1}\sum_{i=1}^{n} y_{ij}^{(t+1)}, \quad j = 1, \ldots, p \\
\sigma_{jk}^{(t+1)} &= n^{-1} E\left(\sum_{i=1}^{n} y_{ij}y_{ik} \mid Y_{(0)}, \theta^{(t)}\right) - \mu_j^{(t+1)}\mu_k^{(t+1)} \\
&= n^{-1}\sum_{i=1}^{n}\left[\left(y_{ij}^{(t+1)} - \mu_j^{(t+1)}\right)\left(y_{ik}^{(t+1)} - \mu_k^{(t+1)}\right) + c_{jki}^{(t+1)}\right], \quad j,k = 1, \ldots, p
\end{aligned} \tag{45.48}$$

Example (Missing Outcomes in Multiple Linear Regression)**.** Suppose a scalar outcome variable Y is regressed on p predictor variables $X_1, \ldots, X_p$, $y_i, i = 1, \ldots, m$ are missing, where

$$\begin{aligned}
E\left(Y \mid X_1, \ldots, X_p\right) &= \beta_0 + \sum_{j=1}^{p}\beta_j X_j \\
\mathrm{Var}\left(Y \mid X_1, \ldots, X_p\right) &= \sigma^2
\end{aligned} \tag{45.49}$$

We assume the joint distribution of the data (including outcomes and predictors) is multivariate normal with

$$\begin{aligned}
\mu &= (\mu_1, \ldots, \mu_p, \mu_y) \\
\Sigma &= \begin{pmatrix} \Sigma_{xx} & \Sigma_{xy} \\ \Sigma_{yx} & \sigma_{yy} \end{pmatrix},
\end{aligned} \tag{45.50}$$

and that the missing data mechanism is ignorable.

Standard regression theory gives

$$\begin{gathered}
\beta = \boldsymbol{\Sigma}_{yx}\boldsymbol{\Sigma}_{xx}^{-1}; \quad \beta_0 = \mu_y - \textstyle\sum_{j=1}^{p}\beta_j\mu_j; \\
\sigma^2 = \sigma_{yy} - \boldsymbol{\Sigma}_{yx}\boldsymbol{\Sigma}_{xx}^{-1}\boldsymbol{\Sigma}_{xy}
\end{gathered} \tag{45.51}$$

The loglikelihood based on the observed data of $\theta = (\beta, \sigma^2)$, where $\beta = (\beta_0, \beta_1, \ldots, \beta_p)$, given observed data $\{(x_i, y_i), i = 1, \ldots, n\}$ is

$$\ell(\beta, \sigma^2 \mid X, Y_{(0)}) = -\frac{n-m}{2}\ln\sigma^2 - \frac{1}{2}\sum_{i=m+1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j\right)^2, \tag{45.52}$$

where only using the observed data.

EM algorithms can be applied to all observations and will obtain iteratively the same ML estimates as would have been obtained noniteratively using only the complete observations.

- E-step:

$$\begin{aligned} E\left(y_i \mid X, Y_{(0)}, \theta^{(t)}\right) &= \begin{cases} y_i, & \text{if } y_i \text{ is observed} \\ \beta^{(t)}\tilde{x}_i^\top & \text{if } y_i \text{ is missing} \end{cases} \\ E\left(y_i^2 \mid X, Y_{(0)}, \theta^{(t)}\right) &= \begin{cases} y_i^2, & \text{if } y_i \text{ is observed} \\ \left(\beta^{(t)}\tilde{x}_i^\top\right)^2 + \sigma^{(t)^2}, & \text{if } y_i \text{ is missing} \end{cases} \end{aligned}, \tag{45.53}$$

  where $\tilde{x}_i = (1, x_i)$.
- M-step:

$$\begin{aligned} \beta^{(t+1)} &= \left(\mathbf{X}^\top X\right)^{-1} \mathbf{X}^\top Y^{(t+1)} \\ \sigma^{(t+1)^2} &= n^{-1}\left[\sum_{i=m+1}^{n}\left(y_i - \beta^{(t)}x_i\right)^2 + m\sigma^{(t)^2}\right], \end{aligned} \tag{45.54}$$

  where $X = (1, X_1, X_2, \ldots, X_p)$

Example (Finite Mixture Linear Regression)**.**

45.3 Missing Not At Random Models

Here, we based on

$$L_{\text{full}}\left(\theta, \psi \mid Y_{(0)}, X, M\right) \propto f\left(Y_{(0)}, M \mid X, \theta, \psi\right) \tag{45.55}$$

regarded as a function of the parameters θ, ψ for fixed observed data $Y_{(0)}$ and missingness pattern M; here $f\left(Y_{(0)}, M \mid X, \theta, \psi\right)$ is obtained by integrating $Y_{(1)}$ out of the joint density $f(Y, M \mid X, \theta, \psi)$ based on a joint model for Y and M given X.

The EM algorithm has the following form for MNAR selection models are as followed,

- E-step:

$$\begin{aligned} Q\left(\theta, \psi \mid \theta^{(t)}, \psi^{(t)}\right) = \int \ell\left(\theta, \psi \mid X, Y_{(0)}, Y_{(1)}, M\right) \\ \cdot f\left(Y_{(1)} \mid X, Y_{(0)}, M, \theta = \theta^{(t)}, \psi = \psi^{(t)}\right) \mathrm{d}Y_{(1)} \end{aligned} \tag{45.56}$$

- M-step:

$$Q\left(\theta^{(t+1)}, \psi^{(t+1)} \mid \theta^{(t)}, \psi^{(t)}\right) \geq Q\left(\theta, \psi \mid \theta^{(t)}, \psi^{(t)}\right) \quad \forall \theta, \psi \tag{45.57}$$

45.3.1 Normal Models for MNAR Missing Data

1. Follow up a sample of nonrespondents and incorporate this information into the main analysis.
2. Adopt a Bayesian approach, assigning the parameters prior distributions. Bayesian inference does not generally require that the data provide information for all the parameters, although inferences tend to be sensitive to the choice of prior distribution.
3. Impose additional restrictions on model parameters.
4. Conduct analysis to assess the sensitivity of inferences for quantities of interest to different choices of the values of parameters poorly estimated from the data.
5. Selectively discard data to avoid modeling the missingness mechanism.

Chapter 46

Treatment-effects Analysis

46.1 Evaluations

46.1.1 Average Treatment Effect

Definition 46.1.1 (Average Treatment Effect)

$$\mathbb{E}\left(Y_1 - Y_2\right) \tag{46.1}$$

46.1.2 Mann-Whitney Statistic

Definition 46.1.2 (Mann-Whitney Statistic)

$$\Pr\left(Y_1 < Y_2\right) \tag{46.2}$$

46.1.3 Distribution-type Index

Definition 46.1.3 (Distribution-type Index)

$$F(x) := \Pr\left(Y_1 - Y_2 = x\right) \tag{46.3}$$

Chapter 47

Graphical Lasso

Before deriving the logarithmic likelihood function, we first introduce the concept of sample covariance matrix, inverse covariance matrix, and probability density function of multivariate normal distribution. Let X be a p-dimensional random vector with mean μ and covariance matrix Σ, then it follows multivariate normal distribution (also known as normal distribution): $X \sim N(\mu, \Sigma)$. Its probability density function is:

$$f_X(x) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)\right),$$

where $|\Sigma|$ represents the determinant of the covariance matrix. The sample covariance matrix S can be calculated from sample data, and its specific expression is:

$$S = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x})^\top$$

where $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ represents the sample mean. The determinant $|\Theta|$ of the inverse covariance matrix Θ represents the independence between variables. L1 regularization can make it equal to zero, so that sparse estimation results can be obtained. The inverse covariance matrix is the precision matrix of a specific multivariate normal distribution, and it is defined as: $\Theta = (\theta_{ij}) = \Sigma^{-1}$. The derivation of the logarithmic likelihood function models the covariance matrix and

inverse covariance matrix. The likelihood function is as follows:

$$\begin{aligned}
L(\Theta|x) &= \log P(x|\Theta) \\
&= \log \prod_{i=1}^{n} P(x_i|\Theta) \\
&= \sum_{i=1}^{n} \log P(x_i|\Theta) \\
&= \sum_{i=1}^{n} \left(-\frac{p}{2} \log(2\pi) - \frac{1}{2} \log|\Sigma| - \frac{1}{2}(x_i - \mu)^\top \Sigma^{-1}(x_i - \mu) \right) \\
&= -\frac{np}{2} \log 2\pi - \frac{n}{2} \log|\Sigma| - \frac{1}{2} \sum_{i=1}^{n} (x_i - \mu)^\top \Sigma^{-1}(x_i - \mu) \\
&= -\frac{np}{2} \log 2\pi - \frac{n}{2} \log|\Theta| - \frac{1}{2} \sum_{i=1}^{n} (x_i - \mu)^\top \Theta(x_i - \mu)
\end{aligned}$$

Taking the negative value of $L(\Theta|x)$ gives the logarithmic likelihood function, which is:

$$\begin{aligned}
\log L(\Theta|x) &= -L(\Theta|x) \\
&= \frac{np}{2} \log 2\pi + \frac{n}{2} \log|\Theta| + \frac{1}{2} \sum_{i=1}^{n} (x_i - \mu)^\top \Theta(x_i - \mu)
\end{aligned}$$

where μ is the sample mean and Σ is obtained from Θ. The Graphical Lasso loss function obtained by regularization combines this logarithmic likelihood term with the L1 regularization term.

Chapter 48

Semi-supervised Learning

48.1 Assumptions

1. **Smoothness assumption**:
2. **Low-density assumption**:
3. **Manifold assumption**:

Remark. For regression problems, the low-density assumption does not hold, since the decision boundary does not exist.

VAT (Virtual Adversarial Training):

$$\mathbb{E}_{(\mathbf{x},y)\sim p_l(\mathbf{x},y)}\ell(\mathbf{x},y)+\lambda\cdot\mathbb{E}_{\mathbf{x}\sim p_u(\mathbf{x})}\mathbb{E}_{\tilde{\mathbf{x}}\sim q(\tilde{\mathbf{x}}|\mathbf{x})}D_{\mathrm{KL}}(p_\theta(y\mid\mathbf{x})\|p_\theta(y\mid\tilde{\mathbf{x}}))$$

Covariance shift may be a point?

Part XIV

Machine Learning

Chapter 49

Support Vector Machine

Theorem 49.0.1

The minimizer of

$$f^*(X) = \arg\min_f \mathbb{E}\left\{[1 - f(X)Y]_+ \mid X = x\right\}$$

is the sign of $f(x) = \log \frac{p(x)}{1-p(x)}$, i.e., $\operatorname{sgn}\left[p(x) - \frac{1}{2}\right]$.

Proof. For the hinge loss function,

$$\begin{aligned}
&E\left\{[1 - Yg(X)]_+ \mid X = x\right\} \\
&= [1 - g(x)]_+ P(Y = 1 \mid X = x) + [1 + g(x)]_+ P(Y = -1 \mid X = x) \\
&= [1 - g(x)]_+ p(x) + [1 + g(x)]_+ [1 - p(x)] \\
&= \begin{cases} [1 - g(x)]\, p(x), & g(x) < -1 \\ 1 + [1 - 2p(x)]\, g(x), & -1 \le g(x) \le 1 \\ [1 + g(x)]\, [1 - p(x)], & g(x) > 1 \end{cases}
\end{aligned}$$

When $g(x) < -1$,

$$\arg\min_g E\left\{[1 - Yg(X)]_+ \mid X = x\right\} = \arg\min_g [1 - g(x)]\, p(x) = -1$$

When $g(x) > 1$,

$$\arg\min_g E\left\{[1 - Yg(X)]_+ \mid X = x\right\} = \arg\min_g [1 + g(x)]\, [1 - p(x)] = 1$$

When $-1 \leq g(x) \leq 1$,

$$\begin{aligned}
&\arg\min_g E\left\{[1 - Yg(X)]_+ \mid X = x\right\} \\
=&\arg\min_g \{1 + [1 - 2p(x)]\, g(x)\} \\
=&\begin{cases} -1, & p(x) < \frac{1}{2} \\ 0, & p(x) = \frac{1}{2} \\ 1, & p(x) > \frac{1}{2} \end{cases}
\end{aligned}$$

Thus, for the $g(x) \in [-1, 1]$ the minimizer of $\arg\min_g E\left\{[1 - Yg(X)]_+ \mid X = x\right\}$ is the sign of $p(x) - \frac{1}{2}$, that is the sign of $f(x) = \log \frac{p(x)}{1-p(x)}$ □

Chapter 50

Linear Discriminant Analysis

Chapter 51

K-Nearest Neighbor

Chapter 52

Decision Tree

Chapter 53

Kalman Filter

Part XV

Causal Inference

Chapter 54

Causal Inference

A fundamental aspect of causal inference is the precise representation of the study design-that is, the method by which data are collected-which directly influences the validity of the causal conclusions. Typically, each study design is characterized by three components: the population, the regime (Observational vs. Experimental), and the sampling method. As schematically illustrated in Figure 1, these elements provide the framework for categorizing causal inference tasks into four main types, each presenting its own set of distinct challenges and corresponding solutions.

Policy Evaluation and Confounding Bias Suppose we aim to predict the outcome distribution after intervening on X, i.e., $Q = P(Y = y \mid do(X = x))$, we consider an observational study where X, Y, Z, and W are measured from a random sample. The challenge is to identify Q from $P(y, x, z, w)$ by adequately controlling for the confounding effects of Z and W.

Instrumental Variables and Experimental Data To estimate the causal effect of $do(X = x)$, suppose that the available data come from an **experimental study**, in which the variable Z-being easier to manipulate than X-is randomized. The key question is: **Under what conditions can the randomization of variable Z be utilized to infer the causal effect of X?** Formally, our objective is to infer $P(Y = y \mid do(X = x))$ from the distribution $P(y, x, w \mid \text{do}(Z = z))$.

Randomized Trials with Non-Representative Samples In the preceding tasks, we assumed that data were obtained via a perfectly random sample from the underlying population. In practice, however, such ideal sampling conditions rarely hold. For instance, in randomized clinical trials (RCTs), the study cohort may only represent a specific subset of the broader population. Suppose that the experimental data are available in the form $P(y, z, w \mid \text{do}(X = x), S = 1)$ and

possibly $P(y, x, z, w \mid S = 1)$, where S denotes the sample selection indicator (with $S = 1$ indicating inclusion in the sample). The central challenge is therefore: **How can we reliably infer causal effects when the data arise under non-ideal, imperfect sampling conditions?**

Extrapolation Bias and Transportability In previous tasks, we assumed that the population from which data were collected coincides with the target population for inference. In practice, however, this assumption may not hold. For instance, biological experiments might employ animal models as stand-ins for humans, or experimental data may originate from studies conducted in the past, when the characteristics of the target population could have been different. Suppose our goal is to estimate the causal effect in a target population defined by $S = s$, namely:

$$P(Y = y \mid \text{do}(X = x), S = s).$$

Yet, the data at our disposal consist solely of the experimental distribution $P(y, z, w \mid \text{do}(X = x))$ and the observational distribution $P(y, x, z, w \mid S = s)$. This scenario poses a central challenge: **How can we effectively combine these data sources to infer the causal effects within the target population?**

These problems are classified as: Confounding Bias, Sample Selection Bias, and Transportability Bias, respectively. Each study design leads to different data structures, which naturally result in different causal inference conclusions.

54.1 Causal Models

54.1.1 Structural Causal Models

54.1.2 Causal Bayesian Networks

Definition 54.1.1 (Bayesian Networks)

A Bayesian network $M = (G, P)$ over a set of variables $\mathbf{V} = \{V_1, \ldots, V_n\}$ is defined as a joint probability distribution $P(\mathbf{V})$ that factorizes according to a directed acyclic graph (DAG) G:

$$P(V_1, \ldots, V_n) = \prod_{i=1}^{n} P(V_i | \text{Pa}_{V_i}), \tag{54.1}$$

where Pa_{V_i} denotes the set of parents of V_i in G.

Remark. A Bayesian network is termed **causal** if the graph G accurately encodes the causal relationships among the variables. Specifically, for any intervention

$\text{do}(X = x)$ (with $X \subseteq \mathbf{V}$), the resulting distribution is determined by the truncated factorization formula:

$$P(v|\text{do}(x)) = \begin{cases} \prod_{i:v_i \notin x} P(v_i|\text{Pa}_{V_i}), & \text{if} v \text{is consistent with} x, \\ 0, & \text{otherwise.} \end{cases} \tag{54.2}$$

Interventions Interventions represent deliberate modifications to the data-generating process, thereby altering the joint distribution.

Definition 54.1.2 (Hard Intervention)

A **hard intervention** $\text{do}(X = x)$ forces the variables in X to assume specific values, resulting in a new distribution:

$$P(\mathbf{V}|\text{do}(X = x)) = P(\mathbf{V}_x), \tag{54.3}$$

where $\mathbf{V}_x$ denotes the set of variables under the intervention.

Definition 54.1.3 (Soft Intervention)

A **soft intervention** replaces the conditional probability distribution for a variable V_i with a new distribution $P'(V_i|Pa_i^*)$, potentially altering its set of parents Pa_i^* (while avoiding causal cycles). The updated distribution under a soft intervention is given by:

$$P(v; \sigma_{v'}) = \prod_{i:v_i \in v'} P'(v_i|pa_{v_i}^*) \prod_{i:v_i \notin v'} P(v_i|\text{pa}_{v_i}). \tag{54.4}$$

For instance, the intervention $\sigma_Y = P'(y|x)$ would be incompatible with a causal structure $Y \to X$ as it would create a cycle.

In our study, which focuses on learning causal models and structures, we primarily consider **local interventions**-a subset of soft interventions that remain compatible with any causal structure.

Definition 54.1.4 (Local Intervention)

A local intervention σ on $V_i \in \mathbf{V}$ modifies only the state of V_i without conditioning on other endogenous variables. Formally, this is represented as $v_i \mapsto f(v_i)$, denoted by $\sigma = \text{do}(V_i = f(v_i))$, and alters the conditional probability distribution as:

$$P(v_i|\text{pa}_i; \sigma) = \sum_{v_i':f(v_i')=v_i} P(v_i'|\text{pa}_i). \tag{54.5}$$

- **Hard interventions** $\text{do}(V_i = v_i')$ represent local interventions where $f(v_i)$ is a constant function.
- **Translations**, as in $\text{do}(V_i = v_i + k)$, are local interventions where the function is defined by $f(v_i) = v_i + k$. Examples include shifting object positions in reinforcement learning environments (Shah et al., 2022) or translating images (Engstrom et al., 2019).
- **Logical NOT operations**, such as mapping $X \mapsto \neg X$ for a Boolean variable X, are also considered local interventions.

We additionally address **stochastic interventions**, which involve mixtures of local interventions and do not require prior knowledge of the graph G.

For example, adding noise to a variable:

$$X = X + \epsilon, \quad \epsilon \sim \mathcal{N}(0, 1),$$

constitutes a soft intervention on X that can be interpreted as a mixture of local interventions (translations).

Definition 54.1.5 (Mixtures of Intervention)

A mixed intervention $\sigma^* = \sum_i p_i \sigma_i$ applies each intervention σ_i with probability p_i (where $\sum_i p_i = 1$), leading to the composite distribution:

$$P(v|\sigma^*) = \sum_i p_i P(v|\sigma_i).$$

54.2 Decision Tasks

Decision tasks involve an agent selecting a policy to optimize an objective function (utility). To model these tasks causally, **Causal Influence Diagrams (CIDs)** extend **Causal Bayesian Networks (CBNs)** by adding **decision nodes** and **utility nodes**. (Howard & Matheson, 2005; Everitt et al., 2021)

A **single-decision, single-utility CID** consists of three types of variables:

- **Decision variables** D (what decisions are made)
- **Utility variables** U (real-valued function defining the objective)
- **Chance variables** C (environment variables governed by causal mechanisms)

The **agent selects a policy** $\pi(d|pa_D)$ to **maximize expected utility**:

$$E_\pi[U] = E[U|do(D = \pi(pa_D))]$$

An **optimal policy** π^* maximizes this utility:

$$E_{\pi^*}[U]$$

However, real-world agents often incur **regret** δ, defined as the **difference between the optimal and actual expected utility**:

$$\delta := E_{\pi^*}[U] - E_\pi[U]$$

To simplify analysis, the focus is on **unmediated decision tasks**, where **the agent's decision does not causally influence the environment**:

$$\text{Desc}_D \cap \text{Anc}_U = \emptyset$$

Examples include **classification and regression**, while **Markov Decision Processes (MDPs)** are excluded since the agent's action affects the environment, influencing utility.

Part XVI

Deep Learning

Chapter 55

Transformers

55.1 Scaled Dot-Product Attention

The attention mechanism is a fundamental component of the Transformer architecture, enabling each token to incorporate contextual information from the entire sequence. The most widely adopted form is the **scaled dot-product attention**.

Given an input sequence $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ with $\mathbf{x}_i \in \mathbb{R}^{d_{\text{model}}}$, each token is projected into three vectors:

$$\mathbf{q}_i = \mathbf{x}_i W^Q, \quad \mathbf{k}_i = \mathbf{x}_i W^K, \quad \mathbf{v}_i = \mathbf{x}_i W^V,$$

where $W^Q, W^K, W^V \in \mathbb{R}^{d_{\text{model}} \times d_k}$ are learnable parameter matrices. The attention score between the i-th and j-th tokens is defined as:

$$\text{score}_{ij} = \frac{\mathbf{q}_i^\top \mathbf{k}_j}{\sqrt{d_k}},$$

where the scores are normalized using the softmax function to obtain attention weights:

$$\alpha_{ij} = \frac{\exp\left(\frac{\mathbf{q}_i^\top \mathbf{k}_j}{\sqrt{d_k}}\right)}{\sum_{l=1}^{n} \exp\left(\frac{\mathbf{q}_i^\top \mathbf{k}_l}{\sqrt{d_k}}\right)}.$$

The output for the i-th token is computed as a weighted sum of the value vectors:

$$\text{Attention}(\mathbf{q}_i, K, V) = \sum_{j=1}^{n} \alpha_{ij} \cdot \mathbf{v}_j.$$

In matrix form, let $Q, K, V \in \mathbb{R}^{n \times d_k}$ represent the matrices of all queries, keys, and values, respectively. The attention output is given by:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right) V.$$

In tasks requiring restricted access (e.g., causal decoding or padding), a mask matrix $M \in \mathbb{R}^{n \times n}$ can be added prior to the softmax operation:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}} + M\right)V.$$

To enhance representational power, the Transformer employs **multi-head attention**, which allows the model to attend to information from multiple representation subspaces:

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \ldots, \text{head}_h)W^O. \tag{55.1}$$

Each attention head is computed independently:

$$\text{head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V), \tag{55.2}$$

where $W_i^Q, W_i^K, W_i^V \in \mathbb{R}^{d_{\text{model}} \times d_k}$ are projection matrices specific to head i, and $W^O \in \mathbb{R}^{hd_v \times d_{\text{model}}}$ is the final output projection.

This mechanism enables the model to dynamically integrate context across the sequence and has proven essential for the success of Transformer-based architectures in natural language processing and beyond.

55.2 Key-Value Caching

Transformer-based autoregressive language models, such as GPT, generate text sequentially by predicting each token conditioned on all previously generated tokens. At each generation step t, the model takes the sequence $\{x_1, x_2, \ldots, x_{t-1}\}$ as input and outputs the next token x_t. A computational challenge arises in this setting: without optimization, the model recomputes representations of all previous tokens at every step, leading to redundant and inefficient computations, especially for long sequences.

To address this inefficiency, **Key-Value (KV) Caching** is employed during the inference phase. This technique caches the *key* and *value* vectors generated by the self-attention mechanism for all previously seen tokens. Recall that in a Transformer, the self-attention operation is defined as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V,$$

where Q, K, and V denote the query, key, and value matrices, respectively, and d_k is the dimension of the key vectors.

During inference, the key and value vectors for past tokens are stored in the KV cache, and only the query vector for the current token needs to be computed.

The attention mechanism then retrieves the necessary context by attending to the cached key-value pairs, eliminating the need to recompute past representations. Furthermore, the current token's key and value vectors are appended to the cache for use in subsequent steps.

KV caching significantly reduces the computational complexity of sequence generation and is a core optimization enabling real-time response in large language models. It is especially crucial for applications involving long-context generation, such as open-ended dialogue, document summarization, and code synthesis.

Chapter 56

Mixture of Experts

Part XVII

Generative Models

Chapter 57

Diffusion Model

57.1 Introduction

57.1.1 Denoising Diffusion Probabilistic Model

ho2020denoising

57.2 Score Matching

The most difficult question in Langevin dynamics is how to obtain $\nabla_{\mathbf{x}} \log p(\mathbf{x})$, because we have no access to the true data distribution $p(\mathbf{x})$.

$$\mathcal{J}_{\mathrm{ESM}}(\boldsymbol{\theta}) := \frac{1}{2}\mathbb{E}_{p(\mathbf{x})}\left[\|s_{\boldsymbol{\theta}}(\mathbf{x}) - \nabla_{\mathbf{x}} \log p(\mathbf{x})\|_2^2\right]. \tag{57.1}$$

Explicit Score Matching

Implicit Score Matching

Denoising Score Matching

$$\mathcal{J}_{\mathrm{DSM}}(\boldsymbol{\theta}) := \frac{1}{2}\mathbb{E}_{p(\mathbf{x}',\mathbf{x})}\left[\|s_{\boldsymbol{\theta}}(\mathbf{x}') - \nabla_{\mathbf{x}} \log p(\mathbf{x}'|\mathbf{x})\|_2^2\right]. \tag{57.2}$$

Theorem 57.2.1

$$\mathcal{J}_{\mathrm{DSM}}(\boldsymbol{\theta}) = \mathcal{J}_{\mathrm{ESM}}(\boldsymbol{\theta}) + C, \tag{57.3}$$

where C is a constant that does not depend on $\boldsymbol{\theta}$.

Proof.

$$\begin{aligned}\mathcal{J}_{\mathrm{ESM}}(\boldsymbol{\theta}) &= \frac{1}{2}\mathbb{E}_{p(\mathbf{x})}\left[\|s_{\boldsymbol{\theta}}(\mathbf{x}) - \nabla_{\mathbf{x}}\log p(\mathbf{x})\|_2^2\right] \\ &= \mathbb{E}_{p(\mathbf{x})}\Big[\frac{1}{2}\|s_{\boldsymbol{\theta}}(\mathbf{x})\|_2^2 - \langle s_{\boldsymbol{\theta}}(\mathbf{x}), \nabla_{\mathbf{x}}\log p(\mathbf{x})\rangle + \underbrace{\frac{1}{2}\|\nabla_{\mathbf{x}}\log p(\mathbf{x})\|_2^2}_{:=C_1,\text{independent of }\boldsymbol{\theta}}\Big]\end{aligned}$$

Let's zoom into the second term, we can show that

$$\begin{aligned}\mathbb{E}_{p(\mathbf{x})}\left[\langle s_{\boldsymbol{\theta}}(\mathbf{x}), \nabla_{\mathbf{x}}\log p(\mathbf{x})\rangle\right] &= \int \left(s_{\boldsymbol{\theta}}(\mathbf{x})^\top \nabla_{\mathbf{x}}\log p(\mathbf{x})\right) p(\mathbf{x})\,\mathrm{d}\mathbf{x} \\ &= \int s_{\boldsymbol{\theta}}(\mathbf{x})^\top \frac{\nabla_{\mathbf{x}} p(\mathbf{x})}{\cancel{p(\mathbf{x})}}\cancel{p(\mathbf{x})}\,\mathrm{d}\mathbf{x} \\ &= \int s_{\boldsymbol{\theta}}(\mathbf{x})^\top \nabla_{\mathbf{x}}\left(\int p(\mathbf{x}\mid\mathbf{x}')p(\mathbf{x}')\,\mathrm{d}\mathbf{x}'\right)\mathrm{d}\mathbf{x} \\ &= \int s_{\boldsymbol{\theta}}(\mathbf{x})^\top \left(\int \nabla_{\mathbf{x}} p(\mathbf{x}\mid\mathbf{x}')p(\mathbf{x}')\,\mathrm{d}\mathbf{x}'\right)\mathrm{d}\mathbf{x} \\ &= \int s_{\boldsymbol{\theta}}(\mathbf{x})^\top \left(\int \nabla_{\mathbf{x}} p(\mathbf{x}\mid\mathbf{x}')p(\mathbf{x}') \times \frac{p(\mathbf{x}\mid\mathbf{x}')}{p(\mathbf{x}\mid\mathbf{x}')}\,\mathrm{d}\mathbf{x}'\right)\mathrm{d}\mathbf{x} \\ &= \int s_{\boldsymbol{\theta}}(\mathbf{x})^\top \left(\int \left(\nabla_{\mathbf{x}}\log p(\mathbf{x}\mid\mathbf{x}')\right) p(\mathbf{x}\mid\mathbf{x}')p(\mathbf{x}')\,\mathrm{d}\mathbf{x}'\right)\mathrm{d}\mathbf{x} \\ &= \mathbb{E}_{p(\mathbf{x},\mathbf{x}')}\left[\langle s_{\boldsymbol{\theta}}(\mathbf{x}), \nabla_{\mathbf{x}}\log p(\mathbf{x}\mid\mathbf{x}')\rangle\right].\end{aligned}$$

So if we substitute this back to the original equation, we have

$$\begin{aligned}\mathcal{J}_{\mathrm{ESM}}(\boldsymbol{\theta}) =& \mathbb{E}_{p(\mathbf{x},\mathbf{x}')}\left[\frac{1}{2}\|s_{\boldsymbol{\theta}}(\mathbf{x}) - \nabla_{\mathbf{x}}\log p(\mathbf{x}\mid\mathbf{x}')\|_2^2\right] + C_1 \\ =& \mathbb{E}_{p(\mathbf{x},\mathbf{x}')}\left[\frac{1}{2}\|s_{\boldsymbol{\theta}}(\mathbf{x}) - \nabla_{\mathbf{x}}\log p(\mathbf{x}\mid\mathbf{x}')\|_2^2 + \frac{1}{2}\|\nabla_{\mathbf{x}}\log p(\mathbf{x}\mid\mathbf{x}')\|_2^2\right] \\ &+ C_1 - \underbrace{\mathbb{E}_{p(\mathbf{x},\mathbf{x}')}\left[\frac{1}{2}\|\nabla_{\mathbf{x}}\log p(\mathbf{x}\mid\mathbf{x}')\|_2^2\right]}_{:=C_2,\text{independent of }\boldsymbol{\theta}} \\ =& \mathcal{J}_{\mathrm{DSM}}(\boldsymbol{\theta}) + C_1 - C_2,\end{aligned}$$

which completes the proof. □

In the special case where $p(\mathbf{x}\mid\mathbf{x}') = \mathcal{N}(\mathbf{x}\mid\mathbf{x}', \sigma^2\mathbf{I})$, we can let $\mathbf{x} = \mathbf{x}' + \sigma\boldsymbol{\varepsilon}$, where $\boldsymbol{\varepsilon}\sim\mathcal{N}(0,\mathbf{I})$, then we have

$$\nabla_{\mathbf{x}}\log p(\mathbf{x}\mid\mathbf{x}') = -\frac{\mathbf{x}-\mathbf{x}'}{\sigma^2} = -\frac{\boldsymbol{\varepsilon}}{\sigma}.$$

As a result, we have

$$\mathcal{J}_{\mathrm{DSM}}(\boldsymbol{\theta}) = \mathbb{E}_{p(\mathbf{x}')}[\|s_{\boldsymbol{\theta}}(\mathbf{x}' + \sigma\boldsymbol{\varepsilon}) + \boldsymbol{\varepsilon}/\sigma\|_2^2].$$

If we replace the dummy variable $\mathbf{x}'$ with $\mathbf{x}$, we have

$$\mathcal{J}_{\mathrm{DSM}}(\boldsymbol{\theta}) = \mathbb{E}_{p(\mathbf{x})}[\|s_{\boldsymbol{\theta}}(\mathbf{x} + \sigma\boldsymbol{\varepsilon}) + \boldsymbol{\varepsilon}/\sigma\|_2^2].$$

Example (SDEs for DDPM)**.** For $i = 1, \ldots, N$, we have

$$\mathbf{x}_i = \sqrt{(1-\beta_i)}\mathbf{x}_{i-1} + \sqrt{\beta_i}\mathbf{z}_i, \text{where } \mathbf{z}_i \sim \mathcal{N}(0, \mathbf{I}).$$

which can be rewritten as an SDE via

$$\mathrm{d}\mathbf{x} = -\frac{1}{2}\beta(t)\mathbf{x}\,\mathrm{d}t + \sqrt{\beta(t)}\,\mathrm{d}\mathbf{w}(t).$$

Proof. We define a step size $\Delta t = \frac{1}{N}$, and consider an auxiliary noise level $\{\tilde{\beta}_i\}_{i=1}^N$ such that $\beta_i = \tilde{\beta}_i \frac{1}{N}$. Then we have

$$\beta_i = \beta\left(\frac{i}{N}\right)\frac{1}{N} = \beta(t + \Delta t)\Delta t.$$

here we assume that as $N \to \infty$, $\tilde{\beta}_i$ converges to $\beta(t)$, which is a continuous function of t for all $t \in [0, 1]$. Similarly, we have

$$\mathbf{x}_i = \mathbf{x}\left(\frac{i}{N}\right) = \mathbf{x}(t + \Delta t), \quad \mathbf{z}_i = \mathbf{z}\left(\frac{i}{N}\right) = \mathbf{z}(t + \Delta t).$$

Then we have

$$\begin{aligned}
&\mathbf{x}_i = \sqrt{1-\beta_i}\mathbf{x}_{i-1} + \sqrt{\beta_i}\mathbf{z}_i \\
&\mathbf{x}(t+\Delta t) = \sqrt{1-\beta(t+\Delta t)\Delta t}\cdot\mathbf{x}(t) + \sqrt{\beta(t+\Delta t)\Delta t}\cdot\mathbf{z}(t+\Delta t) \\
\Rightarrow\quad &\approx \left(1 - \frac{1}{2}\beta(t)\Delta t\right)\mathbf{x}(t) + \sqrt{\beta(t)\Delta t}\cdot\mathbf{z}(t) \\
\Rightarrow\ &\mathrm{d}\mathbf{x} = -\frac{1}{2}\beta(t)\mathbf{x}\,\mathrm{d}t + \sqrt{\beta(t)}\,\mathrm{d}\mathbf{w}(t),
\end{aligned}$$

The approximation arises from the Taylor expansion of $\sqrt{1-\beta(t+\Delta t)\Delta t} \approx 1 - \frac{1}{2}\beta(t)\Delta t$ as $\Delta t \to 0$. For the final expression, the first term is straightforward, and the second term follows from the fact that $\mathbf{w}(t)$ is a Wiener process, and $\sqrt{\Delta t}\mathbf{z}(t) = \mathbf{z}(t + \Delta t) - \mathbf{z}(t)$, thus, $\frac{\sqrt{\Delta t}\mathbf{z}(t)}{\Delta t}$ converges to $\mathrm{d}\mathbf{w}(t)$ as $\Delta t \to 0$. This completes the proof. □

57.3 Classifier and Classifier-Free Guidance

According to the Bayes rule, we have

$$p(\mathbf{x}_t \mid y) = \frac{p(y \mid \mathbf{x}_t)p(\mathbf{x}_t)}{p(y)}. \tag{57.4}$$

Thus,

$$\log p(\mathbf{x}_t \mid y) = \log p(y \mid \mathbf{x}_t) + \log p(\mathbf{x}_t) - \log p(y).$$

Take derivative with respect to $\mathbf{x}_t$ on both sides, we have

$$\nabla_{\mathbf{x}_t} \log p(\mathbf{x}_t \mid y) = \nabla_{\mathbf{x}_t} \log p(\mathbf{x}_t) + \nabla_{\mathbf{x}_t} \log p(y \mid \mathbf{x}_t).$$

Classifier-Free Guidance

57.4 Effort in Inference

Due to the sequential nature of the sampling process, the diffusion model is computationally expensive. Effors have been made to accelerate this process, by resorting to:

Higher-orde Numerical Schemes **dockhorn2022genie**

High Resolution Image Generation lower dimensional latent space (**rombach2022higha**)