9. Bayesian Inference  
  9.1. Conjugate Priors  
  9.2. Bayesian Point and Interval Estimation  
  9.3. Bernstein-von Mises theorem  
  9.4. Proof of Theorem 9.7  
10. Elementary Decision Theory  
  10.1. Risk: the frequentist approaches  
  10.2. Bayes’ risk  
  10.3. Bayes’ rule and admissibility  
  10.4. Admissibility and inadmissibility in Gaussian location models  
11. Nonparametric estimation  
  11.1. Kernel density estimator  
  11.2. Regression: projection estimators  
  11.3. Gaussian nonparametric models and their asymptotic equivalence  
  11.4. Minimax lower bound  
References
1. Prelude

1.1. The Likelihood Function. In a probability problem, we typically have a random variable \( Y \) whose distribution has been specified in some way, and wish to make statements about various probabilities and averages. In an inferential problem, we face an inverse problem: given that \( Y \) has been observed to be \( y \), we wish to make statements about the unknown parameters associated with \( Y \).

Throughout, suppose that we have a model consisting of a collection of distribution functions indexed by an unknown parameter \( \theta \) taking values in some space \( \Theta \). We assume that \( Y \) is sampled according to one of these distributions, resulting in observed data \( y \).

Let \( f(y|\theta) \) be the probability mass function if \( Y \) is discrete, or the probability density function if \( Y \) is continuous (more generally, \( \{f(y|\theta) : \theta \in \Theta\} \) can be a family of densities with respect to some dominating measure, which is normally taken to be counting measure in the discrete case and Lebesgue measure in the continuous case). The model may be parametric (which we will take to mean that \( \Theta \) is finite-dimensional) or nonparametric (e.g., the space of all CDFs, or all unimodal continuous PDFs).

The likelihood function, defined below, gives the probability of the observed data as a function of the unknowns. This gives a graphical way to see what the data say about \( \theta \), and to compare different values of \( \theta \).

Definition 1.1. The likelihood function of the model \( f(y|\theta) \) with observed data \( y \) is the function

\[
L(\theta) = f(y|\theta),
\]

viewed as a function of \( \theta \) with the observed data \( y \) fixed. Since \( L(\theta) \) involves a product, it is often more convenient to work with the logarithm of the likelihood function, so we define the log-likelihood function by

\[
l(\theta) = \log L(\theta).
\]

1.2. Notations. We will use the following generic notations, unless specified otherwise.

1. \( Y = (Y_1, \ldots, Y_n) \) to denote the vector of \( n \) observations \( Y_1, \ldots, Y_n \).
2. While discussing frequentist methods, \( f_\theta(y) \) to denote a probability density/mass function and \( F_\theta(y) \) to denote a cumulative distribution function (CDF) of a random variable \( Y \), where the probability distribution is indexed by parameter \( \theta \).
3. \( P_{\theta}[Y \in A] \) to denote the probability that the observed data belongs to a (measurable) set \( A \), where the underlying probability distribution is indexed by parameter \( \theta \).
2. Sufficiency and Ancillarity

2.1. Sufficient statistics. A sufficient statistic is a statistic that contains all of the information available from the data which is pertinent for inference about \( \theta \). This can serve the purpose of data reduction, which is one major purpose of statistics (indeed, Fisher wrote that “the object of statistical methods is the reduction of data”). Also, we will see that conditioning on a sufficient statistic is useful in creating estimators.

**Definition 2.1.** Let \( Y_1, \ldots, Y_n \overset{iid}{\sim} f_{\theta}(\cdot) \). A statistic \( T(Y) \) is sufficient for \( \theta \) (or, more generally, for the family of distributions indexed by \( \theta \)) if the conditional distribution of \( Y|T \) does not depend on \( \theta \). That is, if we condition on a sufficient statistic \( T \), the distribution of the data no longer involves \( \theta \): the data depend on \( \theta \) only via \( T \). Mathematically, the conditional probability \( P_{\theta}[Y \in A|T = T(Y)] \) is free of \( \theta \) for any measurable set \( A \).

Sufficiency can also be defined via Bayesian or information theoretic perspectives.

**Theorem 2.2.** The following are equivalent.

(a) The statistic \( T \) is sufficient for \( \theta \), i.e., the conditional distribution of \( Y \) given \( T \) does not depend on \( \theta \).

(b) (Bayesian) The conditional distribution of \( \theta \) given \( T \) is the same as the posterior distribution of \( \theta \) given \( Y \), for any prior distribution on \( \theta \).

(c) (information theory) The chain \( \theta \rightarrow T \rightarrow Y \) is Markovian for any distribution on \( \theta \), i.e., \( \theta \) and \( Y \) are conditionally independent given \( T \).

**Proof.** Exercise. \( \square \)

**Example 2.3.** Let \( Y_1, Y_2 \) be i.i.d. Pois(\( \lambda \)), and let \( T \equiv Y_1 + Y_2 \). Then \( (Y_1, Y_2)|T \sim \text{Multinomial}(T, (1/2, 1/2)) \), so \( T \) is sufficient. Likewise, any one-to-one function of \( T \) is sufficient. Check though that, for example, \( T_2 \equiv Y_1 + 2Y_2 \) is not sufficient.

**Example 2.4.** Let \( Y_1, Y_2, \ldots, Y_n \) be i.i.d. Bern(\( p \)). Then \( \sum_i Y_i \) is sufficient since the distribution of \( (Y_1, \ldots, Y_n) \) given that \( \sum_i Y_i = k \) is uniform on the \( \binom{n}{k} \) possible sequences.

Rather than using the definition of sufficiency, it is usually easier to use the following criterion. The result says that \( f_{\theta}(y) \) depends on \( \theta \) only through \( T(y) \).

**Theorem 2.5** (Factorization Theorem). A statistic \( T(Y) \) is sufficient for \( \theta \) if and only if the joint probability function \( f_{\theta}(y) \) can be expressed in the form:

\[
f_{\theta}(y) = g_{\theta}(T(y)) h(y)
\]

for some functions \( g_{\theta}(\cdot) \) and \( h(\cdot) \), where \( g(\cdot) \) involves \( \theta \) and \( h(\cdot) \) is free of \( \theta \).

The Factorization Criterion holds in great generality, but the general proof requires more measure theory then we wish to get embroiled in here. For the discrete case, we can argue as follows.
Proof. Only if part: Let $T$ be sufficient. For any $y$ and the corresponding $T(Y) = T(y)$,
\[
  f_\theta(y) = P_\theta [Y = y] = P_\theta \left[ Y = y \bigcap T(Y) = T(y) \right],
\]
because \( \{ y : Y = y \} \subset \{ y : T(Y) = T(y) \} \)
\[
  = P_\theta [Y = y|T(Y) = T(y)] P_\theta [T(Y) = T(y)].
\]
By definition of sufficiency, the first term is free of \( \theta \) and the second term depends on \( \theta \) and on \( y \) through \( T(y) \). Thus, the joint pmf \( f_\theta(y) \) is of the desired form with \( h(y) = P_\theta [Y = y|T(Y) = T(y)] \) and \( g_\theta (T(y)) = P_\theta [T(Y) = T(y)] \).

If part: We now assume the condition
\[
  f_\theta(y) = P_\theta [Y = y] = g_\theta (T(y)) h(y).
\]
Then, using the general fact from the proof of the only if part that \( P_\theta [Y = y] = P_\theta [Y = y \bigcap T(Y) = T(y)] \), it follows that the conditional probability
\[
  P_\theta [Y = y|T(Y) = T(y)] = \frac{P_\theta [Y = y]}{P_\theta [T(Y) = T(y)]} = \frac{g_\theta (T(y)) h(y)}{\sum_{y:T(Y) = T(y)} g_\theta (T(y)) h(y)} = \frac{h(y)}{\sum_{y:T(Y) = T(y)} h(y)},
\]
which is free of \( \theta \).

We now present a couple of examples to illustrate applications of the factorization theorem:

**Example 2.6.** (Exponential family) Let \( Y_1, \ldots, Y_n \) be iid observations from
\[
  f_\theta(y) = \exp \left( \eta(\theta) T(y) - \psi(\eta) \right) h(y).
\]
The joint pdf is given by
\[
  f_\theta(y) = \exp \left( \eta(\theta) \sum_{i=1}^n T(y_i) - n\psi(\eta) \right) \prod_{i=1}^n h(y_i).
\]
By the factorization theorem, \( T(Y) = \sum_{i=1}^n T(Y_i) \) is a sufficient statistic for \( \theta \). More generally, for \( Y_1, Y_2, \ldots, Y_n \) i.i.d. from a multi-parameter exponential family
\[
  f_\theta(y) = \exp \left( \sum_{j=1}^k \eta_j(\theta) T_j(y) - \psi(\eta) \right) h(y),
\]
the statistic \( (\sum_{i=1}^n T_1(Y_i), \ldots, \sum_{i=1}^n T_k(Y_i)) \) is jointly sufficient for \( \theta \). Note that the dimension of the sufficient statistic matches the number of parameters.
Example 2.7. (Uniform distribution) Let \( Y_1, \ldots, Y_n \overset{iid}{\sim} \text{Unif}[0, \theta] \). Let \( Y_{(n)} = \max(Y_1, \ldots, Y_n) \) be the largest order statistic and \( Y_{(1)} = \min(Y_1, \ldots, Y_n) \) be the smallest order statistic. The joint pdf of \( Y \) is given by

\[
f_\theta(y) = \frac{1}{\theta^n} 1(y_{(n)} \leq \theta) 1(y_{(1)} \geq 0).
\]

By the factorization theorem, \( Y_{(n)} \) is a sufficient statistic for \( \theta \).

2.2. Minimal Sufficient Statistics. A sufficient statistic is not unique. First of all, it is easy to see that any one-to-one transformation of a sufficient statistic gives a sufficient statistic. Moreover, note that the entire data always form a sufficient statistic, which belies the intention of keeping a simplified function of the data without losing information. We therefore define minimality to capture the idea of keeping just what is necessary for inference about \( \theta \).

Definition 2.8. A sufficient statistic \( T \) is minimal if \( T \) is a function of any other sufficient statistic. That is, if \( S \) is sufficient, then there is a function \( g \) such that \( T = g(S) \) a.s. for all \( \theta \).

Some examples follow.

1. For \( Y_1, Y_2, \ldots, Y_n \overset{iid}{\sim} N(\mu, \sigma^2) \) with \( \mu \) and \( \sigma^2 \) unknown, \((\sum_{i=1}^n Y_i; \sum_{i=1}^n Y_i^2)\) is minimal sufficient. Equivalently, \((\bar{Y}, \sum_{i=1}^n (Y_i - \bar{Y})^2)\) is minimal sufficient.

2. For \( Y_1, \ldots, Y_n \overset{iid}{\sim} \text{Unif}(\theta - 1/2, \theta + 1/2) \), the 2-dimensional statistic \((\min Y_i, \max Y_i)\) is minimal sufficient. This family is not an EF since the support depends on \( \theta \).

A minimal sufficient statistic always exists (except in certain pathological measure-theoretic cases). By definition, the minimal sufficient statistic will be unique up to one-to-one transformations. The following Theorem from Casella and Berger (2002) provides us with a guideline to establish minimal sufficiency.

Theorem 2.9. Let \( Y \) constitute a vector of iid observations from a probability distribution with density \( f_\theta(y) \) and let \( T(Y) \) be a sufficient statistic for \( \theta \). If for two distinct data points \( x \neq y \), the ratio \( f_\theta(x)/f_\theta(y) \) is free of \( \theta \) only if \( T(x) = T(y) \), then \( T \) is minimal sufficient.

Proof. Let \( T' \) be any other sufficient statistic. We will show that \( T \) is a function of \( T' \), for which it suffices to show that if any two points \( x \) and \( y \) \((x \neq y)\) satisfy \( T'(x) = T'(y) \), then they also satisfy \( T(x) = T(y) \).

For two such points \( x \) and \( y \) satisfying \( T'(x) = T'(y) \), we have that

\[
\frac{f_\theta(x)}{f_\theta(y)} = \frac{g_\theta(T'(x)) h(x)}{g_\theta(T'(y)) h(y)} = \frac{h(x)}{h(y)},
\]

which is free of \( \theta \). Hence by the condition of the Theorem, we must have \( T(x) = T(y) \).
Example 2.10. Let $Y_1, \ldots, Y_n \overset{iid}{\sim} N(\mu, \sigma^2)$. For $x \neq y$, the ratio

$$\frac{f_\theta(x)}{f_\theta(y)} = \frac{\exp\left\{ -\sum_{i=1}^n (x_i - \mu)^2 / (2\sigma^2) \right\}}{\exp\left\{ -\sum_{i=1}^n (y_i - \mu)^2 / (2\sigma^2) \right\}}$$

$$= \exp\left\{ -\frac{1}{2\sigma^2} \left( \sum_{i=1}^n x_i^2 - \sum_{i=1}^n y_i^2 \right) + \frac{\mu}{\sigma^2} \left( \sum_{i=1}^n x_i - \sum_{i=1}^n y_i \right) \right\},$$

which is free of $\mu$ and $\sigma^2$ (i.e., same for all values of $\mu$ and $\sigma^2$) if and only if

$$\sum_{i=1}^n x_i = \sum_{i=1}^n y_i, \quad \sum_{i=1}^n x_i^2 = \sum_{i=1}^n y_i^2.$$ 

Thus, $(\sum_{i=1}^n Y_i, \sum_{i=1}^n Y_i^2)$ is a minimal sufficient statistic for $(\mu, \sigma^2)$.

2.3. Completeness and minimal sufficiency. We now introduce the notion of completeness and a complete sufficient statistic, which provides us with an alternative route to establishing minimal sufficiency.

Definition 2.11. A statistic $T(Y)$ is called complete for a family of distributions indexed by parameter $\theta$ if it is impossible to construct a non-trivial unbiased estimator of zero from the statistic. That is, the only way to have $E_\theta[h(T)] = 0$ for all $\theta$ is $h(T)$ to be 0 almost surely, i.e., $P_\theta[h(T)] = 0$.

Example 2.12. (Casella and Berger, 2002) Let $Y_1, \ldots, Y_n \overset{iid}{\sim} Ber(p)$. We shall show that the statistic $T = T(Y) = \sum_{i=1}^n Y_i$ is a complete sufficient statistic for $p$. Let $h(T)$ be a function of $T$ that satisfies $E_p[h(T)] = 0$. Since $T \sim Binom(n, p)$, this means

$$\sum_{t=0}^n h(t)\binom{n}{t}p^t (1-p)^{n-t} = 0$$

$$\Rightarrow \ (1-p)^n \sum_{t=0}^n h(t)\binom{n}{t} \left( \frac{p}{1-p} \right)^t = 0 \ \forall \ p \in (0,1).$$

The term $(1-p)^n$ is not zero for any $p \in (0,1)$. Thus we must have $\sum_{t=0}^n h(t)\binom{n}{t}r^t = 0$ where $r = p/(1-p) \in (0, \infty)$. Now $\sum_{t=0}^n h(t)\binom{n}{t}r^t$ is a polynomial in $r$ of degree $t$. To be zero for all $r$, each coefficient must be zero. Thus $P_p[h(T) = 0] = 1$ for all $p \in (0,1)$, implying that $T$ is complete sufficient.

A complete sufficient statistic will automatically be minimal (except in pathological cases), and completeness greatly simplifies the ancillarity issues as discussed below.

Proposition 2.13. Any complete sufficient statistic is also minimal, as long as at least one minimal sufficient statistic exists for the model.
Proof. Let $T$ be complete and sufficient, and let $M$ be minimal sufficient for parameter $\theta$. Consider the quantity $h(T) = E_\theta(T|M) - T$. Then the following claims can be made:

(a) $h(T)$ is a statistic, i.e., a function of $Y$ only that is free of $\theta$.
(b) $h(T)$ is a function of $T$.
(c) $h(T)$ is an unbiased estimator of zero.

Claim (a) is true because the first term $E_\theta(T|M)$ depends on the conditional distribution of $T|M$, which is free of $\theta$ because it is conditioned on the sufficient statistic $M$. Also, $E_\theta(T|M)$ is a function of $M$, which is a function of $Y$. The second term $T$ is a statistic by definition. To see why claim (b) is true, note that the first term $E_\theta(T|M)$ is a function of $M$, which is minimal sufficient, and hence a function of every other sufficient statistic, including $T$. Thus, both terms are functions of $T$. Finally, part (c) holds because $E\left[E_\theta(T|M)\right] = E(T) - E(T) = 0$.

By (c), $E_\theta(h(T)) = 0$. By Definition 2.11, this means $h(T) = 0$ almost surely. Hence $E(T|M) = T$ almost surely, which implies that $T$ is a function of $M$ and hence $T$ is minimal sufficient.

Completeness is often difficult to prove, but it is automatic that the natural sufficient statistic in an exponential family is complete, under mild conditions. However, a complete sufficient statistic may not exist. Consider the following example: let $Y_1, \ldots, Y_n \iddist N(\theta, \theta^2)$ (curved normal family). It can be shown that the statistic $T = (\sum_{i=1}^n Y_i, \sum_{i=1}^n Y_i^2)$ is minimal sufficient but not complete sufficient.

2.4. Ancillarity. A statistic $A(Y)$ is ancillary if its distribution does not depend on $\theta$. Why then isn’t “ancillary” a polite word for “useless”? We have been assuming that the model $f_\theta(y)$ is correct, but if we wish to test the model, or compare different models, then ancillary information may be very valuable. Indeed, the fact that an ancillary statistic is (typically) independent of the minimal sufficient statistic is very convenient in testing the model.

Example 2.14. Some examples of ancillary statistic are given below.

(i) A trivial example of ancillary statistic is a constant.
(ii) Let $Y_1, \ldots, Y_n \iddist N(\mu, 1)$. The statistic $A(Y) = \sum_{i=1}^n (Y_i - \bar{Y})^2 \sim \chi^2_{n-1}$ is ancillary for $\mu$.
(iii) Let $Y_1, \ldots, Y_n \iddist \text{expo}(\lambda)$ with pdf $f_\lambda(y) = \lambda \exp(-\lambda y), \lambda > 0, y \geq 0$. The statistic $A(Y) = Y_n/(Y_1 + \ldots + Y_n)$ is ancillary for $\lambda$. To see why this is true, consider the transformation $U_i = \lambda Y_i$ for $i = 1, \ldots, n$.
   Then $A(Y) = U_n/(U_1 + \ldots + U_n)$, where $U_i \iddist \text{expo}(1)$. This example can be generalized to the case where $Y_1, \ldots, Y_n$ are iid from a scale family of distributions, i.e., with cdf of the form $F(y/\sigma), \sigma > 0$.
   Then, any statistic that depends on $Y$ only through the $n-1$ values
\[ Y_2/Y_1, \ldots, Y_n/Y_1 \text{ (e.g., } Y_1/\sum_{i=1}^n Y_i \text{)} \text{ is an ancillary statistic. Refer to Section 6.2 of Casella and Berger (2002) for further details.} \]

Basu’s Theorem is an extremely elegant and useful result relating completeness to ancillarity. The proof beautifully combines each ingredient.

**Theorem 2.15** (Basu). If \( T \) is a complete sufficient statistic for \( \theta \) and \( A \) is ancillary, then \( A \) and \( T \) are independent.

**Proof.** For any measurable set \( B \), \( h(T) \equiv P_\theta(A \in B|T) - P_\theta(A \in B) \) is a function of \( T \) not depending on \( \theta \) (by sufficiency for the first term, by ancillarity for the second). We have \( E(h(T)) = 0 \), so completeness gives \( h(T) = 0 \) almost surely. \( \square \)

**Example 2.16.** Let \( Y_1, \ldots, Y_n \) be i.i.d. \( N(0, \sigma^2) \). We will use Basu’s Theorem to give an easy proof that \( \bar{Y} \) and \( S = \sum_{i=1}^n (Y_i - \bar{Y})^2 \) are independent. To apply Basu’s Theorem, we first need to specify an inference problem for which \( \bar{Y} \) is complete and sufficient. To do this, we introduce a parameter \( \mu \) for the mean: let \( Y_1, \ldots, Y_n \) be i.i.d. \( N(\mu, \sigma^2) \) with \( \mu \) unknown and \( \sigma^2 \) known. (Even though the mean is 0 in the original problem, we require an unknown parameter for which \( \bar{Y} \) is a complete sufficient statistic.) Then \( \bar{Y} \) is a complete sufficient statistic (this can be shown directly or using NEFs and the fact that the MGF (if it exists) uniquely determines the distribution). The statistic \( S \) is ancillary (indeed, so is the vector of residuals \( (Y_1 - \bar{Y}, \ldots, Y_n - \bar{Y}) \)) because it is translation-invariant (add and subtract \( \mu \) in each term). By Basu’s Theorem, \( \bar{Y} \) and \( S \) are independent for all \( \mu \) and \( \sigma^2 \). In particular, they are independent for \( N(0, \sigma^2) \), regardless of whether \( \sigma^2 \) is known.

**Example 2.17.** Let us revisit example 2.14(iii), where we argued that \( Y_n/(Y_1 + \cdots + Y_n) \) is ancillary for \( \lambda \). It is easy to see that \( \bar{Y} \) is CSS for \( \lambda \). Hence by Basu’s theorem, \( \bar{Y} \) and \( Y_n/(Y_1 + \cdots + Y_n) \) are independent.

**Example 2.18.** Let \( Y_1, \ldots, Y_n \) be i.i.d. \( N(\mu, \sigma^2) \). Two possible estimators for \( \mu \) are the sample mean \( \bar{Y} \) and the sample median \( \text{med}(Y) \). We may wish to know how correlated these two estimators are. Using Basu’s Theorem, it is easy to find their covariance:

\[
\text{Cov}(\bar{Y}, \text{med}(Y)) = \text{Cov}(\bar{Y}, \text{med}(Y) - \bar{Y} + \bar{Y}) \\
= \text{Cov}(\bar{Y}, \text{med}(Y) - \bar{Y}) + \text{Cov}(\bar{Y}, \bar{Y}) \\
= \sigma^2/n,
\]

since \( \text{med}(Y) - \bar{Y} \) is ancillary (for \( \mu \), treating \( \sigma \) as known). Note that the same answer, \( \sigma^2/n \), holds if the median is replaced by any statistic \( T \) such that \( T - \bar{Y} \) is ancillary, e.g., any order statistic!
3. **Unbiased point estimation**

3.1. **Unbiased point estimator.** We now discuss how a sufficient statistic (possibly a complete or a minimal sufficient statistic) or its function can be used as an *unbiased point estimator* of a parameter $\theta$ or a parametric function $g(\theta)$. For example, in the context of iid observations $Y_1, \ldots, Y_n$ from a $\text{Poi}(\lambda)$ distribution, one might be interested in estimating the parameter $\lambda$ itself, or a parametric functions like the true probability of obtaining a zero value from the distribution (i.e., $P_\lambda(Y = 0) = e^{-\lambda}$). We first introduce the concept of unbiasedness.

**Definition 3.1.** A statistic $T(Y)$ is an unbiased estimator of a parametric function $g(\theta)$ if its expected value (with respect to the distribution of the statistic over hypothetical replications of the data $Y$) equals $g(\theta)$, i.e., $E_\theta[T(Y)] = g(\theta)$.

An unbiased estimator is not necessarily a “good” one, as we will soon see through some examples. For example, in the Poisson example, a single observation $Y_1$ is an unbiased estimator of $\lambda$. We will however, argue, that the concepts of sufficiency and completeness will allow us to start from such “trivial” estimators, improve them, and even arrive at estimators that are “best” with respect to some criterion of goodness for point estimators. The criterion that we will consider in this Section is *Mean Squared Error* (MSE), defined as the expectation of the squared difference between the parametric function $g(\theta)$ and its estimator $T$, i.e., $E_\theta[T - g(\theta)]^2$. Note that the MSE can be expressed as

$$E_\theta[T - E_\theta(T)]^2 + [E_\theta(T) - g(\theta)]^2.$$  

The first term in the above expression is the variance of $T$ denoted by $\text{Var}_\theta(T)$, and the second is the square of the bias $E_\theta[T] - g(\theta)$. If $T$ is unbiased, the second term vanishes, and the MSE reduces to the variance. A common and well studied problem in Statistical inference is the “bias-variance” tradeoff. An estimator may have a small bias, but may be preferable to an unbiased estimator with a much larger variance. However, unbiasedness is an attractive (albeit nonsensical in certain situations) property on its own, and for the most part of this Section, we will keep our discussion restricted to unbiased estimators.

Based on the MSE criterion, it makes sense to find an estimator $T(Y)$ of $g(\theta)$ that has the smallest variance in the class of all unbiased estimators. The key idea that we will use to find such an estimator with minimum variance consists of the following steps:

(i) Find *any* (possibly trivial) unbiased estimator.

(ii) Improve this unbiased estimator by conditioning on a sufficient statistic (better unbiased estimator).

(iii) If the sufficient statistic used for conditioning is complete, then we’ve found the “best” estimator.
In this Section we will state and prove two Theorems. Theorem 3.2, due to Rao (1945) and Blackwell (1947), helps accomplish step (ii). Theorem 3.3 can be used to achieve task (iii).

**Theorem 3.2** (Rao-Blackwell Theorem). Let \( W(Y) \) be an unbiased estimator of \( g(\theta) \) and \( T \) be any sufficient statistic. Consider the estimator \( \phi(T) = E_\theta[W|T] \). This estimator satisfies:

(a) The estimator \( \phi(T) \) is an unbiased estimator of \( g(\theta) \), i.e.,
\[
E_\theta[\phi(T)] = g(\theta).
\]

(b) The estimator \( \phi(T) \) is “uniformly” better than \( W \) in the sense
\[
\text{Var}_\theta[\phi(T)] \leq \text{Var}_\theta[W] \quad \forall \theta \in \Theta.
\]

**Proof.** First, note that \( \phi(T) \) is an estimator (a statistic that is free of \( \theta \) by sufficiency of \( T \). We have that,
\[
g(\theta) = E_\theta[W] = E[E[W|T]] = E_\theta[\phi(T)].
\]
This proves part (a). To prove part (b), note that
\[
\text{Var}_\theta[W] = E[\text{Var}(W|T)] + \text{Var}[E(W|T)]
= E[\text{Var}(W|T)] + \text{Var}[\phi(T)]
\geq \text{Var}_\theta[\phi(T)] \quad \forall \theta.
\]
The proof is complete. \( \square \)

Theorem 3.2 tells us that we can take any unbiased estimator \( W \) of \( g(\theta) \) and improve it by conditioning on any sufficient statistic. The question that naturally arises is, does this sequential improvement depend on the initial choice \( W \)? what would have happened if a different estimator, say \( \tilde{W} \) was chosen and improved upon by conditioning? Can we reach a “best” estimator by starting from either of them? Theorem 3.3 provides an answer to these questions.

**Theorem 3.3** (Lehmann-Scheffe Theorem). An unbiased estimator of \( g(\theta) \) that is a function of a complete sufficient statistic is the unique uniformly minimum variance unbiased estimator (UMVUE) of \( g(\theta) \).

**Proof.** Let \( W \) and \( \tilde{W} \) be any two unbiased estimators of \( g(\theta) \), \( T \) a complete sufficient statistic (CSS) for \( \theta \), and \( \phi(T) = E[W|T] \) and \( \tilde{\phi}(T) = E[\tilde{W}|T] \) be the estimators obtained by conditioning \( W \) and \( \tilde{W} \) on the CSS \( T \). Then, by the Rao-Blackwell theorem, \( \phi(T) \) and \( \tilde{\phi}(T) \) are both (i) functions of the CSS \( T \), (ii) unbiased estimators of \( g(\theta) \), and (iii)
\[
\text{Var}_\theta[\phi(T)] \leq \text{Var}[W], \quad \text{Var}_\theta[\tilde{\phi}(T)] \leq \text{Var}[\tilde{W}].
\]
Let \( h(T) = \phi(T) - \tilde{\phi}(T) \). Then
\[
E_\theta[h(T)] = E_\theta[\phi(T) - \tilde{\phi}(T)] = g(\theta) - g(\theta) = 0.
\]
By completeness of $T$, $h(T) = 0$ almost surely, implying that $\phi(T) = \tilde{\phi}(T)$ almost surely.

\begin{remark}[Implications of the Lehmann-Scheffe theorem]
(a) As long as we have not found a CSS, we can keep on improving an unbiased estimator by conditioning on any sufficient statistic. Once a CSS is obtained, and used for conditioning, we cannot improve further because the unique “best” estimator has been obtained by conditioning on the CSS.

(b) If we have found a CSS and a function of the CSS that is unbiased for $g(\theta)$, that function is the UMVUE of $g(\theta)$.
\end{remark}

\begin{example}[UMVUE can be a ridiculous estimator!]
Consider a single observation $Y \sim Poi(\lambda)$. Suppose you are interested in the function $g(\lambda) = e^{-2\lambda}$. It is easy (show this) to see that $T(Y) = (-1)^{Y}$ is an unbiased estimator of $g(\lambda)$. Because $Y$ is a CSS, by the Lehmann-Scheffe theorem, $T(Y)$ is the UMVUE of $e^{-2\lambda}$, which means, irrespective of the observed value of $Y$, the UMVUE is $-1$ or $+1$ depending on whether $Y$ is even or odd!
\end{example}

\begin{remark}
Unbiasedness itself may be a ridiculous concept in many practical situations, see for example, the famous Circus Example by Basu (1971). However, it remains a useful and intuitive concept in statistical inference, and often a starting point in many inference problems.
\end{remark}

### 3.2. The Cramér-Rao Lower Bound
Having discussed unbiased estimators of parametric functions, their construction and properties, we now aim at finding a lower bound of the variance of such unbiased estimators. The famous Cramér-Rao Lower Bound (Rao, 1945; Cramer, 1946) derives an expression for variance which is the best that an unbiased estimator can achieve, in the sense of minimizing variance (which is equivalent to minimizing MSE when working with unbiased estimators). However, in order to state and prove the theorem, we need to introduce some additional concepts related to the likelihood function defined by 1.1 and briefly discussed in Section 1.1. We thus introduce two important terms — the Score function and Fisher information — in the following two subsections and discuss their properties.

#### 3.2.1. The Score Function
The derivative of the log-likelihood function occurs so commonly that it has its own name: the score function.
Definition 3.8. The score function for a single observation $Y$ is defined as

\[ \dot{\ell}_{\theta}(Y) = \dot{\ell}_1(Y, \theta) = \frac{\partial \log f_{\theta}(Y)}{\partial \theta}. \]

Definition 3.9. The score function $\dot{\ell}(Y, \theta)$ for the data $Y = (Y_1, \ldots, Y_n)$ is defined as

\[ \dot{\ell}_n(Y, \theta) = \frac{\partial \log f_{\theta}(Y)}{\partial \theta} = \frac{\partial l(\theta)}{\partial \theta}, \]

where $l(\theta)$ is the log-likelihood function.

Henceforth, for notational simplicity, we will drop the index $n$ and write the score function for data $Y$ as $\dot{\ell}(Y, \theta)$. The score function is often called the “score statistic”, but that terminology falsely suggests that it is a statistic (i.e., a quantity computable from the data); in fact, the score function depends on both $Y$ and $\theta$. Note that, if $Y_1, \ldots, Y_n$ are iid observations, then from (3.2)

\[ \dot{\ell}_n(Y, \theta) = \sum_{i=1}^{n} \frac{\partial \log f_{\theta}(Y_i)}{\partial \theta} = \sum_{i=1}^{n} \dot{\ell}_1(Y_i, \theta). \]

As we shall see later, an immediate application of the score function is in finding the MLE: after observing $Y = y$, set $\dot{\ell}(y, \theta) = 0$ and solve for $\theta$.

Under mild regularity conditions, the score function has zero mean.

Proposition 3.10. Under regularity conditions,

\[ E_{\theta} \dot{\ell}(Y, \theta) = 0 \text{ for all } \theta \in \Theta. \]

Proof. For simplicity, we will assume that the density is absolutely continuous with respect to Lebesgue measure, but the same result holds for discrete or mixed distributions (integrating with respect to the appropriate measure).

\[
E_{\theta} \dot{\ell}(Y, \theta) = \int \dot{\ell}(y, \theta) f_{\theta}(y) dy \\
= \int \frac{L'(\theta)}{L(\theta)} L(\theta) dy \\
= \int \frac{\partial f_{\theta}(y)}{\partial \theta} dy \\
= \frac{\partial}{\partial \theta} \int f_{\theta}(y) dy = 0,
\]

as desired. \qed

Here under regularity conditions typically mean that we allow exchange of derivative and integration, but the readers should be aware of the fact that ‘regularity conditions’ may vary from place to place.
The variance of the score function (equivalently, its second moment) gives an often-convenient way to compute expected Fisher information introduced in the following subsection.

**Proposition 3.11.** Under regularity conditions,
\[ \text{Var}_\theta(\dot{\ell}(Y, \theta)) = E_\theta(\dot{\ell}^2(Y, \theta)). \]

This follows directly from Proposition 3.10.

### 3.2.2. Fisher Information

The “expected” Fisher information, or simply the Fisher information for a single observation \( Y \) is defined as:

\[(3.3)\quad I_1(\theta) = E_\theta \left( \frac{\partial}{\partial \theta} \log f_\theta(Y) \right)^2 = E_\theta(\dot{\ell}_1^2(Y, \theta)), \]

where \( \dot{\ell}_1(Y, \theta) \) is given by (3.1).

Fisher information based on data \( Y = (Y_1, \ldots, Y_n) \) is defined as

\[(3.4)\quad I_n(\theta) = E_\theta \left( \frac{\partial}{\partial \theta} \log f_\theta(Y) \right)^2 = E_\theta(\dot{\ell}^2(Y, \theta)). \]

As in the case of score function, for notational simplicity, we will drop the index \( n \) and write the Fisher information for data \( Y \) as \( I(\theta) \), unless otherwise specified.

**Proposition 3.12.** Under mild regulatory conditions,

\[(3.5)\quad I(\theta) = -E_\theta \left( \frac{\partial^2}{\partial \theta^2} \log f_\theta(Y) \right). \]

**Proof.** Exercise. \( \square \)

We now develop some important properties of Fisher information, which will be useful in computing information and provide some intuition for the highly suggestive term “information”.

A simple but fundamental property of Fisher information is additivity: if \( Y_1 \) and \( Y_2 \) are independent observations, then the information about \( \theta \) from \((Y_1, Y_2)\) (given by \( I_2(\theta) \) defined by (3.4) with \( n = 2 \)) is the sum of the information from \( Y_1 \) and the information from \( Y_2 \). This simple property is both intuitively reasonable (information should accumulate as independent evidence arrives) and very useful for computation. In particular, for \( Y_1, \ldots, Y_n \) i.i.d., we have \( I(\theta) = nI_1(\theta) \), where \( I_1(\theta) \) is given by (3.3)

**Example 3.13** (Poisson). Let \( Y_1, \ldots, Y_n \ iid \sim Poi(\lambda) \). It is easy to see that (check this!) \( I_1(\lambda) = 1/\lambda \) and \( I(\lambda) = n/\lambda \).

**Example 3.14** (Location Families). Let \( f_\theta(y) = f(y - \theta) \) with \( f \) a known density and \( \theta \) a location parameter. Then

\[
I_1(\theta) = \int \dot{\ell}_1^2(y, \theta) f(y - \theta)dy = \int \left( \frac{f'(y - \theta)}{f(y - \theta)} \right)^2 f(y - \theta)dy = \int \left( \frac{f'(u)}{f(u)} \right)^2 f(u)du
\]
is a constant (not depending on \( y \) or \( \theta \)).
Example 3.15 (Scale Families). Let \( f_\theta(y) = \theta^{-1}f(y/\theta) \) with \( f \) a known density and \( \theta > 0 \) a scale parameter. Then a calculation similar to that of the previous example yields \( I_1(\theta) \propto 1/\theta^2 \) (checking this is left as an exercise). This makes sense since we can write \( Y = \theta X \) with \( X \sim f \), and then \( \text{Var}(\theta X) = \theta^2 \text{Var}(X) \) (so small \( \theta \) corresponds to small variance, in which case an observed \( y \) is very informative).

3.2.3. The Cramér-Rao Theorem.

Theorem 3.16 (CRLB). Let \( g(\theta) \) be a parametric function of interest and \( T(Y) \) be an unbiased estimator of \( g(\theta) \). Then under regularity conditions,

\[
\text{Var}(T) \geq \frac{[g'(\theta)]^2}{I(\theta)},
\]

where \( I(\theta) \) is the Fisher information defined by (3.4).

Proof. Consider the Score function \( \ell(Y, \theta) \) defined by (3.2), and the covariance between \( \ell(Y, \theta) \) and \( T(Y) \). By the Cauchy-Schwarz inequality,

\[
\left\{ \text{Cov} \left( \ell(Y, \theta), T(Y) \right) \right\}^2 \leq \text{Var} \left( \ell(Y, \theta) \right) \text{Var} \left( T(Y) \right).
\]

Now,

\[
\text{Cov} \left( \ell(Y, \theta), T(Y) \right) = E \left[ \ell(Y, \theta)T(Y) \right] - E \left[ \ell(Y, \theta) \right] E \left[ T(Y) \right]
\]

\[
= \int \frac{\partial \log f_\theta(y)}{\partial \theta} T(y)f_\theta(y)dy
\]

\[
= \int \frac{\partial f_\theta(y)}{\partial \theta} \frac{1}{f_\theta(y)} T(y)f_\theta(y)dy
\]

\[
= \frac{\partial}{\partial \theta} \int T(y)f_\theta(y)dy
\]

\[
\left(3.8\right)
\]

\[
= \frac{\partial}{\partial \theta} E_\theta \left( T(y) \right) = \frac{\partial}{\partial \theta} g(\theta) = g'(\theta).
\]

Now, by Proposition 3.11, \( \text{Var} \left( \ell(Y, \theta) \right) = E \left[ \ell^2(Y, \theta) \right] \), which equals \( I(\theta) \) by (3.4). Substituting \( \text{Var} \left( \ell(Y, \theta) \right) = I(\theta) \) and \( \text{Cov} \left( \ell(Y, \theta), T(Y) \right) = g'(\theta) \) from (3.8) into the CS-inequality 3.7, the result follows. \( \square \)

Here are two examples in which CRLB is achieved with equality.

Example 3.17. Consider \( n \) i.i.d. \( N(\mu, \sigma^2) \) observations. Then \( \text{Var}(\bar{Y}) = \frac{\sigma^2}{n} = \frac{1}{T_n(\mu)} \).

Example 3.18. Consider \( n \) i.i.d. \( \text{Pois}(\lambda) \) observations. Then \( \text{Var}(\bar{Y}) = \frac{\lambda}{n} = \frac{1}{T_n(\lambda)} \).

So equality holds for the Normal and the Poisson, when estimating the mean parameters; but equality would \( \text{not} \) hold in those examples if estimating a nonlinear function of the mean. You will explore the question of when
in general CRLB can be achieved with equality as a homework problem. We shall see later, that the MLE asymptotically achieves the CRLB, but it does not typically achieve it with equality for any fixed sample size.
4. Method of Moments and Maximum Likelihood Estimation

4.1. Method of Moments Estimation. Method of moments is often a straightforward way of obtaining estimators of parameters or parametric functions by equating sample moments to population moments. Recall that for a random variable $Y$ distributed with a probability density function $f_\theta(y)$, the $r$th raw moment of the distribution of $Y$ is

$$\mu'_r = E_\theta(Y^r), \ r = 1, 2, \ldots.$$ 

The $r$th central moment is

$$\mu_r = E_\theta(Y - \mu'_1)^r = E_\theta(Y - E_\theta(Y))^r, \ r = 1, 2, \ldots.$$ 

Thus, for example, mean of the distribution is the first raw moment and the variance the second central moment.

Definition 4.1. Let $Y_1, \ldots, Y_n$ be an iid random sample from $f_\theta(y)$, where $\theta \in \Theta \subseteq \mathbb{R}^k$. A method of moments estimator of $\theta$ is obtained by equating the first $k$ sample moments $\sum_{i=1}^n Y_i^r/n$ for $r = 1, \ldots, k$ to the corresponding $k$ population (raw) moments $\mu'_r(\theta) = E_\theta(Y^r)$ for $r = 1, \ldots, k$.

Example 4.2. Let $Y_1, \ldots, Y_n \overset{iid}{\sim} N(\mu, \sigma^2)$. The method of moments estimators can be obtained by equating the first two sample and population moments, i.e.,

$$\bar{Y} = \mu, \quad \frac{1}{n} \sum_{i=1}^n Y_i^2 = E(Y_1^2) = \mu^2 + \sigma^2.$$ 

Solving the above two equations, we get the following method of moments (MOM) estimators for $\mu$ and $\sigma^2$:

$$\hat{\mu}_{MOM} = \bar{Y},$$

$$\hat{\sigma}^2_{MOM} = \frac{1}{n} \sum_{i=1}^n Y_i^2 - \bar{Y}^2 = \frac{1}{n} \sum_{i=1}^n (Y_i - \bar{Y})^2.$$ 

Example 4.3. Let $Y_1, \ldots, Y_n \overset{iid}{\sim} Binom(k,p)$, where both $k$ and $p$ are unknown. Equating the first two sample moments to those of the population, we have

$$\bar{Y} = kp,$$

$$\frac{1}{n} \sum_{i=1}^n Y_i^2 = E(Y_1^2) = kp(1-p) + k^2p^2.$$ 

Solving the above two equations, the method of moments estimators of $k$ and $p$ are obtained as:

$$\hat{k}_{MOM} = \frac{\bar{Y}^2}{\bar{Y} - \sum_{i=1}^n (Y_i - \bar{Y})^2},$$

$$\hat{p}_{MOM} = \frac{\bar{Y}}{\hat{k}_{MOM}}.$$
These MOM estimators are not good, because both of them can be negative if the sample mean is smaller than the sample variance. However, obtaining MLEs of $k$ and $p$ is a non-trivial exercise. Like many other situations where the MLE is difficult to obtain, in this example the method of moments estimators provide a reasonable starting point for obtaining good estimators.

**Example 4.4** (MOM estimators for the exponential family). Consider the single parameter exponential family of distributions characterized by the pdf

$$f_\theta(y) = \exp \{ \eta(\theta)T(y) - \psi(\eta) \} h(y).$$

To obtain the method of moments estimator for $\theta$, note that by differentiating both sides of the identity

$$\int \exp \{ \eta(\theta)T(y) - \psi(\eta) \} h(y)dy = 1$$

within the integral with respect to $\eta$ yields

$$\int \exp \{ \eta T(y) - \psi(\eta) \} (T(y) - \psi'(\eta)) h(y)dy = 0$$

where $\psi'(\eta) = \partial \psi(\eta)/\partial \eta$. The above implies

$$\int T(y) \exp \{ \eta T(y) - \psi(\eta) \} h(y)dy = \psi'(\eta) \int \exp \{ \eta T(y) - \psi(\eta) \} h(y)dy,$$

which implies $E_\theta(T(Y)) = \psi'(\eta)$.

Thus, the following equation (4.1) yields the MOM estimator for the parameter $\theta$ of an exponential family:

(4.1) \[ \frac{1}{n} \sum_{i=1}^{n} T(Y_i) = \psi'(\eta). \]

4.2. **Maximum Likelihood Estimation.** A widely used method of obtaining a point estimate for a parameter $\theta$ is to find the maximum likelihood estimate (MLE). As the name suggests, the MLE is defined as any vector $\hat{\theta}_n$ maximizing $L(\theta)$.

The MLE is often a reasonable estimate, but it should not be put to use blindly without inspecting the entire likelihood function. For example, it sometimes happens that other values of $\theta$ are far more centrally located than the MLE (the mode of a distribution need not be near the mean!). Or there may be several peaks in the likelihood function, perhaps negligibly different in height. Careful study of the shape of $L(\theta)$ is more informative than locating a single point with slightly higher likelihood than other points.

**Example 4.5** (MLE in an exponential family). Let

$$Y_1, \ldots, Y_n \iid \exp \{ \eta(\theta)T(y) - \psi(\eta) \} h(y).$$
The log-likelihood equation is
\[ l(\theta, Y) = \left( \sum_{i=1}^{n} T(Y_i) \right) \eta(\theta) - n\psi(\eta) + \sum_{i=1}^{n} \log h(Y_i). \]

Differentiating both sides with respect to \( \eta \) yields
\[ \frac{\partial l}{\partial \eta} = \sum_{i=1}^{n} T(Y_i) - n\psi'(\eta). \]

Equating the above to zero, we obtain, the ML equation as
\[ \frac{1}{n} \sum_{i=1}^{n} T(Y_i) = \psi'(\eta), \]
which is exactly the same as equation (4.1) that yields the MOM estimator.

4.2.1. Computing the MLE. The following Cauchy location problem illustrates many important ideas for finding the MLE. This is not an exponential family, and can have a very complicated likelihood function with many local maxima.

**Example 4.6.** Let \( Y_1, Y_2, \ldots, Y_n \) be i.i.d. as a Cauchy location family, i.e., \( Y_i = C_i + \theta \) with \( C_i \) i.i.d. Cauchy. So
\[ f_\theta(y) = \frac{1}{\pi(1 + (y - \theta)^2)}. \]

We wish to estimate the location parameter \( \theta \). The log-likelihood function is
\[ l(\theta) = -\sum_{j=1}^{n} \ln(1 + (y_j - \theta)^2), \]
so the score function is
\[ \ell(y, \theta) = 2 \sum_{j=1}^{n} W_j(y_j - \theta), \]
where the \( W_j \) are “weights” defined by
\[ W_j(y, \theta) = \frac{1}{1 + (y_j - \theta)^2}. \]

The MLE \( \hat{\theta}_n \) satisfies \( \ell(y, \theta) = 0 \), so it satisfies
\[ \theta = \frac{\sum_{j=1}^{n} W_j y_j}{\sum_{j=1}^{n} W_j}, \]
which is a weighted average of the \( y_j \). Note that the weights depend on both \( \theta \) and on the \( y_j \), with the more extreme values of \( y_j \) given less weight.

A solution to the above equation is a fixed point of the function
\[ h(\theta) = \frac{\sum_{j=1}^{n} W_j(y, \theta)y_j}{\sum_{j=1}^{n} W_j(\theta, y)}, \]
i.e., we are looking for $\hat{\theta}_n$ satisfying $h(\hat{\theta}_n) = \hat{\theta}_n$. In such a setting, iterative techniques are usually used.

We first explain how the *Newton-Raphson* method can be applied here. Start with a guess $\theta_1$ for the solution to $\ell'(y, \hat{\theta}_n) = 0$ (e.g., the median of $y_1, \ldots, y_n$). Expand the score function as a Taylor series (up through the linear term):

$$
\ell(y, \theta_2) \approx \ell(y, \theta_1) + (\theta_2 - \theta_1) \ell'(y, \theta_1).
$$

Setting $\ell(y, \theta_2) = 0$ (the hoped-for value) and solving for $\theta_2$ gives

$$
\theta_2 = \theta_1 + \frac{\ell(y, \theta_1)}{-\ell'(y, \theta_1)}.
$$

(The $'$ denotes the derivative with respect to the $\theta$ component.) To evaluate this, note that

$$
-\ell'(y, \theta) = 2 \sum_{j=1}^n W_j (2W_j - 1) = 2 \sum_{j=1}^n (2W_j^2 - W_j),
$$

so

$$
\theta_2 = \theta_1 + \frac{\sum_{j=1}^n W_j (y_j - \theta)}{\sum_{j=1}^n (2W_j^2 - W_j)}.
$$

The procedure can then be iterated to find $\theta_3, \theta_4, \ldots$ until convergence. Note though that this may not give the *global* maximum (it may even give a local minimum). The same idea can be used for multi-dimensional $\theta$, using gradients and Hessians.

A closely-related alternative is to use *Fisher’s method of scoring*. This method replaces $-\ell'(Y, \theta_1)$ by its average $E_{\theta_1}(-\ell'(Y, \theta_1)) = I(\theta_1)$. Fisher scoring has the advantages that it is sometimes easier to compute and work with $I(\theta_1)$ than $-\ell'(Y, \theta_1)$, and that it is often less sensitive to the initial guess $\theta_1$.

Returning to the specific Cauchy location problem, let us compute $I(\theta)$. As shown above, in a location family $I(\theta)$ is a constant. Thus, we can choose $\theta$ to make the computations as simple as possible. Let us take $\theta = 0$. Then

$$
W_j = \frac{1}{1 + Y_j^2} = \frac{Z_2^2}{Z_2^2 + Z_1^2}
$$

with $Z_1, Z_2$ i.i.d. $N(0, 1)$, using the fact that a Cauchy is the ratio of two standard Normal r.v.s. But if $Z \sim N(0, 1)$, then $Z^2 \sim \chi_1^2 \sim 2\text{Gamma}(1/2)$, so $W_j \sim \text{Beta}(1/2, 1/2)$. In particular, $E_0 W_j = 1/2, \text{Var}_0 W_j = 1/8$. Therefore,

$$
I(\theta) = E_\theta(-\ell'(Y, \theta)) = -n E_\theta(2W_1(2W_1 - 1)) = n E_\theta(4W_1^2 - 2W_1) = \frac{n}{2}.
$$
This does not depend on $\theta$, so $I(\hat{\theta}_n)$ also equals $n/2$. Thus, Fisher scoring uses
\[
\theta_2 = \theta_1 + \frac{4}{n} \sum_{j=1}^{n} W_j(y_j - \theta).
\]

4.2.2. Properties of the MLE. Let $\hat{\theta}_n$ denote the MLE of a parameter $\theta$ obtained from an iid sample $Y_1, \ldots, Y_n$ drawn from $f_\theta(y)$. Assume that there exists a “true value” $\theta_0$ of $\theta$. Then, under mild regularity conditions, the MLE can be shown to be a consistent estimator of $\theta$ (i.e., it converges almost surely (strong convergence) or in probability (weak convergence) to the true value $\theta_0$).

Recall that the MLE is obtained by maximizing the log-likelihood function $f_\theta(Y) = \sum_{i=1}^{n} \log f_\theta(Y_i)$, which is the same as maximizing the “average log-likelihood” $n^{-1} \sum_{i=1}^{n} \log f_\theta(Y_i)$. Denote this quantity by
\[
(4.3) \quad \bar{l}_n(\theta) = n^{-1} \sum_{i=1}^{n} \log f_\theta(Y_i).
\]

Then, by definition,
\[
\hat{\theta}_n = \arg \max_{\theta \in \Theta} \bar{l}_n(\theta).
\]

We further need fix the following notation:
\[
\bar{l}(\theta) = E_{\theta_0} [\log f_\theta(Y)].
\]

**Theorem 4.7.** Suppose that

1. $|\bar{l}(\theta)| < \infty$ for all $\theta$ in the neighborhood of $\theta_0$,
2. $\{f_\theta : \theta \in \Theta\}$ is identifiable, i.e. $\theta_1 \neq \theta_2 \implies f_{\theta_1} \neq f_{\theta_2}$ under $P_{\theta_0}$,
3. $\log f_\theta(y)$ has a continuous derivative with respect to $\theta$ in a neighborhood of $\theta_0$ for each $y$.

Then almost surely there exists $\hat{\theta}_n$ solving the likelihood equation $\partial \bar{l}_n(\theta)/\partial \theta = 0$ such that $\hat{\theta}_n \to_{a.s.} \theta_0$.

**Lemma 4.8.** If the density $f_\theta$ is identifiable, then
\[
(4.4) \quad \bar{l}(\theta_0) > \bar{l}(\theta) \quad \forall \theta \neq \theta_0.
\]

**Proof.** Since log is a concave function, by Jensen’s inequality,
\[
E_{\theta_0} \left[ \log \frac{f_\theta(Y)}{f_{\theta_0}(Y)} \right] \leq \log E_{\theta_0} \left[ \frac{f_\theta(Y)}{f_{\theta_0}(Y)} \right] = \log \int \frac{f_\theta(y)}{f_{\theta_0}(y)} f_{\theta_0}(y) dy = 0.
\]

This implies
\[
E_{\theta_0} \left[ \log \frac{f_\theta(Y)}{f_{\theta_0}(Y)} \right] \leq 0
\]
and consequently,
\[
E_{\theta_0} [\log f_{\theta_0}(Y)] \geq E_{\theta_0} [\log f_\theta(Y)] \quad \forall \theta,
\]
Equality holds in the above if and only if $f_\theta = f_{\theta_0}$. Therefore, under the assumption of identifiability, the equality is strict. \qed
Proof of Theorem 4.7. By the strong law of large numbers (SLLN), it follows that
\[ \frac{\bar{l}_n(\theta)}{\sigma_n} \xrightarrow{a.s.} \bar{l}(\theta), \]
for \( \theta \in \{\theta_0 \pm \delta, \theta_0\}. \) Hence
\[
\lim_n \left( \bar{l}_n(\theta_0) - \bar{l}_n(\theta_0 - \delta) \right) \\
= \lim_n \left( \bar{l}_n(\theta_0) - \bar{l}(\theta_0) + \bar{l}(\theta_0) - \bar{l}(\theta_0 - \delta) + \lim_n \left( \bar{l}(\theta_0 - \delta) - \bar{l}_n(\theta_0 - \delta) \right) \right) \\
= \bar{l}(\theta_0) - \bar{l}(\theta_0 - \delta) > 0 \text{ a.s.}
\]
Similarly, we can show that \( \lim_n \left( \bar{l}_n(\theta_0) - \bar{l}_n(\theta_0 + \delta) \right) > 0 \) almost surely. By condition (3), we know almost surely there exists some \( \hat{\theta}_n \in [\theta_0 - \delta, \theta_0 + \delta] \) that solves \( \partial \bar{l}_n(\theta) / \partial \theta = 0 \) for \( n \) large enough. \( \Box \)

Next we study asymptotic normality of the MLE.

**Theorem 4.9.** Let \( Y_1, \ldots, Y_n \overset{iid}{\sim} f_{\theta_0}(y) \), where \( \theta_0 \) is an interior point of \( \Theta \) and \( f_\theta(y) \) satisfies the following regularity conditions:

(i) \( \frac{\partial^3}{\partial \theta^3} \log f_\theta \) exists locally at \( \theta_0 \), and there exists some \( \delta > 0 \) such that
\[
\sup_{|\theta - \theta_0| < \delta} \left| \frac{\partial^3}{\partial \theta^3} \log f_\theta(y) \right| \leq g(y) \text{ holds for some } g \text{ where } E_{\theta_0}g(Y) < \infty.
\]
(ii)
\[
- E_{\theta_0} \left( \frac{\partial^2}{\partial \theta^2} \log f_\theta(Y) \right) \bigg|_{\theta = \theta_0} = I_1(\theta_0).
\]

Let \( \hat{\theta}_n \) be the MLE of \( \theta_0 \) satisfying:
\[
\bar{l}_n'(\hat{\theta}_n) = \frac{1}{n} \sum_{i=1}^{n} \frac{\partial \log f_{\theta}(Y_i)}{\partial \theta} = 0.
\]

Then if \( \hat{\theta}_n \to_\theta \theta_0 \),
\[
\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} N(0, (I_1(\theta_0))^{-1})
\]
as \( n \to \infty \).

**Proof.** We use a Taylor series expansion of \( \bar{l}_n'(\hat{\theta}_n) \) around \( \theta_0 \):
\[
\bar{l}_n'(\hat{\theta}_n) = \bar{l}_n'(\theta_0) + \bar{l}_n''(\theta_0)(\hat{\theta}_n - \theta_0) + \frac{1}{2} \bar{l}_n'''(\theta^*)(\hat{\theta}_n - \theta_0)^2,
\]
where \( \theta^* \) is between \( \hat{\theta}_n \) and \( \theta_0 \). Substituting \( \bar{l}_n'(\hat{\theta}_n) = 0 \) from (4.6), and rearranging terms, we have that
\[
\hat{\theta}_n - \theta_0 = \frac{-\bar{l}_n'(\theta_0)}{\bar{l}_n''(\theta_0) + \frac{1}{2} \bar{l}_n'''(\theta^*)(\hat{\theta}_n - \theta_0)^2},
\]
which implies
\[
\sqrt{n}(\hat{\theta}_n - \theta_0) = \frac{-\sqrt{n}\bar{l}_n'(\theta_0)}{\bar{l}_n''(\theta_0) + \frac{1}{2} \bar{l}_n'''(\theta^*)(\hat{\theta}_n - \theta_0)^2}.
\]
Note that
\[ \sqrt{n} \bar{l}'_n(\theta_0) \xrightarrow{d} N(0, I_1(\theta_0)) \] by CLT,
\[ \bar{l}''_n(\theta_0) = \frac{1}{n} \sum_{i=1}^{n} \frac{\partial^2 \log f_\theta(Y_i)}{\partial^2 \theta} \bigg|_{\theta=\theta_0} \]
\[ \rightarrow_p E_{\theta_0} \left( \frac{\partial^2 \log f_\theta(Y_i)}{\partial^2 \theta} \bigg|_{\theta=\theta_0} \right) = -I_1(\theta_0), \quad \text{by WLLN}. \]
Furthermore, fix \( \varepsilon > 0 \) and choose \( \delta > 0 \) small enough,
\[ P_{\theta_0} \left( \left| \frac{1}{2} \bar{l}''(\theta^*) (\hat{\theta}_n - \theta_0) \right| > \varepsilon \right) \]
\[ \leq P_{\theta_0} \left( \sup_{\theta:|\theta-\theta_0| \leq \delta} \left| \frac{1}{2} \bar{l}''(\theta) \right| \delta > \varepsilon \right) + P_{\theta_0} (|\hat{\theta}_n - \theta_0| > \delta) \]
\[ \leq P_{\theta_0} \left( \frac{1}{2n} \sum_{i=1}^{n} g(Y_i) > \frac{\varepsilon}{\delta} \right) + P_{\theta_0} (|\hat{\theta}_n - \theta_0| > \delta) \xrightarrow{n \to \infty} 0. \]
as \( n \to \infty \) followed by \( \delta \to 0 \). This shows that
\[ \frac{1}{2} \bar{l}''(\theta^*) (\hat{\theta}_n - \theta_0) \xrightarrow{p} 0. \]
Consequently, applying Slutsky’s theorem, the result is established. \( \square \)

Remark 4.10. The asymptotic normality of MLE established through Theorem 4.9 has the following implications:

1. The MLE is asymptotically unbiased.
2. For large \( n \) the variance of the MLE can be approximated by \( (nI_1(\theta_0))^{-1} \), which is the CRLB for unbiased estimators of \( \theta \). Thus, \( \hat{\theta}_n \) has the smallest asymptotic variance among unbiased estimators, and hence is called asymptotically efficient.
5. General theory for M-estimation

5.1. General framework. The theory for the maximum likelihood estimator discussed in the previous section is only one example for a much larger class of estimation procedures. Suppose we observe $Y_1, \ldots, Y_n$, and an $M$-estimator $\hat{\theta}_n$ is defined by

$$
\hat{\theta}_n \equiv \arg \max_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^{n} m_\theta(Y_i),
$$

where $\{m_\theta : \theta \in \Theta\}$ is a class of criteria/loss functions indexed by the (finite-dimensional) parameter $\theta \in \Theta$.

In many cases, solving (5.1) can be reduced to a class of estimating equations by setting the derivative(s) to be zero:

$$
\frac{1}{n} \sum_{i=1}^{n} \psi_{\hat{\theta}_n}(Y_i) = 0.
$$

In particular, if $\theta \mapsto m_\theta(y)$ is differentiable for all $y$, then we may take $\psi_\theta = m_\theta'$. An estimator defined through (5.2) is called an $Z$-estimator.

Let us fix some notation that will be used frequently throughout this section. Let

$$
M_n(\theta) = \mathbb{P}_n m_\theta \equiv \frac{1}{n} \sum_{i=1}^{n} m_\theta(Y_i),
$$

$$
\Psi_n(\theta) = \mathbb{P}_n \psi_\theta = \frac{1}{n} \sum_{i=1}^{n} \psi_\theta(Y_i),
$$

and let

$$
G_n(f) \equiv \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (f(Y_i) - Pf)
$$

be the empirical process.

**Example 5.1.** Let $Y_1, \ldots, Y_n$ be i.i.d. with a common density $f_\theta$. The maximum likelihood estimator $\hat{\theta}_n$ studied in the previous section is included in the $M$-estimation framework by letting $m_\theta = \log f_\theta$:

$$
\hat{\theta}_n = \arg \max_{\theta \in \Theta} \sum_{i=1}^{n} \log f_\theta(Y_i).
$$

**Example 5.2.** Let $Y_1, \ldots, Y_n$ be i.i.d. samples from a common distribution, and we want to estimate the ‘location’ of this distribution. Here ‘location’ is a vague term, and can have different meanings:

- Solving zero for

$$
\sum_{i=1}^{n} (Y_i - \theta) = 0
$$
leads to the mean for the empirical distribution.

- Solving zero for
  \[ \sum_{i=1}^{n} \text{sign}(Y_i - \theta) = 0 \]
  leads to the median for the empirical distribution.

A general form for the ‘location’ estimator \( \hat{\theta}_n \) is given by the zero of the estimating equation
\[ \sum_{i=1}^{n} \psi(Y_i - \theta) = 0, \]
where \( \psi \) is some monotone and odd function. Here are some further choices:

- (Huber loss) \( \psi_k(y) = (y \wedge k) \vee (-k) \).
- (Quantile function) \( \psi_p(y) = -(1-p)1_{y<0} + p1_{y>0} \). Note that \( p = 1/2 \) corresponds to the median.

5.2. Consistency. First we consider consistency for M-estimators. Since \( \hat{\theta}_n = \arg \max_{\theta \in \Theta} M_n(\theta) \), if \( M_n \to M \) for some function \( M \) ‘nicely’, then we may naturally expect that
\[ \hat{\theta}_n = \arg \max_{\theta \in \Theta} M_n(\theta) \to \arg \max_{\theta \in \Theta} M(\theta) = \theta_0. \]
We make this heuristic argument rigorous below.

**Theorem 5.3.** Let \( M_n \) be random functions and \( M \) be a fixed function defined on a metric space \((\Theta, d)\). Suppose that for any \( \varepsilon > 0 \),

1. \( \sup_{\theta \in \Theta} |M_n(\theta) - M(\theta)| \to_p 0, \)
2. \( \sup_{\theta \neq \theta_0} d(\theta, \theta_0) = \varepsilon \Rightarrow M(\theta) < M(\theta_0). \)

Then \( \hat{\theta}_n \to_p \theta_0. \)

**Proof.** Since
\[ M_n(\hat{\theta}_n) = M_n(\theta_0) \]
by definition of \( \hat{\theta}_n \)
\[ = M(\theta_0) + o_p(1) \]
by (1),

it follows that
\[ M(\theta_0) - M(\hat{\theta}_n) \leq M_n(\hat{\theta}_n) - M(\theta_0) + o_p(1) \]
\[ \leq \sup_{\theta \in \Theta} |M_n(\theta) - M(\theta)| + o_p(1) \to_p 0. \]

The claim now follows from the assumption (2) (why?). \( \square \)

Next we consider Z-estimators. The results are broadly similar in spirit to the above theorem.

**Theorem 5.4.** Let \( \psi_n \) be random functions and \( \psi \) be a fixed function defined on a metric space \((\Theta, d)\). Suppose that for any \( \varepsilon > 0 \),

1. \( \sup_{\theta \in \Theta} \|\psi_n(\theta) - \psi(\theta)\| \to_p 0, \)


(2) \( \inf_{d(\theta, \theta_0) > \varepsilon} \| \psi(\theta) \| > 0 = \| \psi(\theta_0) \| . \)

Then \( \hat{\theta}_n \rightarrow_p \theta_0 . \)

**Proof.** Apply the previous theorem with \( M_{n}(\theta) = -\| \psi_{n}(\theta) \| \) and \( M(\theta) = -\| \psi(\theta) \| . \)

Condition (1) in the above theorems is usually an ‘empirical process condition’: for the \( M \)-estimation framework in the i.i.d. case, condition (1) boils down to

\[
\sup_{\theta \in \Theta} | (P_n - P) m_{\theta} | \rightarrow_p 0 .
\]

The class \( \{ m_{\theta} : \theta \in \Theta \} \) is called \( P \)-Glivenko-Cantelli if the above display holds true. The Glivenko-Cantelli property can typically be checked purely by the geometric properties of the class \( \{ m_{\theta} : \theta \in \Theta \} \), and has been well-known since the 1980s.

5.3. **Asymptotic normality.** For asymptotic normality results, we slightly reverse the order and start with the \( Z \)-estimation framework. Recall that a \( Z \)-estimator \( \hat{\theta}_n \) is given by the zero of the estimating equation

\[
\Psi_n(\hat{\theta}_n) = \frac{1}{n} \sum_{i=1}^{n} \psi_{\theta_n}(Y_i) = 0 .
\]

Using Taylor expansion, we have

\[
0 = \Psi_n(\hat{\theta}_n) = \Psi_n(\theta_0) + (\hat{\theta}_n - \theta_0)\Psi'_n(\theta_0) + \frac{1}{2}(\hat{\theta}_n - \theta_0)^2\Psi''_n(\tilde{\theta}_n),
\]

where \( \tilde{\theta}_n \) is some point in between \( \theta_0 \) and \( \hat{\theta}_n \). Therefore we may rewrite the above display as follows:

\[
\sqrt{n}(\hat{\theta}_n - \theta_0) = \frac{-\sqrt{n}\Psi_n(\theta_0)}{\Psi'_n(\theta_0) + \frac{1}{2}(\hat{\theta}_n - \theta_0)\Psi''_n(\tilde{\theta}_n)} .
\]

Since we expect

\[
\sqrt{n}\Psi_n(\theta_0) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi_{\theta_0}(Y_i) \approx N(0, P_{\psi_{\theta_0}}^2),
\]

\[
\Psi'_n(\theta_0) \rightarrow P_{\psi'_{\theta_0}},
\]

\[
\frac{1}{2}(\hat{\theta}_n - \theta_0)\Psi''_n(\tilde{\theta}_n) = o_p(1) \cdot O_p(1) = o_p(1),
\]

it is natural to expect the following result

\[
\sqrt{n}(\hat{\theta}_n - \theta_0) \rightarrow_d N(0, P_{\psi_{\theta_0}}^2/(P_{\psi'_{\theta_0}})^2).
\]

This is exactly the same development as we have seen previously in the case for the maximum likelihood estimator. One major drawback of this approach, however, is that although the asymptotic normality by itself only uses the first derivative of \( \psi_{\theta} \), the proof requires at least a second derivative
for $\psi$. Such a requirement is often too strong. Below we will use more modern (= ‘empirical process’) approach to alleviate such a strong smoothness requirement.

**Theorem 5.5.** Suppose that:

1. The map $\theta \mapsto P_{\psi_\theta}$ is differentiable at $\theta_0$ with non-singular derivative matrix $V_{\theta_0}$.
2. For any $\delta_n \to 0$, $\sup_{d(\theta, \theta_0) < \delta_n} |G_n(\psi_\theta - \psi_{\theta_0})| \to_p 0$.

If $\hat{\theta}_n \to_p \theta_0$, then

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = -V_{\theta_0}^{-1} \cdot \frac{1}{\sqrt{n}} \sum_{i=1}^n \psi_{\theta_0}(Y_i) + o_p(1)$$

$$\to_d N(0, V_{\theta_0}^{-1} P_{\psi_{\theta_0}} \psi_{\theta_0}^\top V_{\theta_0}^{-1}).$$

**Proof.** By consistency of $\hat{\theta}_n$ and (2), we have

$$o_p(1) = G_n \psi_{\hat{\theta}_n} - G_n \psi_{\theta_0}$$

$$= \sqrt{n}P(\psi_{\theta_0} - \psi_{\hat{\theta}_n}) - G_n \psi_{\theta_0}$$

(since $P_n \psi_{\hat{\theta}_n} = P_{\psi_{\theta_0}} = 0$)

$$= \sqrt{n}V_{\theta_0}(\theta_0 - \hat{\theta}_n) + \sqrt{n} \cdot o_p(||\theta_0 - \hat{\theta}_n||) - G_n \psi_{\theta_0}.$$

This implies

$$\sqrt{n}V_{\theta_0}(\theta_0 - \hat{\theta}_n) + o_p(\sqrt{n}||\theta_0 - \hat{\theta}_n||) = G_n \psi_{\theta_0} + o_p(1).$$

Note that if we can prove $\sqrt{n}||\theta_0 - \hat{\theta}_n|| = O_p(1)$, then (5.3) immediately yields that the claim of the theorem. To see this, by (5.3),

$$\sqrt{n}||\theta_0 - \hat{\theta}_n|| \leq ||V_{\theta_0}^{-1}|| \cdot ||\sqrt{n}V_{\theta_0}(\theta_0 - \hat{\theta}_n)|| \leq O_p(1) + o_p(\sqrt{n}||\theta_0 - \hat{\theta}_n||)$$

$$\Leftrightarrow (1 - o_p(1))\sqrt{n}||\theta_0 - \hat{\theta}_n|| = O_p(1),$$

as desired. \qed

Next consider the $M$-estimators.

**Theorem 5.6.** Suppose that:

1. The map $\theta \mapsto P_{m_\theta}$ admits a second-order Taylor expansion at $\theta_0$ with non-singular second derivative matrix (Hessian) $V_{\theta_0}$.
2. For any $M > 0$,

$$\sup_{||h|| \leq M} \left| G_n \sqrt{n}(m_{\theta_0} + h/\sqrt{n} - m_{\theta_0}) - h^\top m_{\theta_0} \right| \to_p 0.$$

If $\sqrt{n}(\hat{\theta}_n - \theta_0) = O_p(1)$, then

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \to_d N(0, V_{\theta_0}^{-1} P_{m_{\theta_0}} m_{\theta_0}^\top V_{\theta_0}^{-1}).$$
Proof. Condition (2) entails that for any $h_n = O_p(1)$,

$$o_p(1) = \mathbb{G}_n(\sqrt{n}(m_{\theta_0 + h_n/\sqrt{n}} - m_{\theta_0})) - \mathbb{G}_n(h_n^T m_{\theta_0})$$

$$= n^{\mathbb{P}}(m_{\theta_0 + h_n/\sqrt{n}} - m_{\theta_0}) - nP(m_{\theta_0 + h_n/\sqrt{n}} - m_{\theta_0}) - \mathbb{G}_n(h_n^T m_{\theta_0})$$

$$= n^{\mathbb{P}}(m_{\theta_0 + h_n/\sqrt{n}} - m_{\theta_0}) - \frac{1}{2}h_n^T V_{\theta_0} h_n + o(1) - \mathbb{G}_n(h_n^T m_{\theta_0}),$$

which implies that

$$n^{\mathbb{P}}(m_{\theta_0 + h_n/\sqrt{n}} - m_{\theta_0}) = \frac{1}{2}h_n^T V_{\theta_0} h_n + \mathbb{G}_n(h_n^T m_{\theta_0}) + o_p(1).$$

Now let $\hat{h}_n = \sqrt{n}(\hat{\theta}_n - \theta_0)$ and $\tilde{h}_n = -V_{\theta_0}^{-1}\mathbb{G}_n m_{\theta_0}$, we obtain

$$n^{\mathbb{P}}(m_{\theta_0 + h_n/\sqrt{n}} - m_{\theta_0}) = \frac{1}{2}h_n^T V_{\theta_0} h_n + \mathbb{G}_n(h_n^T m_{\theta_0}) + o_p(1)$$

$$n^{\mathbb{P}}(m_{\theta_0 + h_n/\sqrt{n}} - m_{\theta_0}) = -\frac{1}{2}(\mathbb{G}_n m_{\theta_0})^T V_{\theta_0}^{-1}(\mathbb{G}_n m_{\theta_0}) + o_p(1).$$

Using the definition of $\hat{h}_n$, we have

$$\frac{1}{2}h_n^T V_{\theta_0} h_n + \mathbb{G}_n(h_n^T m_{\theta_0}) + o_p(1)$$

$$= n^{\mathbb{P}}(m_{\theta_0 + h_n/\sqrt{n}} - m_{\theta_0})$$

$$\geq n^{\mathbb{P}}(m_{\theta_0 + h_n/\sqrt{n}} - m_{\theta_0})$$

$$= -\frac{1}{2}(\mathbb{G}_n m_{\theta_0})^T V_{\theta_0}^{-1}(\mathbb{G}_n m_{\theta_0}) + o_p(1),$$

which is equivalent to

$$\frac{1}{2}(\hat{h}_n + V_{\theta_0}^{-1}\mathbb{G}_n m_{\theta_0}) V_{\theta_0}(\hat{h}_n + V_{\theta_0}^{-1}\mathbb{G}_n m_{\theta_0}) + o_p(1) \geq 0.$$

This shows that

$$\hat{h}_n = \sqrt{n}(\hat{\theta}_n - \theta_0) = V_{\theta_0}^{-1}\mathbb{G}_n m_{\theta_0} + o_p(1),$$

completing the proof. 

Condition (2) in the above theorems is usually called ‘Donsker property’ for the underlying class of interest. One classical condition verifying such Donsker property is given by the Lipschitz condition:

- In the $Z$-estimation framework, we require the following: for any $\theta_1, \theta_2$ close enough to $\theta_0$, there exists some $\psi$ with $P\psi^2 < \infty$ such that

$$\|\psi_{\theta_1}(y) - \psi_{\theta_2}(y)\| \leq \psi(y)\|\theta_1 - \theta_2\|.$$

- In the $M$-estimation framework, we require the following: for any $\theta_1, \theta_2$ close enough to $\theta_0$, there exists some $\tilde{m}$ with $P\tilde{m}^2 < \infty$ such that

$$\|m_{\theta_1}(y) - m_{\theta_2}(y)\| \leq \tilde{m}(y)\|\theta_1 - \theta_2\|.$$
It is easy to check that this condition requires much less differentiability of the log-likelihood function in the context of maximum likelihood estimation.

**Example 5.7.** Let $Y_1, \ldots, Y_n$ be i.i.d. samples from an unknown distribution $F$, and we are interested in estimating the median $\theta_0$ of $F$. We assume that $F$ is differentiable at $\theta_0$ with positive derivative $f(\theta_0)$. Consider the sample median

$$\hat{\theta}_n = \arg \max_{\theta \in \mathbb{R}} \frac{1}{n} \sum_{i=1}^{n} |Y_i - \theta|.$$

We may apply Theorem 5.6 with $m_\theta(y) = -|y - \theta| + |y|$. By triangle inequality we may take $\dot{m} \equiv 1$. The map $\theta \mapsto m_\theta(y)$ is differentiable at $\theta_0$ except for $y = \theta_0$ with derivative $\dot{m}_{\theta_0}(y) = -\text{sign}(y - \theta_0)$. Furthermore, by a somewhat tedious calculation using integration by parts we have

$$Pm_\theta = -\int |x - \theta| \, dF(x) + \int |x| \, dF(x) = -2 \int_0^\theta F(x) \, dx + \theta.$$

This means that

$$\frac{d}{d\theta} Pm_\theta = -2F(\theta) + 1(=0 \text{ for } \theta = \theta_0),$$

$$\frac{d^2}{d\theta^2} Pm_\theta = -2f(\theta).$$

Hence Theorem 5.6 yields that

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \to_d N\left(0, 1/(4f^2(\theta_0))\right).$$

**Example 5.8.** Suppose we observe i.i.d. samples $(X_1, Y_1), \ldots, (X_n, Y_n)$ from the regression model

$$Y = f_{\theta_0}(X) + \xi,$$

where $E[\xi|X] = 0$ and $\{f_\theta : \theta \in \Theta\}$ is a family of regression functions. Let $\hat{\theta}_n$ be the least squares estimator defined by

$$\hat{\theta}_n = \arg \min_{\theta \in \Theta} \sum_{i=1}^{n} (Y_i - f_\theta(X_i))^2.$$

To fit the $M$-framework, we may take $m_\theta(x, y) = -(y - f_\theta(x))^2$, and we expect that under identifiability condition on $\{f_\theta : \theta \in \Theta\}$,

$$\hat{\theta}_n \to \arg \max_{\theta} Pm_\theta = \arg \min_{\theta} P(f_\theta - f_{\theta_0})^2 = \theta_0.$$

On the other hand,

$$Pm_\theta \approx P((\theta - \theta_0)^\top \dot{f}_{\theta_0})^2 + E\xi^2,$$
so we may expect $V_{\theta_0} = 2P\hat{\theta}_0 \hat{\theta}_0^\top$. Since $m_{\theta_0}(x, y) = -2(y - f_{\theta_0}(x))\dot{f}_{\theta_0}(x)$, we expect the asymptotic variance for $\sqrt{n}(\hat{\theta}_n - \theta_0)$ to be given by

$$
(2P\hat{\theta}_0 \hat{\theta}_0^\top)^{-1} P\xi^2 \hat{\theta}_0 \hat{\theta}_0^\top \cdot (2P\hat{\theta}_0 \hat{\theta}_0^\top)^{-1}
$$

$$
= \frac{1}{4} P\xi^2 \cdot (P\hat{\theta}_0 \hat{\theta}_0^\top)^{-1}
$$

if $X$ is independent of $\xi$.

The distribution convergence follows by additional complexity condition on the class $\{f_\theta : \theta \in \Theta\}$ but we shall not go into technical details for this.
6. Delta method

6.1. Delta method. Suppose \( \hat{\theta}_n \) is an estimator of \( \theta_0 \), and we are interested in the quantity \( \phi(\theta_0) \). A natural plug-in estimator is given by \( \phi(\hat{\theta}_n) \), but the question is, how does \( \phi(\hat{\theta}_n) \) perform as an estimator of \( \theta_0 \)?

The following theorem, known as the delta method, quantifies the distributional performance of the estimator \( \phi(\hat{\theta}_n) \) in terms of the performance of \( \hat{\theta}_n \) itself.

**Theorem 6.1.** Let \( \phi : \mathbb{R}^k \to \mathbb{R}^m \) be differentiable at \( \theta_0 \) in the sense that there exists \( \phi'_\theta \in \mathbb{R}^{m \times k} \) such that

\[
\phi(\theta_0 + h) = \phi(\theta_0) + \phi'_\theta \cdot h + o(\|h\|), \quad \text{as } \|h\| \to 0.
\]

If \( r_n(\hat{\theta}_n - \theta_0) \to_d Z \) for some \( r_n \to \infty \) and some random vector \( Z \), then

\[
r_n(\phi(\hat{\theta}_n) - \phi(\theta_0)) \to_d \phi'_\theta \cdot Z.
\]

**Proof.** The proof is basically Taylor expansion, but we include here some details. Note that

\[
r_n(\phi(\hat{\theta}_n) - \phi(\theta_0)) = r_n\phi'_\theta(\hat{\theta}_n - \theta_0) + r_n o(\|\hat{\theta}_n - \theta_0\|)
= \phi'_\theta(r_n(\hat{\theta}_n - \theta_0)) + o_p(1)
\to_d \phi'_\theta \cdot Z,
\]

as desired. \( \square \)

**Example 6.2.** Let \( Y_1, \ldots, Y_n \) be i.i.d. samples. Let

\[
\bar{S}_n^2 \equiv \frac{1}{n} \sum_{i=1}^n (Y_i - \bar{Y})^2
\]

be the sample variance (modulo the fact that we use the weight \( n \) rather than the usual \( n - 1 \), but this does not matter in asymptotics). Let \( \phi(x, y) = y - x^2 \). Then we may write \( \bar{S}_n^2 = \phi(\bar{Y}, \bar{Y}^2) \). This helps us to derive limit distribution for \( \bar{S}_n \). To this end, let \( \alpha_k = EY_1^k \). By central limit theorem, we have

\[
\sqrt{n} \left( \left( \bar{Y} - \frac{\alpha_1}{\alpha_2} \right) \right) \to_d N \left( \left( \begin{array}{c} 0 \\ 0 \end{array} \right), \left( \begin{array}{cc} \alpha_2 - \alpha_1^2 & \alpha_3 - \alpha_1 \alpha_2 \\ \alpha_3 - \alpha_1 \alpha_2 & \alpha_4 - \alpha_2^2 \end{array} \right) \right).
\]

It is easy to find the derivative of the map \( \phi \), which is given by \( \phi'(\alpha_1, \alpha_2) = (-2\alpha_1, 1)^\top \). Then by delta method, the asymptotic variance for \( \sqrt{n}(\phi(\bar{Y}, \bar{Y}^2) - \phi(\alpha_1, \alpha_2)) \) is given by

\[
\begin{pmatrix} -2\alpha_1 & 1 \\ \alpha_3 - \alpha_1 \alpha_2 & \alpha_4 - \alpha_2^2 \end{pmatrix}
\begin{pmatrix} \alpha_2 - \alpha_1^2 & \alpha_3 - \alpha_1 \alpha_2 \\ \alpha_3 - \alpha_1 \alpha_2 & \alpha_4 - \alpha_2^2 \end{pmatrix}
\begin{pmatrix} -2\alpha_1 \\ 1 \end{pmatrix}
= \alpha_4 - \alpha_2^2 \text{ if } \alpha_1 = 0.
\]

Hence by shift invariance, with \( \mu_k = E(Y_1 - EY_1)^k \),

\[
\sqrt{n}(\bar{S}_n^2 - \mu_2) \to_d N(0, \mu_4 - \mu_2^2).
\]
6.2. **Variance stabilizing.** Asymptotic results like the ones we discussed above are crucial for building confidence intervals or testing hypotheses about the parameters (to be discussed later), but typically suffer from the following problem: the asymptotic variance is a function of the unknown parameter in the sense that

\[
\sqrt{n}(\hat{\theta}_n - \theta_0) \rightarrow_d N(0, \sigma^2(\theta_0)).
\]

Therefore, we either need to use a naive plug-in approach (i.e., substituting MLEs into the variance expressions) or use a variance stabilizing approach: transform the parameter such that the asymptotic variance is free of the unknown parameter (i.e. pivotal). The basic idea for finding such a transform is as follows: by delta method,

\[
\sqrt{n}(\phi(\hat{\theta}_n) - \phi(\theta_0)) \rightarrow_d N(0, (\phi'(\theta_0))^2 \sigma^2(\theta_0)).
\]

So we may take

\[
\phi(\theta) = \int \frac{1}{\sigma(\theta)} \, d\theta,
\]

so that the asymptotic distribution of \(\sqrt{n}(\phi(\hat{\theta}_n) - \phi(\theta_0))\) is pivotal.

**Example 6.3.** Let \((X_1, Y_1), \ldots, (X_n, Y_n)\) be i.i.d. bivariate normally distributed with correlation \(\rho\). Then the sample correlation

\[
r_n = \frac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n} (X_i - \bar{X})^2 \sum_{i=1}^{n} (Y_i - \bar{Y})^2}}
\]

satisfies

\[
\sqrt{n}(r_n - \rho) \rightarrow_d N(0, (1 - \rho^2)^2).
\]

The variance of the limiting normal distribution unfortunately depends on the unknown correlation \(\rho\) so we cannot directly use \(r_n\) to conduct inference on \(\rho\). Following the variance stabilizing idea above, we may consider the following transform:

\[
\phi(\rho) = \int \frac{1}{1 - \rho^2} \, d\rho = \frac{1}{2} \log \frac{1 + \rho}{1 - \rho} = \text{arctanh} \rho.
\]

6.3. **Asymptotic normality of moment estimators.** Consider the following general form of moment estimators:

\[
E_{\hat{\theta}_n} f_j(Y) = \frac{1}{n} \sum_{i=1}^{n} f_j(Y_i), \quad 1 \leq j \leq k.
\]

The classical method of moment estimator we have seen in Section 4.1 corresponds to the case \(f_j(y) = y^j\).

For notational simplicity, we write \(f = (f_1, \ldots, f_k)\).
Theorem 6.4. Suppose that the map \( e(\theta) = P_\theta f \) is one-to-one on an open set \( \Theta \subset \mathbb{R}^k \), and is continuously differentiable at \( \theta_0 \) with non-singular derivative matrix \( e'_\theta \). Further assume that \( P_{\theta_0}\|f\|^2 < \infty \). Then \( \hat{\theta}_n \) exists with probability tending to one, and
\[
\sqrt{n}(\hat{\theta}_n - \theta_0) \to_d N\left(0, (e'_\theta)_{\theta_0}^{-1}(\text{Cov}_{\theta_0}f)(e'_\theta)_{\theta_0}^{-\top}\right).
\]

Proof. By inverse function theorem, there exists an open set \( U \subset \mathbb{R}^k \) containing \( \theta_0 \), and an open set \( V \subset \mathbb{R}^k \) containing \( P_{\theta_0}f \) such that \( e : U \to V \) is \( C^1 \) homeomorphism. So \( \hat{\theta}_n = e^{-1}(P_n f) \) as soon as \( P_n f \in V \), which happens with probability tending to 1 according to the weak law of large numbers. The claim of the theorem now follows directly from the delta method. \( \square \)
7. Efficiency lower bound: asymptotic theory

7.1. Hodge’s estimator and superefficiency. From the Cramér-Rao lower bound, we know that if an estimator $T_n$ is unbiased for estimating $g(\theta)$, then we necessarily have

$$\text{Var}_\theta(T_n) \geq (g'(\theta))^2/I(\theta),$$

for all $\theta \in \mathbb{R}$. The theory for MLE tells us that under sufficient regularity conditions, the MLE will asymptotically achieve the CR lower bound. It is therefore very natural to expect that we may be able to prove a statement of the following type: suppose $\sqrt{n}(T_n - g(\theta)) \rightarrow_d L_\theta$ under $\theta$, then

$$\text{Var}_\theta(L_\theta) \geq (g'(\theta))^2/I(\theta),$$

for all $\theta \in \mathbb{R}$. However, as we will see very soon from the following famous counter-examples that such a statement is not true.

Let $Y_1, \ldots, Y_n$ be i.i.d. observations from the normal location family $N(\theta, 1)$ (so the Fisher information is simply $I(\theta) \equiv 1$. We know that the sample mean is asymptotically efficient:

$$\sqrt{n}(\bar{Y} - \theta) \rightarrow_d N(0, 1).$$

Now consider the following Hodge’s estimator

$$T_n = \begin{cases} \bar{Y} & \text{if } |ar{Y}| > n^{1/4}, \\ a\bar{Y} & \text{if } |ar{Y}| \leq n^{1/4}. \end{cases}$$

Then the asymptotic distribution of $T_n$ is given by the following

(7.1) $$\sqrt{n}(T_n - \theta) \rightarrow_d \begin{cases} N(0, 1) & \text{if } \theta \neq 0, \\ N(0, a^2) & \text{if } \theta = 0. \end{cases}$$

Note that when the truth is $\theta = 0$, the limiting variance of $T_n$ can be arbitrarily small! This would imply that a naive asymptotic version of the CR lower bound described above will not work.

First let us prove (7.1). To see this, note that

$$\sqrt{n}(T_n - \theta) = \sqrt{n}(\bar{Y} - \theta)1_{|\bar{Y}| > n^{-1/4}} + \sqrt{n}(a\bar{Y} - \theta)1_{|\bar{Y}| \leq n^{-1/4}}$$

$$= \sqrt{n}(\bar{Y} - \theta)1_{\sqrt{n}|\bar{Y}-\theta| > n^{1/4}}$$

$$+ \left[a\sqrt{n}(\bar{Y} - \theta) + \sqrt{n}\theta(a-1)\right]1_{\sqrt{n}|\bar{Y}-\theta| \leq n^{1/4}}$$

$$= \sqrt{n}(\bar{Y} - \theta) + \sqrt{n}\theta(a-1)1_{\sqrt{n}|\bar{Y}-\theta| \leq n^{1/4}}$$

$$\rightarrow_d \begin{cases} Z & \text{if } \theta \neq 0, \\ aZ & \text{if } \theta = 0. \end{cases}$$

Here $Z \sim N(0, 1)$ and the proof for (7.1) is complete.

One may naturally wonder if the Hodge’s estimator indeed improves the sample mean because we know a lot of optimality about the sample mean in the normal location model. The problem here is that although Hodge's
estimator improves the efficiency at one point \( \theta = 0 \), it behaves much more poorly in a local \( n^{-1/2} \)-ball in the following sense: Let \( \theta_n = cn^{-1/2} \), then
\[
\sqrt{n}(T_n - \theta_n) \approx \theta_n Z_{|Z+c| > n^{1/4}} + (aZ + c(a-1))1_{|Z+c| \leq n^{1/4}} \\
\sim \theta_n Z + c(a-1) \sim N(c(a-1), a^2).
\]
So in the local perturbation of radii \( n^{-1/2} \), the Hodge’s estimator pays a huge price in the sense it is asymptotically biased at level \( n^{-1/2} \) (depending on \( c \)). This amount to saying that although Hodge’s estimator is super-efficient at 0, the price for this is that its behavior is highly irregular against \( n^{-1/2} \)-perturbation.

The major goal of this section is to show that the asymptotic version of CR lower bound holds as long as we exclude counter-examples like Hodge’s estimator which is not locally regular in a sense to be defined precisely later on.

7.2. Le Cam’s contiguity theory. The major question for Le Cam’s contiguity theory is the following: suppose we know that the estimators \( T_n \) converge in distribution under \( P_n \) to \( T \), then what can we say about the asymptotic distribution of \( T_n \) under another sequence of probability measures \( Q_n \)?

We first some basic facts if \( P_n = P \) and \( Q_n = Q \), where \( P, Q \) are defined on the same measurable space \( (\Omega, \mathcal{A}) \) with density \( p, q \) with respect to the dominating measure \( \mu \). We say that \( Q \) is absolutely continuous with respect to \( P \), notationally \( Q \ll P \) if \( P(A) = 0 \) implies \( Q(A) = 0 \). Furthermore, we may write \( dQ/dP \) to be the Radon-Nikodym derivative so
\[
E_Q f(X) = E_P f(X) \frac{dQ}{dP}.
\]
As a minor technicality we may allow \( Q \) not necessarily absolutely continuous with respect to \( P \) and define \( dQ/dP \) by only considering the ‘absolute continuous’ component of \( Q \) with respect to \( P \) in the Lebesgue decomposition, but for technical convenience we will assume \( P_n \) and \( Q_n \) are mutually absolutely continuous in the sequel.

Now we want to consider the asymptotic version where allowing \( P_n \) and \( Q_n \) changing with \( n \).

**Definition 7.1.** Let \( P_n, Q_n \) be probability measures defined on \( (\Omega_n, \mathcal{A}_n) \). We say that \( \{Q_n\} \) is contiguous with respect to \( \{P_n\} \), notationally \( Q_n \ll P_n \) if and only if any \( P_n(A_n) \to 0 \) implies \( Q_n(A_n) \to 0 \). We say that \( \{P_n\}, \{Q_n\} \), notationally \( P_n \ll Q_n \) and \( Q_n \ll P_n \).

**Theorem 7.2** (Le Cam’s first lemma). The following are equivalent.

1. \( Q_n \ll P_n \).
2. If \( \frac{dP_n}{dQ_n} \sim P_n U \) along a subsequence, then \( P(U > 0) = 1 \).
3. If \( \frac{dQ_n}{dP_n} \sim P_n V \) along a subsequence, then \( E V = 1 \).
4. For any statistics \( T_n : \Omega_n \to \mathbb{R}^k \) if \( T_n \to P_n 0 \), then \( T_n \to Q_n 0 \).
Proof. \((1) \Rightarrow (4)\). Simply let \(A_n = \{ \| T_n \| > \epsilon \} \).

\((4) \Rightarrow (1)\). Let \(T_n = 1_{A_n} \).

\((1) \Rightarrow (2)\). For notational convenience we simply use \(\{n\}\) for the subsequence. For any \(\epsilon > 0\),

\[
g_n(\epsilon) \equiv Q_n \left( \frac{dP_n}{dQ_n} < \epsilon \right) - P(U < \epsilon).
\]

By Portmanteau’s theorem, we have

\[
\liminf_n g_n(\epsilon) \geq 0, \\forall \epsilon > 0.
\]

This means that we can find a slowly decreasing \(\epsilon_n \downarrow 0\) such that

\[
\liminf_n g_n(\epsilon_n) \geq 0.
\]

Hence

\[
P(U = 0) = \lim_n P(U < \epsilon_n) \leq \liminf_n Q_n \left( \frac{dP_n}{dQ_n} < \epsilon_n \right).
\]

On the other hand,

\[
P_n \left( \frac{dP_n}{dQ_n} < \epsilon_n \right) = \int_{\frac{dP_n}{dQ_n} < \epsilon_n} \frac{dP_n}{dQ_n} dQ_n \leq \int \epsilon_n dQ_n \to 0.
\]

By contiguity assumed in \((1)\), we conclude that

\[
Q_n \left( \frac{dP_n}{dQ_n} < \epsilon_n \right) \to 0,
\]

and hence \(P(U = 0) = 0\).

\((3) \Rightarrow (1)\). Suppose \(P_n(A_n) \to 0\), and we want to prove \(Q_n(A_n) \to 0\). Then

\[
\left( \frac{dQ_n}{dP_n} 1_{\Omega_n \setminus A_n} \right) \to_{P_n} (V, 1).
\]

Since convergence in probability implies convergence in distribution, Portmanteau’s theorem applies to see that

\[
\liminf_n Q_n(\Omega_n \setminus A_n) = \liminf_n \int 1_{\Omega_n \setminus A_n} \frac{dQ_n}{dP_n} dP_n \geq E1 \cdot V = 1,
\]

where the last equality uses \((3)\). This means that \(Q(A_n) \to 0\), proving the claim.

\((2) \Rightarrow (3)\). Let \(\mu_n \equiv \frac{1}{2}(P_n + Q_n)\) (so it dominates both \(P_n\) and \(Q_n\)). By Prohorov’s theorem, there exists a subsequence, simply denoted \(\{n\}\), such that there exists some random variables \(U, V, W\),

\[
\frac{dP_n}{dQ_n} \Rightarrow_{Q_n} U, \quad \frac{dQ_n}{dP_n} \Rightarrow_{P_n} V, \quad W \equiv \frac{dP_n}{d\mu_n} \Rightarrow_{\mu_n} W.
\]
The last one follows by the fact that the density of $dP_n/d\mu_n$ under $\mu_n$ is uniformly bounded so uniform tightness automatically follows. Now it is easy to see that
\[
\frac{dP_n}{dQ_n} = \frac{W_n}{2-W_n}, \quad \frac{dQ_n}{d\mu_n} = 2-W_n, \quad EW = 1.
\]
For any continuous and bounded $f$, define
\[
g(w) \equiv f\left(\frac{w}{2-w}\right)(2-w), \quad \forall 0 \leq w < 2,
\]
and $g(2) = 0$. Then $g$ is continuous and bounded on $[0, 2]$. By Portmanteau's theorem,
\[
Ef(U) = \lim_n E_{Q_n} f\left(\frac{dP_n}{dQ_n}\right) = \lim_n E_{\mu_n} f\left(\frac{dP_n}{dQ_n}\right)\frac{dQ_n}{d\mu_n}
\]
\[
= \lim_n E_{\mu_n} f\left(\frac{W_n}{2-W_n}\right)(2-W_n)
\]
\[
= \lim_n E_{\mu_n} g(W_n) = g(W) = Ef\left(\frac{W}{2-W}\right)(2-W).
\]
Take a sequence $f_m \leq 1$ such that $f_m \downarrow 10$ pointwise. Then by dominated convergence theorem,
\[
P(U = 0) = \lim_m Ef_m(U) = \lim_m Ef_m\left(\frac{W}{2-W}\right)(2-W)
\]
\[
= E1_0\left(\frac{W}{2-W}\right)(2-W) = 2P(W = 0).
\]
Similarly we may show that
\[
Ef(V) = Ef\left(\frac{2-W}{W}\right)W
\]
holds for all continuous and bounded function. Take $0 \leq f_m \uparrow id$ pointwise. Then by monotone convergence theorem,
\[
EV = E\left(\frac{2-W}{W}\right)W = E(2-W)1_{W>0} = 2P(W > 0) - EW
\]
\[
= 2P(W > 0) - 1.
\]
Combining (7.2) and (7.3) we see that
\[
P(U = 0) + EV = 2(P(W = 0) + P(W > 0)) - 1 = 1.
\]
The claim of (3) follows by the assumption of (2). \qed

The following example will be repeated used later on.

**Example 7.3.** Let $\{P_n\}, \{Q_n\}$ be such that
\[
\frac{dP_n}{dQ_n} \overset{\sim}{\underset{Q_n}{\to}} e^{N(\mu, \sigma^2)}.
\]
Then by (2) of Le Cam’s first lemma, \( Q_n \ll P_n \). Furthermore, by (3), \( Q_n \ll \|P_n \) iff \( \mu = -\frac{1}{2} \sigma^2 \) (since \( E \exp(N(\mu, \sigma^2)) = 1 \) iff \( \mu = -\frac{1}{2} \sigma^2 \)).

From Le Cam’s first lemma, we immediately get the following generalized version of Le Cam’s third lemma.

**Theorem 7.4.** Let \( P_n, Q_n \) be probability measures defined on \( (\Omega_n, A_n) \) and \( Q_n \ll P_n \). Let \( X_n : \Omega_n \to \mathbb{R}^k \) be random vectors. Suppose that

\[
\left( X_n, \frac{dQ_n}{dP_n} \right) \rightsquigarrow P_n \left( X, V \right).
\]

Then \( L(B) \equiv E[1_B(X) \cdot V] \) defines a probability measure and \( X_n \rightsquigarrow Q_n L \).

**Proof.** By Le Cam’s first lemma, \( EV = 1 \) and hence \( L \) is a probability measure. By a standard approximation argument we see that

\[
E_L f(X) = \int f \, dL = E[f(X)V].
\]

Take continuous and non-negative \( f \). By Portmanteau’s theorem,

\[
\lim \inf E_{Q_n} f(X_n) = \lim \inf \int f(X_n)\frac{dQ_n}{dP_n} \, dP_n \geq E[f(X)V] = E_L f(X),
\]

as desired. \( \square \)

**Corollary 7.5.** Suppose that

\[
\left( X_n, \log \frac{dQ_n}{dP_n} \right) \rightsquigarrow P_n \left( N_{k+1} \left( \left( \begin{array}{c} \mu \\ -\frac{1}{2} \sigma^2 \end{array} \right), \left( \begin{array}{cc} \Sigma & \tau \\ \tau^\top & \sigma^2 \end{array} \right) \right) \right).
\]

Then

\[
X_n \rightsquigarrow Q_n \left( N_k(\mu + \tau, \Sigma) \right).
\]

**Proof.** Let \( (X, W) \) equal to the limiting distribution given by the assumption. Then by continuous mapping theorem,

\[
\left( X, \frac{dQ_n}{dP_n} \right) \rightsquigarrow P_n \left( X, e^W \right).
\]

Since \( e^W = d \exp(N(-\sigma^2/2, \sigma^2)) \), we see that \( P_n \ll Q_n \). By the above theorem, we know that \( X_n \rightsquigarrow Q_n L \) with \( L(B) = E1_B(X)e^W \). The characteristic function for \( L \) is

\[
\int \exp(i t \cdot x) \, dL(x) = E e^{i t \cdot x} e^W = E e^{i(t-i)(X+W)\top} = \exp \left( i t \cdot \mu - \frac{\sigma^2}{2} t - \frac{1}{2} (t-i)\top \left( \begin{array}{cc} \Sigma & \tau \\ \tau^\top & \sigma^2 \end{array} \right) \left( \begin{array}{c} t \\ -i \end{array} \right) \right) = \ldots = e^{i t \cdot (\mu + \tau) - \frac{1}{2} t\top \Sigma t},
\]

as desired. \( \square \)
7.3. **Convergence to normal models.** The idea to formulate an asymptotic version of the CR lower bound is to map the statistical model in hand to its limiting Gaussian model. In particular, since the difficulty of the estimation lies in a $n^{-1/2}$ neighborhood of the truth, it is very natural to approximate the empirical model

$$(Y_1, \ldots, Y_n) \sim \left( P_{\theta_0 + h/\sqrt{n}} = \otimes^n P_{\theta_0 + h/\sqrt{n}} : h \in \mathbb{R}^k \right)$$

by that of the limiting normal location model

$$Y \sim \left( N(h, I_\theta^{-1}) : h \in \mathbb{R}^k \right)$$

where $I_\theta$ is the Fisher information of the model. To do this, we need

- a notion quantifying the ‘normality’ of the empirical model,
- an operation that maps the estimator in the empirical model to an estimator in the limiting normal location model.

Once these steps are done, we can basically analyze the lower bound problem in the limiting Gaussian model.

Now we start with an appropriate notion for normality for the empirical model.

**Definition 7.6.** We say $(P_\theta : \theta \in \mathbb{R}^k)$ is local asymptotically normal (LAN) iff there exists some $\hat{\ell}_\theta$ for which $P_{\hat{\ell}_\theta} = 0$ and well-defined Fisher information $I_\theta$ such that

$$\log \frac{dP_{\theta_0 + h/\sqrt{n}}}{dP_{\theta_0}} = G_n(h^\top \hat{\ell}_\theta) - \frac{1}{2} h^\top I_\theta h + o_{P_\theta}(1).$$

If the above $o_{P_\theta}(1)$ is uniform for all $\|h\| \leq M$ for any $M > 0$, then we say that $(P_\theta : \theta \in \mathbb{R}^k)$ is uniformly LAN.

Here $I_\theta$ is used as abbreviation for $I_1(\theta)$, but in a multidimensional setting. Recall that $G_n(f) = \sqrt{n} \left( P_n - P \right)(f) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \left( f(Y_i) - Pf \right)$ is the empirical process. To see why this definition is natural, consider the log likelihood ratio process for the limiting Gaussian models:

$$\log \frac{dN(h, I_\theta^{-1})}{dN(0, I_\theta^{-1})} = G_{P_\theta}(h^\top \hat{\ell}_\theta) - \frac{1}{2} h^\top I_\theta h,$$

where $G_{P_\theta}$ is the limiting Brownian-bridge process (for here it can be simply understood as a centered normal random variable with variance $h^\top I_\theta h$).

Now it is clear that our definition of LAN requires the model $(P_\theta : \theta \in \mathbb{R}^k)$ is almost its Gaussian counterpart. This is true in many cases. In particular, it is easy to check by Taylor expansion that if the densities $f_\theta$ corresponding to $P_\theta$ is sufficiently differentiable (e.g. twice differentiable), then LAN holds. As we have already seen in the $M$-estimation theory, requiring differentiability on the likelihood is often too restrictive. Below is a classical condition for LAN that considerably allows much less differentiability than doing a Taylor expansion.
Theorem 7.7. Suppose that $\Theta$ is an open set of $\mathbb{R}^k$ and that $\{P_\theta : \theta \in \Theta\}$ is quadratically differentiable at $\theta$ in the sense that
\[
\int \left( \sqrt{f_{\theta+h}} - \sqrt{f_\theta} - \frac{1}{2} h^\top \ell_\theta \sqrt{f_\theta} \right)^2 d\mu = o(\|h\|^2)
\]
as $\|h\| \to 0$, for some $\ell_\theta$. Here $f_\theta$ is the density of $P_\theta$ with respect to a dominating measure $\mu$. Then $P_\theta \ell_\theta = 0$ and $I_\theta = P_\theta \ell_\theta^\top$ exists, and $\{P_\theta : \theta \in \Theta\}$ is LAN.

Proof. Write $f_n = f_{\theta+h}/\sqrt{n}$, $f = f_\theta$ and $g = h^\top \ell_\theta$ for notational convenience.

We first prove that $P_\theta \ell_\theta = 0$. By quadratic differentiability, we have
\[
\sqrt{n}(\sqrt{f_n} - \sqrt{f}) \to \frac{1}{2} g \sqrt{f} \text{ in } L_2(\mu).
\]
This in particular implies that $\sqrt{f_n} \to \sqrt{f}$ in $L_2(\mu)$. Hence
\[
P_\theta h^\top \ell_\theta = P_\theta g = \int \frac{1}{2} g \sqrt{f} \cdot 2 \sqrt{f} d\mu
\]
\[
= \lim_n \int \sqrt{n}(\sqrt{f_n} - \sqrt{f})(\sqrt{f_n} + \sqrt{f}) d\mu
\]
\[
= \lim_n \sqrt{n} \left( \int f_n - \int f \right) = \lim_n \sqrt{n}(1 - 1) = 0
\]
holds for all $h$. So we have proved $P_\theta \ell_\theta = 0$.

That $I_\theta$ is well-defined is trivial. Now we prove that $\{P_\theta : \theta \in \Theta\}$ is LAN. Let
\[
W_{ni} = 2 \left[ \sqrt{f_n}(Y_i) - 1 \right].
\]
Then using the expansion $\log(1 + x) = x - x^2/2 + o(|x|^2)$, the log-likelihood ratio can be written as
\[
\log \prod_{i=1}^n \frac{f_n}{f}(Y_i) = 2 \sum_{i=1}^n \log \left( 1 + \frac{1}{2} W_{ni} \right)
\]
\[
= \sum_{i=1}^n W_{ni} - \frac{1}{4} \sum_{i=1}^n W_{ni}^2 + o \left( \sum_{i=1}^n W_{ni}^2 \right)
\]
\[
= (I) + (II) + \text{Remainder}.
\]
For $(I)$, first note that
\[
E_\theta \sum_{i=1}^n W_{ni} = 2n \left( \int \sqrt{f_n} \sqrt{f} d\mu - 1 \right)
\]
\[
= - \int (\sqrt{n}(\sqrt{f_n} - \sqrt{f}))^2 d\mu \to -\frac{1}{4} P_\theta g^2.
\]
The convergence in the above display follows from (7.4) (i.e. quadratic differentiability). On the other hand,

\[
\text{Var}_\theta \left( \frac{1}{\sqrt{n}} \sum_{i=1}^{n} W_{ni} - 1 \right) = \sum_{i=1}^{n} \text{Var}_\theta \left( \frac{1}{\sqrt{n}} g(Y_i) \right)
\]

\[
= \text{Var}_\theta \left( \sqrt{n} W_{n1} - g(Y_1) \right)
\]

\[
\leq E_\theta \left( \sqrt{n} W_{n1} - g(Y_1) \right)^2 \to 0
\]

where the convergence follows from quadratic differentiability. Combining the above two displays, we find

\[
(I) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} g(Y_i) - \frac{1}{4} P_\theta g^2 + o_P(1).
\]

For (II), note that by the arguments in (7.6), we have

\[
n W_{ni}^2 = g^2(Y_i) + A_{ni}
\]

where \( E_\theta |A_{ni}| \to 0 \) uniformly in \( i \). This means

\[
(II) = -\frac{1}{4} \sum_{i=1}^{n} W_{ni}^2 = -\frac{1}{4n} \sum_{i=1}^{n} g^2(Y_i) - \frac{1}{4n} \sum_{i=1}^{n} A_{ni} = -\frac{1}{4} P_\theta g^2 + o_P(1).
\]

Combining (7.5), (7.7) and (7.8) we obtain the claim of the theorem. \( \square \)

Now we turn to the second task. First we need one technical notion.

**Definition 7.8.** Let \( Y \) be an observation. We say that \( T \) is randomized statistics based on \( Y \) iff \( T = T(Y, U) \) where \( U \) is uniformly distributed on \([0,1]\) independent of \( Y \).

**Theorem 7.9.** Suppose that \((P_\theta : \theta \in \Theta)\) is LAN with non-singular Fisher information matrix \( I_\theta \). Let \( \{T_n\} \) be statistics in \((P^n_{\theta+h/\sqrt{n}} : h \in \mathbb{R}^k)\) such that \( T_n \) converges in distribution under \( P^n_{\theta+h/\sqrt{n}} \). Then there exists a randomized statistic \( T \) in \((N(h, I_\theta^{-1}) : h \in \mathbb{R}^k)\) such that \( T_n \rightsquigarrow P^n_{\theta+h/\sqrt{n}} T \) for all \( h \in \mathbb{R}^k \).

**Proof.** Let \( P_{n,h} \equiv P^n_{\theta+h/\sqrt{n}} \) and \( \Delta_n = \frac{1}{n} \sum_{i=1}^{n} \hat{e}_\theta(Y_i) \). Then \((T_n, \Delta_n)\) is uniformly tight under \( P_{n,0} \) and hence by Prohorov’s theorem there exists a subsequence \( \{n\} \) and random vector \( (S, \Delta) \) such that

\[
(T_n, \Delta_n) \rightsquigarrow P_{n,0} (S, \Delta).
\]

In particular, this means that marginally \( \Delta \sim N(0, I_\theta) \). On the other hand, by LAN we have

\[
\left( T_n, \log \frac{dP_{n,h}}{dP_{n,0}} \right) \rightsquigarrow P_{n,0} \left( S, h^\top \Delta - \frac{1}{2} h^\top I_\theta h \right).
\]
By Example 7.3, we have $P_{n,h} \ll P_{n,0}$. Now by Le Cam's third lemma, the probability measure $L_h$ corresponding to the law of the limit of $T_n$ under $P^*_{n,h}$ is given by

$$L_h(B) = E_1 B(S)e^{h^\top \Delta - \frac{1}{2}h^\top I_\theta h}.$$ 

It is possible to construct $T(\Delta, U)$ such that $(T(\Delta, U), \Delta) =_d (S, \Delta)$ (so we may define $T$ by requiring $T(\delta, U) =_d S|\Delta = \delta$, which can be done by quantile transformation). We want to verify that such constructed $T$ has distribution $L_h$ under $P_h = dN(h, I_\theta^{-1})$. Let $Y$ be an observation in $(N(h, I_\theta^{-1} : h \in \mathbb{R}^k))$. Then $I_\theta Y \sim_0 N(0, I_\theta) =_d \Delta$. Hence for any Borel set $B$,

$$P_h(T(I_\theta Y, U) \in B)$$

$$= E_U \int 1(T(I_\theta y, U) \in B) \cdot e^{-\frac{1}{2}(y-h)^\top I_\theta (y-h)} \sqrt{\frac{|I_\theta|}{(2\pi)^{k}}} \, dy$$

$$= E_U \int 1(T(I_\theta y, U) \in B) \cdot e^{h^\top I_\theta y - \frac{1}{2}h^\top I_\theta h} \cdot \left(e^{-\frac{1}{2}y^\top I_\theta y} \sqrt{\frac{|I_\theta|}{(2\pi)^{k}}} \right) \, dy$$

$$= P_0 1_B(T(I_\theta Y, U))e^{h^\top I_\theta y - \frac{1}{2}h^\top I_\theta h} = E_1 B(S)e^{h^\top \Delta - \frac{1}{2}h^\top I_\theta h} = L_h(B),$$

as desired. \qed

**Corollary 7.10.** Suppose that $(P_0 : \theta \in \Theta)$ is LAN with non-singular Fisher information matrix $I_\theta$. Let $\psi$ be differentiable at $\theta$ and $\{T_n\}$ be estimators in $(P^n_{\theta + h/\sqrt{n}} : h \in \mathbb{R}^k)$ such that

$$\sqrt{n}(T_n - \psi(\theta + h/\sqrt{n})) \overset{p}{\rightarrow} P^n_{\theta + h/\sqrt{n}} L_{\theta,h}, \quad \forall h.$$ 

Then there exists a randomized statistics $T$ in $(N(h, I_\theta^{-1} : h \in \mathbb{R}^k))$ such that $T =_d L_{h,\theta} + \psi_\theta h$ under $P_h$.

**Proof.** Apply the previous theorem to

$$\sqrt{n}(T_n - \psi(\theta))$$

$$= \sqrt{n}(T_n - \psi(\theta + h/\sqrt{n})) + \sqrt{n}(\psi(\theta + h/\sqrt{n}) - \psi(\theta))$$

$$\overset{p}{\rightarrow} P^n_{\theta + h/\sqrt{n}} L_{\theta,h} + \psi_\theta h,$$

as desired. \qed

The corollary says that most estimators $T_n$ are matched with a randomized estimator $T$ in the Gaussian model in the sense that

$$\sqrt{n}(T_n - \psi(\theta + h/\sqrt{n})) \approx_{\theta + h/\sqrt{n}} T - \psi_\theta h.$$
7.4. Convolution theorem. Now we are at the position to formulate the CR lower bound in the limiting Gaussian model. As we have seen in the counter-example (e.g. Hodge’s estimator), any theory of formulating such a lower bound result must necessarily exclude locally irregular estimators. We make this precise in the following definition.

**Definition 7.11.** Let $Y \sim N(h, \Sigma)$. We call $T$ regular for estimating $Ah$ iff the distribution $T - Ah$ under $h$ does not depend on $h$.

A classical example for regular $T$ is given by $T = Ay$. This is the limit version of the sample mean in the univariate normal location model.

**Theorem 7.12.** Let $T$ be a regular estimator for estimating $Ah$, and $L$ be the distribution of $T$ under $0$. Then

$$L(L) = L(Z + W),$$

where $Z \sim N(0, A\Sigma A^\top)$ and $W$ is independent of $Z$.

The above theorem is the Gaussian version of the convolution(lower bound) theorem. Since adding an independent random variable does not decrease variance, the theorem says that any regular estimator in the Gaussian model for estimating $Ah$ must have variance bounded below by $A\Sigma A^\top$.

**Proof of Theorem 7.12.** The basic idea here is a decomposition of the target statistics $T - Ah$ as follows:

$$T - Ah = T - Ay + A(y - h) \equiv \tilde{W} + \tilde{Z}.$$  

The random variable $\tilde{Z}$ has the required distribution, but unfortunately $\tilde{W}, \tilde{Z}$ are not independent. The idea now is to randomize $h$, i.e. use a Bayesian argument. Let $H \sim N(0, \Lambda)$ where $\Lambda = \lambda I$ be the prior on the location parameter $h$, and $\lambda > 0$ is introduced to make the prior proper (in fact in the end we will let $\lambda \to \infty$ so the prior becomes uninformative). We may then define $Y$ conditionally on $H$ by $Y|H = h \sim N(h, \Sigma)$. Let $U$ be a uniform distribution on $[0, 1]$ independent of $(H,Y)$.

It is standard to calculate the posterior distribution of $H$ given $Y$:

$$H|Y \sim N((\Sigma^{-1} + \Lambda^{-1})\Sigma^{-1}Y, (\Sigma^{-1} + \Lambda^{-1})^{-1}).$$

Define

$$W_\Lambda \equiv T - A(\Sigma^{-1} + \Lambda^{-1})^{-1}\Sigma^{-1}Y,$$

$$G_\Lambda \equiv -A(H - (\Sigma^{-1} + \Lambda^{-1})^{-1}\Sigma^{-1}Y).$$

Then $T - Ah = W_\Lambda + G_\Lambda$. The crucial fact here is that $W_\Lambda$ and $G_\Lambda$ are independent due to randomization of $h$. To see this, note that $W_\Lambda$ on depends on $(Y,U)$. On the other hand, $G_\Lambda|Y \sim N(0, A(\Sigma^{-1} + \Lambda^{-1})^{-1}A^\top)$ is independent of $Y$. This proves the claim that $W_\Lambda$ and $G_\Lambda$ are independent. Note in the middle we also proved that $G_\Lambda$ converges in distribution to a normal random vector with covariance $A\Sigma A^\top$ as $\lambda \to \infty$. 
Now by regularity of $T$, the distribution of $T - AH$ is invariant for any choice of $\Lambda$. We denote this distribution as $L$. By the independence of $W_\Lambda$ and $G_\Lambda$, we have

$$\int e^{it \cdot y} \, dL(y) = E e^{it \cdot W_\Lambda} \cdot E e^{it \cdot G_\Lambda}.$$ 

Now let $\lambda \to \infty$, $E e^{it \cdot G_\lambda}$ converges to the characteristic function of a normal random vector $G$ with covariance $A \Sigma A^\top$, and therefore the characteristic function of $G_\Lambda$ converges pointwise to a continuous function. By Levy’s continuity theorem (recorded below), it follows that $W_\Lambda$ converges to some random variable with the characteristic function given by $E L e^{it \cdot Y}/E e^{it \cdot G}$.

The independence is preserved in taking the limits so we have proved the claim. □

**Lemma 7.13** (Lévy’s continuity theorem). Let $X_n$ be a sequence of random variables with characteristic functions $\phi_{X_n}$. If $\phi_{X_n}$ converges pointwise to some function $\phi$ that is continuous at 0, then $X_n$ converges in distribution to some random variable $X$ with characteristic function $\phi_X = \phi$.

*Proof.* Proof can be found in standard probability textbooks, e.g. Probability with Martingales, D. Williams, Section 18.1. □

Finally we are able to translate back the convolution theorem in the Gaussian model to the empirical model. Similarly we will need a notion of local regularity:

**Definition 7.14.** \{\(T_n\)\} is called regular for estimating $\dot{\psi}(\theta)$ at $\theta$ iff the limit distribution of $\sqrt{n}(T_n - \psi(\theta + h/\sqrt{n}))$ under $P_{\theta + h/\sqrt{n}}$ does not depend on $h$.

Note that our definition of regular estimator here rules out the Hodge’s estimator.

**Theorem 7.15.** Suppose that \((P_\theta : \theta \in \Theta)\) is LAN with non-singular Fisher information matrix $I_\theta$. Let $\psi$ be differentiable at $\theta$ and \{\(T_n\)\} be regular estimators in \((P_\theta : \theta \in \Theta)\) with limit distribution $L_\theta$. Then $L(L_\theta) = L(Z_\theta + W_\theta)$, where $Z_\theta \sim N(0, \dot{\psi}_\theta I_\theta^{-1} \dot{\psi}_\theta^\top)$ and $W_\theta$ is independent of $Z_\theta$.

*Proof.* By Corollary 7.10, we know that there exists a randomized estimator $T$ in the Gaussian model $(N(h, I_\theta^{-1}) : h \in \Theta)$ such that $T = L_\theta + \psi(h)$ under $P_h$. This means that $T$ is regular for estimating $\dot{\psi}_\theta h$. Now Theorem 7.12 applies to conclude. □

Summarizing the discussion so far, we have established that as long as the estimator is regular at $\theta$, then the CR lower bound continuous to apply in an asymptotic sense. The basic approach we take here is to use a Gaussian proxy model to approximate the empirical model, but the necessity of
regularity of the estimator that requires distributional uniformity in a local $n^{-1/2}$ neighborhood of the truth, makes Le Cam’s contiguity theory a critical technical tool for theoretical development.
8. Hypothesis Testing

8.1. **MP and UMP tests.** A hypothesis is a statement or assertion about a parameter (scalar) or vector. Typically we test a “null hypothesis” denoted as $H_0$ versus an “alternative hypothesis” $H_1$. The nature and complexity of these hypotheses depend on the type of question asked. Consider the following types of hypotheses for a scalar parameter $\theta$:

1. $H_0 : \theta = \theta_0$ versus $H_1 : \theta = \theta_1$ (point or simple null versus point or simple alternative).
2. $H_0 : \theta = \theta_0$ versus $H_1 : \theta > \theta_0$ (point null versus composite, one-sided alternative).
3. $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$ (point null versus composite, two-sided alternative).
4. $H_0 : \theta \leq \theta_0$ versus $H_1 : \theta > \theta_0$ (composite null versus composite alternative).
5. $H_0 : \theta = \theta_0$ versus $H_1 : \theta \in \mathbb{R}$ (point null nested within a composite alternative).

In the rest of this section, we will denote the parameter space specified by $H_0$ by $\Theta_0$ and that specified by $H_1$ by $\Theta_1$.

**Definition 8.1.** (1) A test is a function $\phi = \phi(Y) \in \{0, 1\}$ (i.e. we reject the null hypothesis if $\phi = 1$ and accept otherwise) and formally $R \equiv \phi^{-1}(1)$ is the rejection region which is typically determined by a test statistic $T(Y)$.

2. Type-I error (rejecting $H_0$ when actually true) = $\sup_{\theta \in \Theta_0} P_\theta \phi$;
   Type-II error (accepting $H_0$ when actually false) = $\sup_{\theta \in \Theta_1} P_\theta (1-\phi)$.

3. Test $\phi$ has level $\alpha$ iff $\sup_{\theta \in \Theta_0} P_\theta \phi \leq \alpha$.
   Test $\phi$ has power $\beta(\theta) \equiv P_\theta \phi$ at $\theta \in \Theta_1$.

Ideally, we want a test have both small type I and II errors. One approach for doing this is to find a test that has maximum power among all level $\alpha$ tests for all $\theta \in \Theta_1$. If such a test exists, it is called a uniformly most powerful (UMP) test of level-$\alpha$. A well-known tool used to construct MP and UMP tests is the Neyman-Pearson lemma described below.

**Theorem 8.2** (Neyman-Pearson lemma). Let $Y_1, \ldots, Y_n$ be a random sample from a distribution $P_\theta$ with pdf $f_\theta$ with respect to $\mu$. Consider the problem of testing the point null hypothesis $H_0 : \theta = \theta_0$ versus an alternative hypothesis $H_1 : \theta = \theta_1$. Then the test $\phi = 1_{R_c}$ with rejection region

\[ R_c = \left\{ Y : \frac{f_{\theta_1}(Y)}{f_{\theta_0}(Y)} \geq c \right\} \]

is the most powerful level $\alpha$ test if $c$ satisfies the size-$\alpha$ condition, i.e.,

\[ P_{\theta_0} \phi = P_{\theta_0}(Y \in R_c) = \alpha. \]

**Proof.** Consider the following diagrammatic representation of the sample space.
where \( R_1 \cup R_2 \) is the optimal Neyman-Pearson rejection region and \( R_2 \cup R_3 \) is the rejection region for any other level-\( \alpha \) test. Then \( R_1 \cup R_2 = \{ Y : \frac{f_{\theta_1}(Y)}{f_{\theta_0}(Y)} \geq c \} \). Since \( R_1 \cup R_2 \) satisfies the size-\( \alpha \) condition and \( R_2 \cup R_3 \) satisfies the level-\( \alpha \) condition, we must have
\[
P_{\theta_0}(R_1 \cup R_2) = \alpha \geq P_{\theta_0}(R_2 \cup R_3),
\]
which implies that
\[
\int_{R_1 \cup R_2} f_{\theta_0} \, d\mu \geq \int_{R_2 \cup R_3} f_{\theta_0} \, d\mu,
\]
which, in turn, implies
\[
(8.1) \quad \int_{R_1} f_{\theta_0} \, d\mu \geq \int_{R_3} f_{\theta_0} \, d\mu,
\]
We need to prove that \( P_{\theta_1}(Y \in R_1 \cup R_2) \geq P_{\theta_1}(Y \in R_2 \cup R_3) \), for which it suffices to show that \( P_{\theta_1}(Y \in R_1) \geq P_{\theta_1}(Y \in R_3) \). To that effect,
\[
P_{\theta_1}(Y \in R_1) = \int_{R_1} \left[ \frac{f_{\theta_1}}{f_{\theta_0}} \right] f_{\theta_0} \, d\mu
\geq c \int_{R_3} f_{\theta_0} \, d\mu, \quad \text{by the condition of the N-P rejection region}
\geq c \int_{R_3} f_{\theta_0} \, d\mu, \quad \text{by (8.1)}
\geq = c \int_{R_3} \left[ \frac{f_{\theta_0}}{f_{\theta_1}} \right] f_{\theta_1} \, d\mu \geq \int_{R_3} f_{\theta_1} \, d\mu = P_{\theta_1}(R_3),
\]
where the last inequality follows since \( \frac{f_{\theta_0}}{f_{\theta_1}} \geq 1/c \) on \( R_3 \), the region not contained in the optimal N-P rejection region. 

**Example 8.3.** Let \( Y_1, \ldots, Y_n \overset{iid}{\sim} N(\mu, 1) \). Consider testing \( H_0^{(1)} : \mu = \mu_0 \) versus \( H_1^{(1)} : \mu = \mu_1 (> \mu_0) \). Then,
\[
\frac{f_{\mu_1}(Y)}{f_{\mu_0}(Y)} = \exp \left\{ n \bar{Y}(\mu_1 - \mu_0) - (n/2)(\mu_1^2 - \mu_0^2) \right\},
\]
after some algebra. Thus \( \frac{f_{\mu_1}(Y)}{f_{\mu_0}(Y)} \geq c \) implies that \( \bar{Y} \geq c \), since \( \mu_1 > \mu_0 \). Thus the rejection region is \( R = \{ Y : \bar{Y} \geq c \} \), where \( c \) must satisfy \( P_{\mu_0} \{ \bar{Y} \geq c \} = \alpha \) and hence should be equal to \( \mu_0 + z_\alpha / \sqrt{n} \).
In Example 8.3, we saw that for the i.i.d. data from $N(\mu, 1)$, the MP test (rejection rule) for testing $H_0^{(1)}: \mu = \mu_0$ versus $H_1^{(1)}: \mu = \mu_1 (\mu > \mu_0)$ is

\begin{equation}
\phi_1 = 1_{R_1}, \quad R_1 = \{ \bar{Y} > \mu_0 + z_\alpha / \sqrt{n} \}.
\end{equation}

Note that the rejection region $R_1$ in (8.2) also gives a UMP test of level $\alpha$ for testing $H_0^{(1)}$ versus the composite alternative $H_1^{(2)}: \mu > \mu_0$, because it does not depend on $\mu_1$. Using a similar argument, a UMP rejection rule of level $\alpha$ for testing $H_0^{(1)}$ versus the composite alternative $H_1^{(3)}: \mu < \mu_0$ will be:

\begin{equation}
R_2 : \bar{Y} < \mu_0 - z_\alpha / \sqrt{n}.
\end{equation}

Now suppose, we wish to test the composite null $H_0^{(2)}: \mu \leq \mu_0$ versus the composite alternative $H_1^{(2)}: \mu > \mu_0$. We will argue that $\phi_1 \equiv \phi$ given by (8.2), which is UMP for testing $H_0^{(1)}$ versus $H_1^{(2)}$ is also UMP for testing $H_0^{(2)}$ versus $H_1^{(2)}$. To show this, we only need to establish the following:

(a) $\sup_{\mu \leq \mu_0} P_{P_{\mu}} \phi \leq \alpha$.

(b) For any test $\tilde{\phi}$ such that $\sup_{\mu \leq \mu_0} P_{P_{\mu}} \tilde{\phi} \leq \alpha$, it holds for all $\mu > \mu_0$ that $P_{P_{\mu}} \phi \geq P_{P_{\mu}} \tilde{\phi}$.

To establish (a), note that

\[
\begin{align*}
P_{P_{\mu}} \phi &= \int_{-\infty}^{\infty} \Phi\left( \Phi^{-1}(P_{P_{\mu}} \phi) - z_\alpha \sqrt{n} \right) \cdot \frac{1}{\sqrt{2\pi}} \exp\left( -\frac{1}{2} \left( \frac{\bar{Y} - \mu}{\sqrt{n}} \right)^2 \right) \, d\mu \\
&= P_{P_{\mu}} \{ Z \leq \sqrt{n}(\mu - \mu_0)/\sqrt{n} \}, \text{ where } Z \sim N(0,1) \\
&= P_{P_{\mu}} \{ Z \leq \sqrt{n}(\mu - \mu_0)/\sqrt{n} \}, \text{ where } Z \sim N(0,1),
\end{align*}
\]

where $\Phi(\cdot)$ is the CDF of a standard normal variable. Thus $P_{P_{\mu}} \phi$ is a non-decreasing function of $\mu$. Thus, for all $\mu \leq \mu_0$, $P_{P_{\mu}} \phi \leq P_{P_{\mu_0}} \phi = \alpha$ and (a) is satisfied.

For (b), we have already seen that $\phi$ is UMP for testing $H_0^{(1)}$ against $H_1^{(2)}$, i.e. for any test $\tilde{\phi}$ such that $\sup_{\mu \leq \mu_0} P_{P_{\mu}} \tilde{\phi} \leq \alpha$, it holds for all $\mu > \mu_0$ that $P_{P_{\mu}} \phi \geq P_{P_{\mu}} \tilde{\phi}$. Any test $\tilde{\phi}$ in (b) certainly satisfies the condition of the previous sentence.

This example for testing hypotheses on the mean of the normal distribution can be generalized by (a) introducing a monotonicity condition on the likelihood ratio, and (b) applying the Karlin-Rubin Theorem (Theorem 8.5) stated later.

**Definition 8.4 (Monotone Likelihood Ratio).** The family of distributions $\{f_{\theta} : \theta \in \Theta\}$, has **monotone likelihood ratio** (MLR) in a statistic $T$ if the ratio $f_{\theta_2}/f_{\theta_1}$ is non-decreasing in $T$ for each pair of fixed $\theta_1 < \theta_2$.

**Theorem 8.5 (Karlin-Rubin).** Consider testing $H_0 : \theta \leq \theta_0$ versus $H_1 : \theta > \theta_0$. Let $T$ be a sufficient statistic for $\theta$, and suppose $\{f_{\theta} : \theta \in \Theta\}$ has MLR in $T$. Then, for any $t_0$, the test $\phi = 1_{T > t_0}$ is UMP level-$\alpha$, where $\alpha = P_{\theta_0}(T > t_0)$.
The proof of the Karlin-Rubin Theorem is similar to our argument for the normal example, the details of which can be found in, e.g. Chapter 8 of Casella and Berger (2002). The key element of the proof involves showing that the MLR condition leads to a non-decreasing power function.

**Example 8.6.** Consider the exponential family with pdf

\[ f_\theta(Y) = \exp\{T(Y) \eta(\theta) - \psi(\eta)\} h(Y). \]

For \( \theta_1 < \theta_2 \), the log of the likelihood ratio is

\[ \lambda = T(Y) \{ \eta(\theta_2) - \eta(\theta_1) \} - \{ \psi(\eta)|_{\theta_2} - \psi(\eta)|_{\theta_1} \}, \]

which implies \( \partial \lambda / \partial T = \eta(\theta_2) - \eta(\theta_1) \geq 0 \) if \( \eta(\theta) \) is a non-decreasing function of \( \theta \).

**Example 8.7.** Let \( Y_1, \ldots, Y_n \overset{iid}{\sim} \text{Bernoulli}(p) \). Consider testing \( H_0 : p \leq p_0 \) versus \( H_1 : p > p_0 \). The likelihood ratio

\[ \frac{f_{p_1}(Y)}{f_{p_0}(Y)} = \left( \frac{p_1}{p_0} \right)^T \left( \frac{1 - p_0}{1 - p_1} \right)^T, \]

where \( T = \sum_{i=1}^n Y_i \), is increasing in \( T \) for any \( p_1 > p_0 \), and hence the family has an MLR in \( T \). Thus, by Karlin-Rubin theorem, the UMP test is \( \phi = 1_{T > c} \) with level

\[ \alpha = P_{p_0}(T > c) = P_{p_0}(T \geq c + 1) = \sum_{j=c+1}^{n} \binom{n}{j} p_0^j (1-p_0)^j. \]

**8.2. Likelihood ratio test.** Consider testing \( H_0 : \theta \in \Theta_0 \) versus \( H_1 : \theta \in \Theta_1 \), where \( \Theta_0 \cup \Theta_1 = \Theta \) based on i.i.d. observations \( (Y_1, \ldots, Y_n) \) from \( f_\theta \). If \( \Theta_0 \) and \( \Theta_1 \) are singleton, then Neyman-Pearson tells us the ‘optimal’ test is given by

\[ \log \frac{f_{\theta_1}(Y)}{f_{\theta_0}(Y)} = \log \prod_{i=1}^n \frac{f_{\theta_1}(Y_i)}{f_{\theta_0}(Y_i)}. \]

Now we consider a general version:

\[ \tilde{\Lambda}_n \equiv \log \sup_{\theta \in \Theta_1} \prod_{i=1}^{n} \frac{f_{\theta}(Y_i)}{f_{\theta_0}(Y_i)}. \]

The test statistic \( \tilde{\Lambda} \) has a somewhat complicated limit distribution so for technical reason we often refer to the following as the likelihood ratio statistic:

\[ \Lambda_n = 2 \log \frac{\sup_{\theta \in \Theta} \prod_{i=1}^{n} f_{\theta}(Y_i)}{\sup_{\theta \in \Theta_0} \prod_{i=1}^{n} f_{\theta}(Y_i)} = 2(\tilde{\Lambda}_n \lor 0). \]

Before we formally derive the limit distribution for \( \Lambda_n \), let us make some heuristic arguments to see what should be expected. To this end, let \( H_n = \)
\( \sqrt{n}(\Theta - \theta) \) and \( H_{n,0} = \sqrt{n}(\Theta_0 - \theta) \) be the ‘local sets’. Then if \( H_n, H_{n,0} \) converge suitably to some set \( H \) and \( H_0 \), using the Gaussian proxy idea,

\[
\Lambda_n = 2 \sup_{h \in H_n} \log \frac{\prod_{i=1}^n f_{\theta + h/\sqrt{n}}(Y_i)}{\prod_{i=1}^n f_{\theta}(Y_i)} - 2 \sup_{h \in H_{n,0}} \log \frac{\prod_{i=1}^n f_{\theta + h/\sqrt{n}}(Y_i)}{\prod_{i=1}^n f_{\theta}(Y_i)} \\
\approx 2 \sup_{h \in H} \frac{dN(h, I_{\theta}^{-1})}{dN(0, I_{\theta}^{-1})}(Y) - 2 \sup_{h \in H_0} \frac{dN(h, I_{\theta}^{-1})}{dN(0, I_{\theta}^{-1})}(Y) \\
= \inf_{h \in H_0} (Y - h)^\top I_{\theta}(Y - h) - \inf_{h \in H} (Y - h)^\top I_{\theta}(Y - h) \\
= \| I_{\theta}^{1/2}(Y - H_0) \|^2 - \| I_{\theta}^{1/2}(Y - H) \|^2.
\]

Here \( Y \) is a single observation from the limit Gaussian location model \( (N(h, I_{\theta}^{-1}) : h \in \mathbb{R}^k) \). So under \( h = 0 \), and the assumption that \( H = \mathbb{R}^k \) and \( H_0 \) is an \( \ell \)-dimensional subspace of \( \mathbb{R}^k \), the distribution on the right hand side of the above play is \( \chi^2_{k-\ell} \).

**Definition 8.8.** We say that sets \( H_n \) converge to \( H \), notationally \( H_n \to H \), iff for any \( h \in H \), there exist \( \{ h_n \in H_n \} \) such that \( h_n \to h \) in \( \mathbb{R}^k \).

**Theorem 8.9** (Wilk’s phenomenon). Suppose that \( (P_\theta : \theta \in \Theta) \) is uniformly LAN. Further assume that the constrained and unconstrained MLEs \( \hat{\theta}_{n,0} \) and \( \hat{\theta}_n \) (i.e. MLEs over \( \Theta_0 \) and \( \Theta \) respectively) are \( \sqrt{n} \)-consistent for estimating \( \theta \). If \( H_{n,0} \to H_0 \) and \( H_n \to H \), then for any \( h \),

\[
\Lambda_n \leadsto P_{\hat{\theta} + h/\sqrt{n}} \Lambda
\]

where

\[
\Lambda = \| I_{\hat{\theta}}^{1/2}Y - I_{\hat{\theta}}^{-1/2}H_0 \|^2 - \| I_{\hat{\theta}}^{1/2}Y - I_{\hat{\theta}}^{-1/2}H \|^2
\]

for a standard Gaussian vector \( Y \equiv d N(h, I_{\theta}^{-1}) \).

We mention that the classical conditions verifying the assumptions of Theorem 8.9 are

- \( (P_\theta : \theta \in \Theta) \) is differentiable in quadratic mean, and
- there exists \( \ell \) with \( P_\theta \ell^2 < \infty \) such that for \( \theta_1, \theta_2 \) close enough to \( \theta \),

\[
| \log f_{\theta_1}(y) - \log f_{\theta_2}(y) | \leq \ell(y) \| \theta_1 - \theta_2 \|
\]

These conditions are rather mild, and in particular do not assume strong differentiability on the log likelihood (give a proof of the above theorem when assuming sufficiently many regularity conditions!).

Now we turn to the proof of Theorem 8.9.

**Lemma 8.10.** Suppose the sets \( H_n \to H \) and the random vectors \( X_n \to d X \). Then \( \| X_n - H_n \| \to \| X - H \| \).

**Proof.** Since the map \( x \mapsto \| x - H \| \) is continuous, by continuous mapping theorem, \( \| X_n - H \| \to d \| X - H \| \). So we only need to prove that \( \| X_n - H_n \| - \| X_n - H \| \to p 0 \).
By the uniform LAN, for every slowly growing other direction follows from similar arguments so we omit the details. □

This means the above display is also valid along the original sequence. The other direction follows from similar arguments so we omit the details.

**Proof of Theorem 8.9.** By the uniform LAN, for every slowly growing $M_n \to \infty$,

$$
\sup_{\|h\| \leq M_n} \left| n \mathbb{P}_n \log \frac{f_{\theta + h/\sqrt{n}}}{f_{\theta}} - h^\top \mathbb{E}_n \hat{\epsilon}_{\theta} + \frac{1}{2} h^\top I_0 h \right| \to_{P_{\theta}} 0.
$$

By contiguity $P_{\theta}^n \triangleq P_{\theta + h/\sqrt{n}}^n$ as will be shown below, the above convergence can be replaced under $P_{\theta + h/\sqrt{n}}^n$. By $\sqrt{n}$-consistent of $\hat{\theta}_{n,0}$ and $\hat{\theta}_n$, we may replace the supremum over $H_n, H_{n,0}$ in the definition of the likelihood ratio statistic by that of $H_n \cap B(0, M_n)$ and $H_{n,0} \cap B(0, M_n)$ respectively. Rigorously,

$$
\tilde{\Lambda}_n \equiv 2 \sup_{h \in H_n \cap B(0, M_n)} n \mathbb{P}_n \log \frac{f_{\theta + h/\sqrt{n}}}{f_{\theta}} - 2 \sup_{h \in H_{n,0} \cap B(0, M_n)} n \mathbb{P}_n \log \frac{f_{\theta + h/\sqrt{n}}}{f_{\theta}} = 2 \mathbb{E}_n \mathbb{E}_n \hat{\epsilon}_{\theta} - \frac{1}{2} h^\top I_0 h
$$

$$
= 2 \sup_{h \in H_{n,0} \cap B(0, M_n)} \left( h^\top \mathbb{E}_n \hat{\epsilon}_{\theta} - \frac{1}{2} h^\top I_0 h \right) + o_{P_{\theta + h/\sqrt{n}}} \left( \mathbb{E}_n \hat{\epsilon}_{\theta} - \frac{1}{2} h^\top I_0 h \right).
$$

Now we wish to find out the limit distribution of $\mathbb{E}_n \hat{\epsilon}_{\theta}$ under $P_{\theta + h/\sqrt{n}}^n$ via Le Cam’s third lemma. First note that

$$
(\mathbb{E}_n \hat{\epsilon}_{\theta}, \mathbb{E}_n \hat{\epsilon}_{\theta}) \sim_{P_{\theta}} (\Delta, \Delta)
$$

where $\Delta \sim N(0, I_0)$. By LAN and Slutsky’s theorem, it follows that

$$
(\mathbb{E}_n \hat{\epsilon}_{\theta}, \log \frac{dP_{\theta}^n}{dP_{\theta}}) \sim_{P_{\theta}} (\Delta, h^\top \Delta - \frac{1}{2} h^\top I_0 h)
$$

$$
=d N \left( \left( \begin{array}{c} 0 \\ -\frac{1}{2} h^\top I_0 h \end{array} \right), \left( \begin{array}{cc} I_0 & I_0 \\ h^\top I_0 & h^\top I_0 \end{array} \right) \right).
$$
This shows in particular that $P_n^{\theta} \overset{\triangle}{=} P_{\theta+h/\sqrt{n}}$. Now using Le Cam’s third lemma, we conclude that

$$G_n \overset{\triangle}{=} P_{\theta+h/\sqrt{n}} Y.$$

By Lemma 8.10 and continuous mapping theorem, we conclude that

$$\bar{\Lambda}_n \overset{\triangle}{=} P_{\theta+h/\sqrt{n}} \Lambda,$$

and the claim of the theorem follows from the fact that $P_{\theta+h/\sqrt{n}}(\bar{\Lambda}_n \neq \Lambda_n) \to 0$.

\[\text{Example 8.11.}\] Consider the location-scale problem: let $Y_1, \ldots, Y_n$ be i.i.d. samples from $f((\cdot - \mu)/\sigma)$ where $f$ is a known density function and $\theta = (\mu, \sigma)$ is the unknown parameter.

1. Suppose we want to test $H_0 : \mu = 0$ vs $H_1 : \mu \neq 0$. Then $\Theta_0 = \{0\} \times \mathbb{R}_-$, and hence for any $\theta = (0, \sigma)$, $H_{n,0} = \sqrt{n}(\Theta_0 - \theta) = \{0\} \times (-\infty, \infty) \to \{0\} \times \mathbb{R}$. This shows that the LR statistic converges to the square distance between a standard 2-dimensional Gaussian vector and $\{0\} \times \mathbb{R}$, which equals in distribution to $\chi^2_1$ (under regularity conditions).

2. Suppose we want to test $H_0 : \mu \leq 0$ vs $H_1 : \mu > 0$. Then $\Theta_0 = (-\infty, 0) \times \mathbb{R}_+$, and hence for any $\theta = (0, \sigma)$, $H_{n,0} = \sqrt{n}(\Theta - \theta) \to (-\infty, 0] \times \mathbb{R} = H_0$. This shows that

$$\Lambda_n \to_d \|Z - I_{\theta}^{1/2}H_0\|^2 = (Z_1 \lor 0)^2.$$

Here $Z = (Z_1, Z_2) \sim N(0, I)$. Note that the right hand side of the above display is not chi-squared distributed. On the other hand, if $\theta \in \text{int}(\Theta_0)$, then $H_{n,0} \to \mathbb{R} \times \mathbb{R}$ and hence $\Lambda_n \to_p 0$.

\[\text{Example 8.12.}\] Consider testing $H_0 : \|\theta\| \leq 1$ vs $H_1 : \|\theta\| > 1$. If the truth $\theta \in \text{int}B(0, 1)$ (here $B(0, 1)$ denotes the closed ball centered at 0 with radius 1), then $\sqrt{n}(\Theta_0 - \theta) \to \mathbb{R}^k$ and therefore the likelihood ratio statistics $\Lambda_n \to_p 0$. If the truth $\theta \in \partial B(0, 1)$, then $\sqrt{n}(\Theta_0 - \theta) \to \{h : h^\top \theta \leq 0\}$ converges to half space. By similar arguments as in the previous example, we see that $\Lambda_n \to_d (Z_1 \lor 0)^2$ where $Z_1 \sim N(0, 1)$.

8.3. \textbf{Efficiency of tests.} Consider a general testing problem where we would like to test $H_0 : \psi(\theta) \leq 0$ versus $H_1 : \psi(\theta) > 0$. Here $\theta \in \mathbb{R}^k$ and the function $\psi : \mathbb{R}^k \to \mathbb{R}$ vanishes under the truth $\theta_0$, i.e. $\psi(\theta_0) = 0$.

The question that we would like to address in this section is the following: Given a sequence of (asymptotically) level-$\alpha$ tests $\phi_n$, what is the best possible power of the test against alternatives?

Apparently, the problem is mostly degenerate when considering fixed alternative $\theta_1$ since the best possible power converges to 1. So below we will focus on \textit{local alternatives} in the similar sense as in the development of efficiency theory in Section 7, that is, we will consider alternatives in a local ball with radius on the order of $n^{-1/2}$. Now we can formulate the problem...
in a more transparent mathematical way: suppose that the tests $\phi_n$ are at level-$\alpha$ asymptotically, i.e. $\limsup_n P_{\theta_0} \phi_n \leq \alpha$, then what is the best possible power of the tests $\phi_n$ under local alternative $\theta_0 + h/\sqrt{n}$, i.e. what is the lowest possible upper bound for $\limsup_n P_{\theta_0+h/\sqrt{n}} \phi_n$?

To answer this question, we will adopt a similar approach that we have seen in Section 7: we would like to consider the problem purely in the limiting Gaussian location model, the benefit of which is that we know very well by Neyman-Pearson lemma that the likelihood ratio test is UMP in such a model. To implement this idea, we need

- construct a test in the limiting Gaussian location model that matches with $\phi_n$ in some quantifiable sense, and
- calculate the lowest possible upper bound for the power of UMP test (= likelihood ratio test) in the Gaussian location model under constant shift.

We first consider the first step. For notational convenience, let us define the power function of $\phi_n$ under local alternatives by $\beta_n$, i.e.

$$\beta_n(h) \equiv P^n_{\theta+h/\sqrt{n}} \phi_n.$$  

(8.4)

**Proposition 8.13.** Let $(P_{\theta} : \theta \in \Theta)$ be LAN. Let $\phi_n$ be tests in $(P_{\theta+h/\sqrt{n}} : h \in \mathbb{R}^k)$ with power functions $\beta_n$ such that $\beta_n$ converges pointwise to some function $\beta$. Then there exists some randomized test $\phi$ in the Gaussian location model $(N(h,I_{\theta}^{-1}) : h \in \mathbb{R}^k)$ such that $\beta$ is the power function for $\phi$.

**Proof.** We will apply Theorem 7.9. To this end, we only need to verify that $\phi_n$ converges under $P^n_{\theta+h/\sqrt{n}}$ for any $h$. To see this, take any subsequence \{n\}, since $(\phi_n, G_n, \ell_n)$ is uniformly tight under $P^n_{\theta}$, Prohorov’s theorem asserts that $(\phi_n, G_n, \ell_n)$ converges in distribution along some subsequence \{nk\}. Now by LAN and Slutsky’s theorem, $(\phi_n, \log dP^n_{\theta+h/\sqrt{n}}/dP^n_{\theta})$ also converges in distribution under $P^n_{\theta}$ along the subsequence. Apply Le Cam’s third lemma we see that $\phi_n$ converges for any $P^n_{\theta+h/\sqrt{n}}$ along the subsequence \{nk\}, and hence the original sequence. Now Theorem 7.9 applies to see that there exists a randomized statistics $\phi \in \{0,1\}$ such that $\phi_n \overset{d}{\to} P^n_{\beta+h/\sqrt{n}} \phi$ for every $h$. Since $\phi_n$ is uniformly bounded (by 1), converges in distribution implies converges in mean (this can be proved via e.g. Skorokhod’s representation plus dominated convergence theorem), i.e.

$$\beta_n(h) = P^n_{\theta+h/\sqrt{n}} \phi_n \to P_h \phi,$$

where $P_h$ is the measure corresponding to the law $N(h, I_{\theta}^{-1})$. By our assumption that $\beta_n$ converges pointwise, it follows that $\beta(h) = P_h \phi$. \qed

Next we consider the second step.
**Proposition 8.14.** Let $Y \sim N(h, \Sigma)$ and $c \in \mathbb{R}^k$ be such that $c^\top \Sigma c > 0$. Consider the testing problem $H_0 : c^\top h \leq 0$ versus $c^\top h > 0$. Then the test $\phi = 1_{c^\top Y > z_0 \sqrt{c^\top \Sigma c}}$ is UMP at level $\alpha$ (over all potentially randomized tests).

**Proof.** Since $c^\top Y \sim N(c^\top h, c^\top \Sigma c)$, it follows by Neyman-Pearson lemma (note that here we use a strengthened version that also works for randomized tests, check this by conditioning arguments through the original proof!) that $\phi$ is UMP for two-point testing problem $H_0^{(1)} : c^\top h = 0$ versus $H_1^{(1)} : c^\top h = \mu_1 > 0$. Since $\phi$ does not depend on $\mu_1$, it is UMP for testing $H_0^{(2)} : c^\top h = 0$ versus $H_1^{(2)} : c^\top h > 0$. The claim follows by a monotone likelihood ratio arguments as we have seen before Karlin-Rubin Theorem (cf. Theorem 8.5). 

Now completing the two steps outlined in the beginning of this section, we are able to state the main efficiency result.

**Theorem 8.15.** Let $(P_\theta : \theta \in \Theta)$ be LAN with an open set $\Theta$. Let $\psi$ be differentiable at $\theta_0$ with gradient $\psi_{\theta_0}$ and $\psi(\theta_0) = 0$. Let $\phi_n$ be level-$\alpha$ tests for $H_0 : \psi(\theta) \leq 0$ versus $H_1 : \psi(\theta) > 0$. Then for any $h$ such that $\dot{\psi}_{\theta_0} h > 0$,

$$\limsup_{n \to \infty} \beta_n(h) = \limsup_{n \to \infty} P_{\theta_0 + h/\sqrt{n}} \phi_n \leq 1 - \Phi\left(z_\alpha - \frac{\dot{\psi}_{\theta_0} h}{\sqrt{\psi_{\theta_0}^\top \Sigma^{-1} \psi_{\theta_0}}} \right).$$

Here $\Phi$ is the normal cdf.

**Proof.** Fix $h$ such that $\dot{\psi}_{\theta_0} h > 0$ and a subsequence $\{n\}$ for which $\limsup_n \beta_n(h)$ is taken (so the sequence depends on the choice of $h$). By the proof of Proposition 8.13 (tightness + Le Cam third lemma), there exists a further subsequence along which $\beta_n$ converges pointwise. For notational convenience, we still use $\{n\}$ for this subsequence. Now by Proposition 8.13, there exists a randomized test $\phi$ in the Gaussian local model $(N(h, I^{-1}_\theta) : h \in \mathbb{R}^k)$ such that the power function of $\phi$ is given by $\beta = \lim \beta_n$.

We claim that $\phi$ is a level-$\alpha$ test for testing $H_0 : \dot{\psi}_{\theta_0} h \leq 0$ versus $H_1 : \dot{\psi}_{\theta_0} h > 0$. To this end, we only need to verify for any $h'$ such that $\dot{\psi}_{\theta_0} h' \leq 0$, $\beta(h') \leq \alpha$. A further reduction is that we only need to consider $h'$ for which $\dot{\psi}_{\theta_0} h' < 0$ whereas the case for equality follows from continuity of $\beta(\cdot)$. Now by assumption,

$$\psi(\theta_0 + h'/\sqrt{n}) = \psi(\theta_0) + (\dot{\psi}_{\theta_0} h')/\sqrt{n} + o(n^{-1/2}) < 0$$

for $n$ large enough. Since $\phi_n$ are level-$\alpha$ tests for $H_0 : \psi(\theta) \leq 0$, we have for $n$ large enough $P_{\theta_0 + h'/\sqrt{n}} \phi_n \leq \alpha$. Therefore it follows that

$$\beta(h') = \lim_n \beta_n(h') \leq \limsup_n P_{\theta_0 + h'/\sqrt{n}} \phi_n \leq \alpha,$$

proving the claim.
By Proposition 8.14, \( \tilde{\phi} = 1_{\psi_0^\top Y > z_\alpha \sqrt{\psi_0^\top I_0^{-1} \psi_0}} \) is a UMP test. This means that

\[
\limsup_n \beta_n(h) = \beta(h) \leq P_h \tilde{\phi}
\]

\[
= P_h \left( \psi_0^\top Y > z_\alpha \sqrt{\psi_0^\top I_0^{-1} \psi_0} \right)
\]

\[
= P \left( N \left( \frac{\psi_0^\top h}{\sqrt{\psi_0^\top I_0^{-1} \psi_0}}, 1 \right) > z_\alpha \right)
\]

\[
= P \left( N(0, 1) > z_\alpha - \frac{\psi_0^\top h}{\sqrt{\psi_0^\top I_0^{-1} \psi_0}} \right)
\]

\[
= 1 - \Phi \left( z_\alpha - \frac{\psi_0^\top h}{\sqrt{\psi_0^\top I_0^{-1} \psi_0}} \right),
\]

as desired. \( \square \)

In summary, we have shown that under mild regularity conditions, any sequence of level-\( \alpha \) tests \( \phi_n \) for the problem \( H_0 : \psi(\theta) \leq 0 \) versus \( H_1 : \psi(\theta) > 0 \) must have asymptotic power bounded by

\[
1 - \Phi \left( z_\alpha - \frac{\psi_0^\top h}{\sqrt{\psi_0^\top I_0^{-1} \psi_0}} \right)
\]

over \( h/\sqrt{n} \) local alternatives. The theoretical development resembles that of the efficiency lower bound theory, where the key idea is to analyze the problem in the limiting Gaussian location model.

8.3.1. **Optimality of likelihood ratio test for simple null hypothesis in one dimension.** In this section, we will use Theorem 8.15 as a benchmark to get some further insight about the power of likelihood ratio test.

Suppose that the truth \( \theta \in \text{int} \Theta \), then Wilk’s phenomenon entails that the likelihood ratio statistic

\[
\Lambda_n \sim_{F_{\theta + h/\sqrt{n}}} \Lambda = \|Z + I_0^{1/2} h - I_0^{1/2} H_0\|^2
\]

where \( H_0 = \lim_n \sqrt{n}(\Theta_0 - \theta) \) and \( Z \sim N(0, I) \). Further assume that \( H_0 \) is a linear subspace with dimension \( \ell \), then \( \Lambda \sim \chi^2_{k-\ell} \), and the null hypothesis is rejected at \( \chi^2_{k-\ell, \alpha} \). This means that

\[
\beta_n(h) = P_{\theta + h/\sqrt{n}}(\Lambda_n > \chi^2_{k-\ell, \alpha})
\]

\[
\rightarrow P \left( \|N(I_0^{1/2} h, I) - I_0^{1/2} H_0\|^2 > \chi^2_{k-\ell, \alpha} \right)
\]

\[
= P \left( \chi^2_{k-\ell} \left( \|I_0^{1/2} h - I_0^{1/2} H_0\| \right) > \chi^2_{k-\ell, \alpha} \right)
\]
where \( \chi^2_m(\delta) \) is the non-central chi-squared distribution with \( m \) degrees of freedom.

- Consider the one-dimension test \( H_0 : \theta = 0 \) versus \( H_1 : \theta > 0 \). Then \( H_0 = \{ 0 \} \) and hence
  \[
  \beta_n(h) \rightarrow P\left( |N(\sqrt{I_\theta}h, 1)|^2 > \chi^2_{k-\ell, \alpha} \right)
  = P\left( N(\sqrt{I_\theta}h, 1) > z_\alpha \right) = 1 - \Phi(z_\alpha - h\sqrt{I_\theta}).
  \]
  This shows the likelihood ratio test achieves the efficiency bound in the one-dimensional simple point testing problem.

- Let us still consider the simple point testing problem as above but now in a multi-dimensional setting. Then
  \[
  \beta_n(h) \rightarrow \beta(h) = P\left( \chi^2_k(\sqrt{h^\top I_\theta}h) > \chi^2_{k, \alpha} \right).
  \]
  The non-central chi-squared distribution is stochastically increasing in \( \delta \rightarrow \chi^2_k(\delta) \) so \( \sup_{\|h\| \leq 1} \beta(h) \) is attained at the vector \( h \) for which \( h \rightarrow h^\top I_\theta h \) attains its maximum. This is certainly determined by the eigen-structure of the Fisher information matrix \( I_\theta \), so the likelihood ratio statistic has better power in certain direction of local alternatives than others.

8.3.2. Tests achieving the efficiency bound in any dimension. We have seen the likelihood test is efficient in the one-dimensional simply point testing problem, but its power in a higher-dimensional setting depends on the direction of the local alternative and the eigen-structure of the Fisher information. The natural question then is, is there a universally efficient test in the sense of Theorem 8.15 in any dimension?

**Theorem 8.16.** Suppose that \( (P_\theta : \theta \in \Theta) \) is LAN, and that \( \psi \) is differentiable at \( \theta \) with gradient \( \dot{\psi}_\theta \) and \( \psi(\theta_0) = 0 \). Let
  \[
  T_n = \frac{\dot{\psi}_\theta^\top I_{\theta_0}^{-1} G_n \dot{\theta}}{\sqrt{\dot{\psi}_\theta^\top I_{\theta_0}^{-1} \dot{\psi}_\theta}}
  \]
  be the score statistics and \( \phi_n = 1_{T_n > z_\alpha} \). Then for any \( h \),
  \[
  P^{n}_{\theta+h/\sqrt{n}} \phi_n \rightarrow 1 - \Phi \left( z_\alpha - \frac{\dot{\psi}_{\theta_0}^\top h}{\sqrt{\dot{\psi}_{\theta_0}^\top I_{\theta_0}^{-1} \dot{\psi}_{\theta_0}}} \right).
  \]

**Proof.** We only need to find out the limit distribution of \( G_n \dot{\theta} \) under \( P^{n}_{\theta+h/\sqrt{n}} \) and then use continuous mapping theorem. This has already be done in the proof of Theorem 8.9 so we omit the details. \( \square \)

Another way to obtain an efficient test is to use an efficient estimator.
Example 8.17. For the model \((P_\theta : \theta \in \Theta)\), under regularity conditions, we know that the MLE \(\hat{\theta}_n\) is efficient at \(\theta\):
\[
\sqrt{n}(\hat{\theta}_n - \theta) = I_\theta^{-1}G_n\hat{\theta}_n + o_P(1).
\]
Under LAN of the model \((P_\theta : \theta \in \Theta)\) and Le Cam’s third lemma, we can easily deduce the law of \(\sqrt{n}(\hat{\theta}_n - \theta_0)\) under \(P^n_{\theta + h/\sqrt{n}}\). If \(\theta \mapsto I_\theta\) is continuous and \(\psi\) is \(C^1\), then the test rejects \(H_0 : \psi(\theta) \leq 0\) for large values of
\[
\frac{\sqrt{n}\psi(\hat{\theta}_n)}{\sqrt{\psi_{\hat{\theta}_n}^{-1}T_{\hat{\theta}_n}^{-1}\psi_{\hat{\theta}_n}}}
\]
is asymptotically efficient for any truth \(\theta\) at the boundary of the null \(H_0\).

8.4. Pitman’s relative efficiency. Let us consider a one-dimensional simple point testing problem: \(H_0 : \theta = 0\) versus \(H_1 : \theta = \theta_t > 0\). Here \(t\) is regarded as time and we will consider asymptotics as \(t \to \infty\), whereas the alternative \(\theta_t \not\to 0\).

The general question we would like to address in this section is the following: Given two sequences of tests \(\{\phi_{n,1}\}\) and \(\{\phi_{n,2}\}\) with the same asymptotic level \(\alpha\) and power \(\beta\), how do we compare the two tests?

Let
\[
n_{t,i} \equiv \min\{n : P_\theta\phi_{n,i} \leq \alpha, P_{\theta_t}\phi_{n,i} \leq \beta\}
\]
be the number of observations that is needed to guarantee both level \(\alpha\) and power \(\beta\) for the test \(\{\phi_{n,i}\}\) at time \(t\).

Definition 8.18. The Pitman efficiency between \(\{\phi_{n,1}\}\) and \(\{\phi_{n,2}\}\) is defined by
\[
p(\phi_{n,1}, \phi_{n,2}) \equiv \lim_{t \to \infty} \frac{n_{t,2}}{n_{t,1}}
\]
if the limit on the right hand side of the above display is well-defined.

If \(p(\phi_{n,1}, \phi_{n,2}) > 1\), then using the tests \(\{\phi_{n,1}\}\) requires smaller samples to achieve level \(\alpha\) and power \(\beta\) so in this sense it is better.

To facilitate theoretical analysis, we consider the following special form of \(\phi_{n,i}\) that is based on the statistics \(T_{n,i}\) satisfying
\[
\frac{\sqrt{n}(T_{n,i} - \mu_i(\theta_0))}{\sigma_i(\theta_0)} \Rightarrow_{\theta_0} N(0,1)
\]
for any \(\theta_0 \not\to 0\). So \(\phi_{n,i} = 1_{T_{n,i} > \mu_i(0) + \sigma_i(0)z_\alpha/\sqrt{n}}\).

Theorem 8.19. Let \(\beta > \alpha\). Let \((P_\theta : \theta \in \Theta)\) be such that \(d_{TV}(P^n_\theta, P^n_0) \to 0\) as \(\theta \to 0\) for every \(n\). Let \(\phi_{n,i}(i = 1, 2)\) be tests based on statistics \(\{T_{n,i}\}\) as described above. Further assume that \(\mu_i\) is differentiable at \(0\) with \(\mu_i'(0) > 0\) and \(\sigma_i\) is continuous at \(0\) with \(\sigma_i(0) > 0\) for \(i = 1, 2\). Then if \(\theta_t \to 0\),
\[
p(\phi_{n,1}, \phi_{n,2}) = \left(\frac{\mu_1'(0)/\sigma_1(0)}{\mu_2'(0)/\sigma_2(0)}\right)^2.
\]
Proof. We first show that \( n_{t,i} \to \infty \) as \( t \to \infty \) for \( i = 1, 2 \). To see this, with \( R_{n,i} = \phi_{n,i}^{-1}(1) \), if \( \{n_{t,i}\} \) stays bounded, then

\[
1 > \alpha + (1 - \beta) \geq P_{n_{t,0}}(\phi_{n,i} + P_{n_{t},\theta_i}(1 - \phi_{n,i})
\geq 1 + \int_{R_{n,i}} (p_{n_{t,0}} - p_{n_{t},\theta_i}) \, d\mu
\geq 1 - \frac{1}{2}d_{TV}(P_{n_{t,0}}, P_{n_{t},\theta_t}) \to 1,
\]

a contradiction, so the claim is proved. This means that \( \{\phi_{n_{t,i}}\} \) have asymptotic size \( \alpha \) and power \( \beta \) respectively. So \( \phi_{n_{t,i}} = 1 \) if

\[
\sqrt{n_{t,i}}(T_{n_{t,i}} - \mu_i(0)) > \sigma_i(0)z_\alpha,
\]

which implies that

\[
P_{\theta,\phi_{n_{t,i}}} = P_{\theta_t}\left(\sqrt{n_{t,i}}(T_{n_{t,i}} - \mu_i(\theta_t)) + \sqrt{n_{t,i}}(\mu_i(\theta_t) - \mu_i(0)) > \sigma_i(0)z_\alpha\right)
= 1 - \Phi\left(z_\alpha - \sqrt{n_{t,i}}\theta_t \cdot \frac{\mu'_i(0)}{\sigma_i(0)}(1 + o(1))\right) + o(1)
\to 1 - \Phi(z_\beta).
\]

Therefore we necessarily have

\[
z_\alpha - \sqrt{n_{t,i}}\theta_t \cdot \frac{\mu'_i(0)}{\sigma_i(0)}(1 + o(1)) \to z_\beta.
\]

Now we can take limit:

\[
\lim_{t \to \infty} \frac{n_{t,2}}{n_{t,1}} = \lim_{t \to \infty} \frac{(z_\alpha - z_\beta)^2(\theta_t \frac{\mu'_i(0)}{\sigma_i(0)})^{-2}(1 + o(1))}{(z_\alpha - z_\beta)^2(\theta_t \frac{\mu'_i(0)}{\sigma_i(0)})^{-2}(1 + o(1))} = \left(\frac{\mu'_i(0)/\sigma_i(0)}{\mu'_i(0)/\sigma_i(0)}\right)^2,
\]

as desired. \( \square \)

**Example 8.20.** In this example, we compute Pitman efficiency between the sign test and \( t \)-test. The setup is as follows. Let \( Y_1, \ldots, Y_n \) be i.i.d. observations from a location family \( \{f(\cdot - \theta) : \theta \in \mathbb{R}\} \). Here \( f \) is a symmetric known density function (w.r.t. Lebesgue measure on \( \mathbb{R} \)) with finite variance. We would like to test \( H_0 : \theta = 0 \) versus \( H_1 : \theta > 0 \).

- Consider the sign statistic

\[
S_n = \frac{1}{n} \sum_{i=1}^{n} 1_{Y_i > 0}.
\]

Then

\[
\mu_S(\theta) = E_\theta S_n = P_\theta(Y_1 > 0) = 1 - F(-\theta)
\]

\[
\sigma^2_S(\theta) = \text{Var}_\theta(S_n) = \frac{1}{n} \text{Var}_\theta(1_{Y_1 > 0})
\]

\[
= \frac{1}{n} \left( P_\theta(Y_1 > 0) - P_\theta^2(Y_1 > 0) \right)
\]
\begin{align*}
&= \frac{1}{n} F(-\theta) (1 - F(-\theta)). \\
\text{Hence} \\
\frac{\sqrt{n}(S_n - \mu_S(\theta_n))}{\sigma_S(\theta_n)} \sim_{\theta_n} N(0,1), \\
\text{and} \\
\frac{\mu'_S(0)}{\sigma_S(0)} = 2f(0).
\end{align*}

- Consider the t-statistic

\[ T_n = \frac{\bar{Y}}{S}. \]

Then with \( \mu_T(\theta) = \theta/\sigma \) and \( \sigma_T(\theta) = 1, \)

\[
\frac{\sqrt{n}(T_n - \mu_T(\theta_n))}{\sigma_T(\theta_n)} = \frac{\sqrt{n}(\frac{\bar{Y}}{S} - \frac{\theta_n}{\sigma})}{\frac{\sqrt{n} \theta_n}{S}} + \frac{\sqrt{n} \theta_n}{S} \left( \frac{1}{S} - \frac{1}{\sigma} \right) \\
= \frac{\sqrt{n}(\bar{Y} - \theta_n)}{S} \cdot \frac{\sigma}{\sigma} + \theta_n \cdot \frac{\sqrt{n}(\sigma - S)}{\sigma S} \\
\sim_{\theta_n} N(0,1).
\]

So

\[
\frac{\mu'_T(0)}{\sigma_T(0)} = \frac{1}{\sqrt{\int y^2 f(y) \, dy}}.
\]

Hence the Pitman efficiency between the sign test and the t-test is given by

\[ p_{S,T}(f) = 4f^2(0) \int y^2 f(y) \, dy. \]

By choosing different density function \( f, \) we may calculate

\[
p_{S,T}(\text{unif}[-1,1]) = 1/3, \\
p_{S,T}(N(0,1)) = 2/\pi, \\
p_{S,T}(\text{Laplace}) = 2,
\]

so the relative efficiency between the sign test and the t-test depends highly on the properties of \( f. \)

8.5. **Confidence interval.**

**Definition 8.21.** (1) Let \( L(Y) \) and \( U(Y) \) be a pair of functions of the data \( Y \) satisfying \( L(Y) \leq U(Y) \) for all \( Y. \) The random interval \([L(Y), U(Y)]\) is called an interval estimator of a parameter \( \theta \) if upon observing \( Y = y, \) the inference \( L(y) \leq \theta \leq U(y) \) is made for parameter \( \theta. \)
(2) The probability \( P_\theta [L(Y) \leq \theta \leq U(Y)] \), i.e., the probability that the random interval \([L(Y), U(Y)]\) contains the true value of \( \theta \) is called the coverage probability. The coverage probability may depend on \( \theta \).

(3) The probability
\[
1 - \alpha = \inf_{\theta \in \Theta} P_\theta [L(Y) \leq \theta \leq U(Y)]
\]
is known as the confidence coefficient.

(4) An interval estimator with a confidence coefficient \( 1 - \alpha \) is referred to as a confidence interval (CI) with confidence coefficient \( 1 - \alpha \), or sometimes a \( 100(1 - \alpha)\% \) confidence interval.

**Example 8.22.** \( Y_1, \ldots, Y_n \sim N(\mu, \sigma^2) \). We are interested in estimating a 95\% confidence interval for \( \mu \). If \( \sigma^2 \) is known, then a \( 100(1 - \alpha)\% \) CI for \( \mu \) is \( \bar{Y} \pm \frac{z_{\alpha/2} \sigma}{\sqrt{n}} \), whereas if \( \sigma^2 \) is unknown, then a \( 100(1 - \alpha)\% \) CI for \( \mu \) is \( \bar{Y} \pm \frac{t_{n-1, \alpha/2} s}{\sqrt{n}} \), where \( s^2 \) is the sample standard deviation.

8.5.1. **Method-I for constructing interval estimators: inverting tests of hypotheses.** Starting with the hypothesis testing procedure, we can construct a confidence interval. For each \( \theta_0 \) in the parameter space, consider testing the null hypothesis \( H_0 : \theta = \theta_0 \) versus the alternative \( H_1 : \theta \neq \theta_0 \), and let \( A(\theta_0) \) denote a level-\( \alpha \) acceptance rejection (i.e., accept \( H_0 \) when \( Y \in A(\theta_0) \)), such that

\[
P_{\theta_0}(Y \in A(\theta_0)) \geq 1 - \alpha.
\]

Now, construct the set
\[
C(Y) = \{ \theta : Y \in A(\theta) \}.
\]
By construction, \( P_\theta \{ \theta \in C(Y) \} = P_\theta \{ Y \in A(\theta) \} \) for all \( \theta \). Since inequality (8.5) holds for any \( \theta \), we have
\[
\inf_{\theta} P_\theta(\theta \in C(Y)) = \inf_{\theta} P_\theta(Y \in A(\theta)) \geq 1 - \alpha.
\]
By definition, \( C(Y) \) is a \( 100(1 - \alpha)\% \) confidence interval for \( \theta \).

**Example 8.23.** So how can we derive the confidence intervals of Example 8.22 by inverting a hypothesis test? Consider testing \( H_0 : \mu = \mu_0 \) versus \( H_1 : \mu \neq \mu_0 \) when \( \sigma^2 \) is unknown. A level \( \alpha \) acceptance region is
\[
A(\mu_0) = \left\{ Y : \left| \frac{\sqrt{n}(\bar{Y} - \mu_0)}{s} \right| \leq t_{n-1, \alpha/2} \right\}.
\]
For any \( \mu, Y \in A(\mu) \) is equivalent to \( \bar{Y} - (s/\sqrt{n})t_{n-1, \alpha/2} \leq \mu \leq \bar{Y} + (s/\sqrt{n})t_{n-1, \alpha/2} \). Thus we arrive at the \( 100(1 - \alpha)\% \) CI of Example 8.22.
8.5.2. Method-II for constructing interval estimators: using pivotal quantities.

Definition 8.24 (Pivotal quantity or Pivot). A pivotal quantity, or simply a “pivot” is a function \( Q(Y, \theta) \) whose distribution does not depend on \( \theta \).

Note that a pivot is not a statistic as it involves \( \theta \), and thus should not be confused with an ancillary statistic.

To find a \( 100(1-\alpha) \%) \) CI for \( \theta \) using the pivot \( Q(Y, \theta) \), find a set \( A \) that satisfies
\[
P_\theta \{ Q(Y, \theta) \in A \} \geq 1 - \alpha.
\]
Thus the set
\[
\{ \theta : Q(Y, \theta) \in A \}
\]
is a \( 100(1-\alpha) \%) \) CI for \( \theta \), provided \( Q(Y, \theta) \) is a monotone function of \( \theta \) for fixed \( Y \) (why is monotonicity needed?)

Example 8.25. In example 8.22, \( Q(Y, \mu) \equiv \frac{(Y-\mu)\sqrt{n}}{\sigma} \sim N(0,1) \) is a pivot if \( \sigma^2 \) is known. Now we can make obtain a frequentist confidence claim using
\[
P(-z_{\alpha/2} < Q(Y, \mu) < z_{\alpha/2}) = 1 - \alpha.
\]
Solving for \( \mu \) then returns the same interval as before, \( Y \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \). If \( \sigma^2 \) is unknown, for inference about \( \mu \) it is very useful having a pivot which does not involve \( \sigma^2 \), and then a very widely-used choice is
\[
Q(Y, \mu) \equiv \frac{(\bar{Y} - \mu)\sqrt{n}}{s} \sim t_{n-1},
\]
where \( s^2 \) is the (unbiased) sample variance.

8.5.3. Method-III for constructing interval estimators: inverting the CDF of a sufficient statistic. Let \( T = T(Y) \) be a sufficient statistic having a continuous distribution with a cdf \( F_\theta(t) \). Assume that \( F_\theta(t) \) is a monotonically decreasing function of \( \theta \) for every \( t \).

Choose \( 0 < \alpha_1 < 1 \) and \( 0 < \alpha_2 < 1 \), such that \( \alpha_1 + \alpha_2 = \alpha \). Invert the interval \( \alpha_1 \leq F_\theta(T) \leq 1 - \alpha_2 \) to obtain \( L(T) \leq \theta \leq U(T) \), where \( F_{U(T)}(t) = \alpha_1 \) and \( F_{L(T)}(t) = 1 - \alpha_2 \) (see Figure 1). Then \( [L(T), U(T)] \) is a \( 100(1-\alpha) \%) \) CI for \( \theta \) because
\[
P_\theta \{ L(T) \leq \theta \leq U(T) \} = P_\theta \{ \alpha_1 \leq F_\theta(T) \leq 1 - \alpha_2 \} = 1 - \alpha_2 - \alpha_1 = 1 - \alpha.
\]
Note that \( P_\theta \{ \alpha_1 \leq F_\theta(T) \leq 1 - \alpha_2 \} = 1 - \alpha_2 - \alpha_1 \) because by the probability integral transform (PIT), \( F_\theta(T) \) is Unif[0, 1].

If \( F_\theta(t) \) is a monotonically increasing function of \( \theta \) for every \( t \), then the same procedure will lead to generation of the interval \( L(T) \leq \theta \leq U(T) \), where \( F_{U(T)}(t) = 1 - \alpha_2 \) and \( F_{L(T)}(t) = \alpha_1 \).

Example 8.26. Let \( Y_1, \ldots, Y_n \) be an iid random sample from a location exponential family of distributions having pdf
\[
f_\mu(y) = e^{-(y-\mu)}, \quad y \geq \mu.
\]
To find a CI for $\mu$ using the method described above, we first need to find a sufficient statistic for $\mu$. It is easy to see that $T = Y_{(1)} = \min\{Y_1, \ldots, Y_n\}$ is sufficient for $\mu$. Now, we have to find the CDF of $T$. Denoting this CDF by $F_\mu(t)$, for $t \geq \mu$,

$$F_\mu(t) = P_\mu \{T \leq t\} = P_\mu \{Y_{(1)} \leq t\} = 1 - P_\mu \{Y \geq t\}^n = 1 - e^{-n(t-\mu)},$$

where the last equality follows by noting that the CDF of the data-generating distribution is $1 - e^{-(y-\mu)}$, $y \geq \mu$. Now note that for fixed $t$, $F_\mu(t)$ is a decreasing function of $\mu$. Using the construction method described above, choose $\alpha_1 = \alpha_2 = \alpha/2$, where $\alpha \in (0, 1)$, and define $L(T)$ and $U(T)$ such that $F_{U(T)}(t) = \alpha_1 = \alpha/2$ and $F_{L(T)}(t) = 1 - \alpha_2 = 1 - \alpha/2$. This yields

$$L(Y) = T + \frac{1}{n} \log(\alpha/2),$$
$$U(Y) = T + \frac{1}{n} \log(1 - \alpha/2)$$

and $[L(T), U(T)]$ is a $100(1 - \alpha)$% CI for $\mu$ by the argument earlier in this subsection.

8.5.4. Asymptotic intervals. We now consider some asymptotic approaches to construct CIs. Analogous to the two approaches for finding CIs described in Sections 8.5.1 and 8.5.2, the following two approaches can be adopted to find asymptotic confidence intervals:
(i) Find approximate or large sample distribution of pivots and use them to construct intervals.

(ii) Invert approximate level-α tests (e.g., using asymptotic distribution of the likelihood ratio statistic).

Example 8.27. Consider the problem of finding an approximate $100(1 - \alpha)\%$ interval for a Poisson parameter $\lambda$ using an iid random sample $Y_1, \ldots, Y_n$.

The asymptotic distribution of the MLE $\hat{\lambda} = \bar{Y}$ can be used for this purpose. Recall that for the Pois($\lambda$) distribution, $I_1(\lambda) = 1/\lambda$ and hence

$$\sqrt{n}(\hat{\lambda} - \lambda) \xrightarrow{d} N(0, \lambda).$$

Asymptotic CI for $\lambda$ can be obtained by using the fact that

$$\frac{\sqrt{n}(\hat{\lambda} - \lambda)}{\sqrt{\lambda}} \sim N(0, 1).$$

This follows by (a) $\hat{\lambda} \xrightarrow{p} \lambda$ by consistency of MLE and (b) Slutsky’s theorem. Thus, an approximate $100(1 - \alpha)\%$ can be obtained as:

$$C(Y) = \left\{ \lambda : \left| \frac{\sqrt{n}(\hat{\lambda} - \lambda)}{\sqrt{\lambda}} \right| \leq z_{\alpha/2} \right\}.$$

$$= \left\{ \lambda : \bar{Y} - \sqrt{\bar{Y}/n} z_{\alpha/2} \leq \lambda \leq \bar{Y} + \sqrt{\bar{Y}/n} z_{\alpha/2} \right\}. $$
9. Bayesian Inference

Unlike frequentist inference, in Bayesian inference the unknown parameter $\theta$ is viewed as a random variable. Any information available about the parameter is captured through a so-called “prior” distribution denoted by $\pi(\theta)$. In a Bayesian setting, we will denote the probability model by $f(y|\theta)$ (instead of the frequentist notation of $f_\theta(y)$), to denote that the model is conditional on the random variable $\theta$. Let $Y_1, \ldots, Y_n \sim f(y|\theta)$ be the observed data from the model. Then the likelihood function $L(\theta|Y)$ is the joint distribution of $f(Y|\theta)$. By application of Baye’s theorem, the posterior distribution of $\theta$ given the data $Y$ is

$$
\pi(\theta|Y) = \frac{L(\theta|Y)\pi(\theta)}{\int L(\theta|Y)\pi(\theta)d\theta} \propto L(\theta|Y)\pi(\theta).
$$

(9.1)

The denominator of (9.1) is known as the normalizing constant. Inference about $\theta$ is made on the basis of the posterior distribution $\pi(\theta|Y)$. The data is expected to “shrink” the prior distribution, thereby leading to more precise inference from the posterior distribution.

9.1. Conjugate Priors. Loosely speaking, a prior is conjugate to a certain statistical model if it yields the same posterior distribution as the prior. This definition can be formalized as follows:

**Definition 9.1.** The class of priors $\Pi = \{\pi : \theta \in \Theta\}$ is conjugate to the model $\mathcal{F} = \{f(\cdot|\theta) : \theta \in \Theta\}$ if and only if $\pi(\cdot|y) \in \Pi$ for any $y$.

**Example 9.2.** Let $Y \sim \text{Bin}(n, \theta)$ with $n$ known. The probability model is $f(y|\theta) = \binom{n}{y} \theta^y (1-\theta)^{n-y}$. Consider the prior $\theta \sim \text{Beta}(\alpha, \beta)$, that is

$$
\pi(\theta) = \frac{1}{\text{Beta}(\alpha, \beta)} \theta^{\alpha-1} (1-\theta)^{\beta-1}, \quad 0 < \theta < 1, \alpha > 0, \beta > 0.
$$

Then

$$
\pi(\theta|y) \propto \pi(\theta)f(y|\theta) \propto \theta^{y+\alpha-1}(1-\theta)^{n-y+\beta-1},
$$

which is the pdf of a $\text{Beta}(y+\alpha, n-y+\beta)$ distribution.

Conjugate priors have the following advantage:

(i) Posterior distributions are tractable and can be obtained in closed-form. This is, however, no longer a strong advantage due to the extraordinary development of MCMC methods.

(ii) Natural appeal: Prior and posterior distributions are alike, and this leads to a clear visualization of shrinkage.

(iii) Historical significance.

(iv) The mathematical theory for exponential family of distributions is elegant, leading to nice interpretations of the results.

**Example 9.3.** Consider the natural exponential family (NEF) with pdf $f(y|\eta) = \exp(\eta y - \psi(\eta))h(y)$, where $y$ is the natural observation, and let $Y_1, \ldots, Y_n \sim f(y|\eta)$. Then $T = \sum_{i=1}^n Y_i$ is sufficient for $\eta$. Suppose our
interest lies in inferring about the mean $\mu = E(Y|\eta) = \psi'(\eta)$. Recall that $\text{var}(Y|\eta) = \psi''(\eta) = V(\mu)$, say.

The likelihood can be written in terms of the sufficient statistic $T$ as

$$L(\eta|Y) \propto \exp \left\{ \eta \sum_{i=1}^{n} Y_i - n\psi(\eta) \right\} \propto \{n(T - \psi(\eta))\}.$$  

Consider the prior pdf

$$\pi(\eta) \propto \exp \left\{ r(\eta\mu_0 - \psi(\eta)) \right\}.$$  

From (9.2) and (9.3), it follows that

$$\pi(\eta|Y) \propto \exp \left\{ (nT + r\mu_0)\eta - (n + r)\psi(\eta) \right\}.$$  

Note that the prior (9.3) and the posterior (9.3) are of the same form. Let us study the properties of this family of PDFs more carefully. Note that, from (9.3),

$$\pi(\mu) \propto \exp \left\{ r \int (\mu - \psi'(\eta)) d\eta \right\}.$$  

Let us write this PDF in terms of the mean $\mu = \psi'(\eta)$. Because $d\mu/d\eta = \psi''(\eta) = V(\mu)$, $d\eta = d\mu/V(\mu)$, and the Jacobian of transformation is $1/V(\mu)$. Thus,

$$\pi(\mu) \propto \frac{1}{V(\mu)} \exp \left\{ -r \int \frac{\mu - \mu_0}{V(\mu)} d\mu \right\},$$

which is a location-scale family with location parameter $\mu_0$ (the prior mean of $\mu$). We may want to interpret the prior as previously observed data with mean $\mu_0$ and sample size $r$.

To obtain the posterior distribution, note that from (9.4), we can write

$$\pi(\eta|Y) \propto \exp \left[ \int \{ (nT + r\mu_0) - (n + r)\psi'(\eta) \} d\eta \right].$$

Transforming $\mu = \psi'(\eta)$, we get

$$\pi(\mu|Y) \propto \frac{1}{V(\mu)} \exp \left[ \int \frac{-n + r + (nT + r\mu_0)}{V(\mu)} \left\{ \frac{\mu - r\mu_0 + nT}{n + r} \right\} d\mu \right].$$

Hence, the posterior mean is

$$E(\mu|Y) = \frac{r\mu_0 + nT}{n + r} = \left( \frac{r}{n + r} \right) \mu_0 + \left( \frac{n}{n + r} \right) T$$

$$= B\mu_0 + (1 - B)T,$$

where

$$B = \frac{r}{r + n}.$$
Note that the posterior mean is a convex combination of prior means $\mu_0$ and observed mean $T = \bar{Y}$. In fact the weighting represents the relative Fisher information of the ‘estimates’ $T$ and $\mu_0$ of the posterior mean i.e.

$$B = \frac{\frac{\tau}{\sigma^2}}{\frac{\tau}{\sigma^2} + \frac{n}{\sigma^2}}$$

where $\frac{\tau}{\sigma^2}$ is the Fisher information in the prior and $\frac{n}{\sigma^2}$ is the Fisher information in the data. Intuitively $B$ is the shrinkage factor and tells us how much the the prior mean $\mu_0$ contributes to the posterior mean compared to the data $Y$.

**Example 9.4.** In the case where $Y$ is normally distributed we know that $V(\mu) = c$, so we have

$$\int \frac{\mu - \mu_0}{V(\mu)} \, d\mu = \frac{1}{2c} \mu^2 - \frac{\mu_0}{c} \mu$$

and we see that after completing the square our conjugate prior on $\mu$ will be a Normal distribution.

**Example 9.5.** If $Y$ has a Poisson distribution then $V(\mu) = \mu$ and

$$\int \frac{\mu - \mu_0}{V(\mu)} \, d\mu = \mu - \mu_0 \log \mu.$$ 

Thus we can see the conjugate prior will be a Gamma distribution on $\mu$.

### 9.2. Bayesian Point and Interval Estimation.

The Bayesian approach focuses on $p(\theta|Y)$ instead of $p(Y|\theta)$; that is, on the probability of parameters given data instead of the probability of data given parameters. Thus, a point estimator of $\theta$ is a single quantity that “best represents” the posterior distribution. The expectation or the mode of the posterior distribution are two popular and common choices of point estimators of $\theta$.

A common Bayesian interval estimate is a *probability interval* (probability intervals are sometimes called by the weak name “credible intervals” or the wordy name “Bayesian confidence intervals”; we prefer to simply call them “probability intervals”). This means that our interval again includes $\theta$ with some set probability; however, we now take this probability over the posterior distribution of $\theta$ (holding $Y$ fixed), as opposed to the distribution of $Y|\theta$. Thus, while we still evaluate $P(\theta \in T(Y))$, this is now a probability statement about $\theta$ instead of $Y$. It should be noted that, while these intervals are constructed in a Bayesian way, they can still be evaluated using frequentist criteria.

**Example 9.6** (Numerical examples for point and interval estimation for the normal model). To illustrate the ideas of point and interval estimation, we consider a single observation $y = 120$ assumed to be obtained from a normal population with known variance $\sigma^2 = 100$ but unknown mean $\mu$ (to be inferred).
The frequentist point estimator of \( \mu \) is the observed value \( y = 120 \). However, if one assumes a conjugate prior \( \mu \sim N(\mu_0, \tau^2) \) where \( \mu_0 = 100 \) and \( \tau^2 = 225 \), then the “prior sample size” \( r \) defined the context of conjugate priors for NEF distributions is \( \sigma^2/\tau^2 = 100/225 = 0.44 \). Consequently \( B = r/\tau \approx 0.31 \) and the point estimate of \( \mu \) equals \( B\mu_0 + (1-B)y = .31 \times 100 + .69 \times 120 = 113.8 \). The variance of the posterior distribution of \( \mu \) is \( \sigma^2/(r + n) = 100/1.44 = 69.44 \) (notice the shrinkage from the prior variance of 225). Moving to interval estimation, a 95% frequentist confidence interval for \( \mu \) is \( y \pm 1.96\sigma \), i.e., \([100.4, 139.6]\). With the conjugate prior, the posterior distribution of \( \mu \) given \( y \) is normal with mean 113.8 and variance 69.44 as seen in the previous paragraph, a 95% probability interval is \( 113.8 \pm 1.96 \times \sqrt{69.44} \) or \([97.56, 130.22]\).

### 9.3. Bernstein-von Mises theorem

The classical Bernstein-von Mises theorem is a Bayesian version of asymptotic normality for the posterior distribution in parametric models from a frequentist point of view. It is usually referred to as a frequentist theory since a frequentist model is assumed (as we did in previous sections, but a pure Bayesian does not necessarily need to work with this apriori modeling). We will not go into unfruitful philosophical discussion as to if such a frequentist framework is reasonable in terms of ‘statistical thinking’ and ‘scientific interpretation’, but will work under this frequentist model assumption to see how far we can draw useful distributional approximation results from a large sample point of view.

Suppose samples \( Y_1, \ldots, Y_n \sim_{i.i.d.} P_{\theta_0} \) are from the model \( (P_\theta : \theta \in \Theta) \) with density functions \( (f_\theta : \theta \in \Theta) \) (note that we switch here from \( f(\cdot|\theta) \) to \( f_\theta \) for notational convenience) with respect to a dominating measure \( \mu \), and a prior \( \Pi \) on \( \Theta \) with density \( \pi \) with respect to the Lebesgue measure. Then the posterior distribution is given by (more explicitly)

\[
\pi(\theta|Y) = \frac{f_\theta^N(Y)\pi(\theta)}{\int f_\theta^N(Y)\pi(\theta) \, d\theta}.
\]

What asymptotics should we expect for the posterior distribution \( \Pi(:,|Y) \)?

Let us reparametrize \( \theta = \theta_0 + h/\sqrt{n} \), where \( h \) is a local parameter. Then the posterior distribution for \( h \) is given by

\[
\pi^h(h|Y) = \frac{\prod_{i=1}^n f_{\theta_0+h/\sqrt{n}}(Y_i)\pi(\theta_0 + h/\sqrt{n})}{\int \prod_{i=1}^n f_{\theta_0+h/\sqrt{n}}(Y_i)\pi(\theta_0 + h/\sqrt{n}) \, dh} = \frac{d\Pi_{\theta_0+h/\sqrt{n}}(Y)\pi(\theta_0 + h/\sqrt{n})}{\int d\Pi_{\theta_0+h/\sqrt{n}}(Y)\pi(\theta_0 + h/\sqrt{n}) \, dh} \approx \frac{dN(h,I_{\theta_0}^{-1})(Y)\pi(\theta_0)}{\int dN(h,I_{\theta_0}^{-1})(Y)\pi(\theta_0) \, dh}.
\]
\[
\frac{dN(h, I_{\theta_0}^{-1})(Y)}{\int dN(h, I_{\theta_0}^{-1})(Y) \, dh} = dN(h, I_{\theta_0}^{-1})(Y).
\]

So we expect the posterior distribution of \(\sqrt{n}(\theta - \theta)\) is \(N(Y, I_{\theta_0}^{-1})\) (in a sense to be rigorously defined below).

Now we formally state the Bernstein-von Mises theorem.

**Theorem 9.7.** Suppose that \((P_\theta : \theta \in \Theta)\) is differentiable in quadratic mean, and the density of the prior \(\pi\) on \(\Theta\) (w.r.t. Lebesgue measure) is continuously positive around \(\theta_0\). Further suppose that for every \(\varepsilon > 0\), there exists a sequence of tests \(\{\phi_n\}\) such that

\[
P_{\theta_0} \phi_n \to 0, \quad \sup_{\|\theta - \theta_0\| \geq \varepsilon} P_\theta(1 - \phi_n) \to 0.
\]

Then

\[
\sup_B |\Pi(\sqrt{n}(\theta - \theta_0) \in B | Y) - P\left(N(I_{\theta_0}^{-1}G_n \dot{\theta}_0, I_{\theta_0}^{-1}) \in B\right)\| \to P_{\theta_0} 0,
\]

where the supremum is taken over all Borel measurable sets \(B\).

Since the MLE \(\hat{\theta}_n^{MLE}\) has an asymptotic expansion

\[
\sqrt{n}(\hat{\theta}_n^{MLE} - \theta_0) = I_{\theta_0}^{-1}G_n \dot{\theta}_0 + o_P(1),
\]

under regularity conditions, by rescaling the total variational metric, we may write the result more compactly:

\[
d_{TV}\left(\Pi(\cdot | Y), N(\hat{\theta}_n^{MLE}, I_{\theta_0}^{-1} / n)\right) \to P_{\theta_0} 0.
\]

This leads to the usual interpretation of Bernstein-von Mises theorem: the posterior distribution resembles the maximum likelihood estimator in large samples.

A classical condition of verifying (9.6) is to use a uniformly consistent estimator \(\hat{\theta}_n\) in the sense that for any \(\varepsilon > 0\),

\[
\sup_\theta P\left(\|\hat{\theta}_n - \theta\| > \varepsilon\right) \to 0.
\]

Then we may define \(\phi_n = 1(\|\hat{\theta}_n - \theta_0\| > \varepsilon/2)\) to verify (9.6).

**Proposition 9.8.** Suppose \(\Theta \subseteq \mathbb{R}^k\) is compact. Further assume that \((P_\theta : \theta \in \Theta)\) is identifiable and the map \(\theta \mapsto P_\theta\) is continuous with respect to the total variational distance \(d_{TV}\). Then there exists a sequence of uniformly consistent estimators.

**Lemma 9.9.** Suppose there exists a class \(\mathcal{F}\) containing functions defined on \(\Theta\) such that the following hold for any \(\varepsilon > 0\):

\[
(1) \inf_{\|\theta - \theta'\| > \varepsilon} \sup_{f \in \mathcal{F}} |(P_\theta - P_{\theta'})(f)| > 0;
\]
sup_{\theta \in \Theta} P_{\theta} \left( \sup_{f \in \mathcal{F}} |(P_{\hat{\theta}_n} - P_{\theta})(f)| > \varepsilon \right) \to 0, \text{ where the supremum is taken over all probability measures on } \Theta^1.

Then there exists a sequence of uniformly consistent estimators.

Proof. Let \( \hat{\theta}_n \) be the near minimizer of the map \( \theta \mapsto \sup_{f \in \mathcal{F}} |(P_{\theta} - P_{\hat{\theta}_n})(f)| \) in the sense that
\[
\sup_{f \in \mathcal{F}} |(P_{\theta} - P_{\hat{\theta}_n})(f)| \leq \inf_{\theta \in \Theta} \sup_{f \in \mathcal{F}} |(P_{\theta} - P_{\hat{\theta}_n})(f)| + \frac{1}{n}.
\]
Then
\[
\sup_{f \in \mathcal{F}} |(P_{\theta} - P_{\hat{\theta}_n})(f)| \leq \sup_{f \in \mathcal{F}} |(P_{\theta} - P_{\hat{\theta}_n})(f)| + \frac{1}{n}.
\]
For fixed \( \varepsilon > 0 \), let
\[
\delta_\varepsilon \equiv \inf_{\|\theta - \theta'\| > \varepsilon} \sup_{f \in \mathcal{F}} |(P_{\theta} - P_{\theta'})(f)| > 0.
\]
Then for any \( \theta \in \Theta \), and any \( \varepsilon > 0 \),
\[
P_{\theta}(\|\hat{\theta}_n - \theta\| > \varepsilon)
\leq P_{\theta}\left( \sup_{f \in \mathcal{F}} |(P_{\hat{\theta}_n} - P_{\theta})(f)| \geq \delta_\varepsilon \right)
\leq P_{\theta}\left( \sup_{f \in \mathcal{F}} |(P_{\theta} - P_{\theta})(f)| \geq \frac{1}{2} \left( \delta_\varepsilon - \frac{1}{n} \right) \right)
\leq \sup_{P\in \mathcal{P}} \sup_{f \in \mathcal{F}} |(P_{\theta} - P)(f)| \geq \frac{1}{2} \left( \delta_\varepsilon - \frac{1}{n} \right) \to 0,
\]
where the last inequality follows from condition (2). The proof is complete. \qed

Lemma 9.10. Let \( X,Y \) be metric space and \( X \) is furthermore compact. If \( f : X \to Y \) is a continuous bijection, then \( f^{-1} \) is uniformly continuous.

Proof. Since \( X \) is compact and \( f \) is continuous, \( Y = f(X) \) is compact. So we only need to show that \( f \) is a closed map, i.e. for any \( K \subset X \) closed, we need to show that \( f(K) \subset Y \) is closed. This can be argued as follows: since \( K \) is closed in a compact space, it is compact, and therefore \( f(K) \) is compact by continuity of \( f \). On the other hand, \( Y \) is a metric space and hence Hausdorff. Any compact set in Hausdorff space must be closed, so we have proved the claim. \qed

\(^1\)This is usually called the uniform Glivenko-Cantelli property since it requires a weak law of large number holds uniformly in \( P \in \mathcal{F} \) and in the underlying law \( P \).
Proof of Proposition 9.8. We will construct a class $\mathcal{F}$ that verifies the conditions of Lemma 9.9.

We first claim that there exist $\{A_i \subset \mathbb{R}^k\}_{i=1}^{\infty}$ such that for any $\theta, \theta' \in \Theta$, if $P_{\theta}(A_i) = P_{\theta'}(A_i)$ for all $i \in \mathbb{N}$, then $\theta = \theta'$. To see this, we may take $A_i = (-\infty, t]$ where $t$ ranging over all vectors in $\mathcal{Q}^k$. Then by continuity of measure, if the condition of the claim holds, then $P_{\theta} = P_{\theta'}$ on all $(-\infty, t]$ with $t$ ranging over $\mathbb{R}^k$, and therefore on all rectangles. Now $\pi - \lambda$ theorem implies that $P_{\theta} = P_{\theta'}$ over all Borel sets. By identifiability, $\theta = \theta'$, proving the claim.

Let $\mathcal{F} \equiv \{f_i : f_i(x) = i^{-1}1_{A_i}(x)\}$. Then the map

$$h : (\Theta, ||\cdot||) \rightarrow (\ell^\infty(\mathcal{F}), ||\cdot||_{\infty})$$

$$\theta \mapsto (f \mapsto P_{\theta}f)$$

is continuous and one-to-one. One-to-one follows immediately from the construction of the sets $\{A_i\}$. Continuity follows from the assumption that $\theta \mapsto P_{\theta}$ is continuous under total variational norm. Now Lemma 9.10 implies that $h^{-1} : \ell^\infty(\mathcal{F}) \supset h(\Theta) \rightarrow \Theta$ is uniformly continuous. This means that for any $\varepsilon > 0$, there exists some $\delta > 0$ such that if $\theta, \theta'$ are such that

$$\sup_{f \in \mathcal{F}} |(h(\theta) - h(\theta'))(f)| = \sup_{f \in \mathcal{F}}|(P_{\theta} - P_{\theta'})(f)| \leq \delta,$$

then $||\theta - \theta'|| \leq \varepsilon$. Equivalently, for any $\theta, \theta'$ such that $||\theta - \theta'|| > \varepsilon$,

$$\sup_{f \in \mathcal{F}}|(P_{\theta} - P_{\theta'})(f)| > \delta.$$ This verifies condition (1) of Lemma 9.9.

For condition (2) of Lemma 9.9, for any $\varepsilon > 0$,

$$Pr_P\left(\sup_{f \in \mathcal{F}} |(P_n - P)(f)| > \varepsilon\right)$$

$$\leq \sum_{i=1}^{\infty} Pr_P\left(|(P_n - P)(f_i)| > \varepsilon\right)$$

$$\leq \sum_{i=1}^{\infty} \frac{1}{\varepsilon^2} E_P\left[(P_n - P)^2\right] \leq \sum_{i=1}^{\infty} \frac{1}{n t^2 \varepsilon^2} \rightarrow 0$$

as $n \rightarrow \infty$ uniformly in $P$. Hence the conditions of Lemma 9.9 are verified and the proof is complete. \qed

9.4. Proof of Theorem 9.7. We need the following lemma to get an exponential control for the type II error.

Lemma 9.11. Suppose the conditions of Theorem 9.7 hold. Then for every $M_n \nearrow \infty$, there exists a sequence of tests $\{\phi_n\}$ and a constant $c > 0$ such that

$$P_{\theta_0}\phi_n \rightarrow 0,$$

$$\sup_{||\theta - \theta_0|| \geq M_n/\sqrt{n}} P_{\theta}(1 - \phi_n) \leq e^{-cn(||\theta - \theta_0||^2 \wedge 1)}.$$
To prove the lemma, we need the following well-known concentration inequality. The proof of this inequality can be found in any textbook on concentration inequality; we provide one for completeness.

**Lemma 9.12** (Hoeffding’s inequality). Suppose \( X_1, \ldots, X_n \) are independent random variables taking values in \([a, b]\), then

\[
P(\bar{X} - E\bar{X} > t) \leq \exp \left( -\frac{2nt^2}{(b-a)^2} \right).
\]

**Proof.** We assume that \( EX_1 = 0 \) and \( a < X < b \) without loss of generality. Let \( L(\lambda) \equiv \log Ee^{\lambda X_1} \) be the logarithm of the moment generating function of \( X \). Then \( L(0) = \log 1 = 0 \), and \( L'(\lambda) = (EX_1e^{\lambda X_1})/Ee^{\lambda X_1} \). Hence

\[
L'(0) = EX_1 = 0.
\]

Furthermore,

\[
L''(\lambda) = \int x^2 \cdot \frac{e^{\lambda x}}{Ee^{\lambda X_1}} \, dP(x) - \left( \int x \cdot \frac{e^{\lambda x}}{Ee^{\lambda X_1}} \, dP(x) \right)^2,
\]

where the measure \( Q \) is defined by \( dQ(x) = \frac{e^{\lambda x}}{Ee^{\lambda X_1}} \, dP(x) \). It is easy to verify that \( Q \) is a probability measure. On the other hand, since \( |X(\omega) - EQX| \leq (b-a)/2 \) for all \( \omega \in \Omega \), it follows that \( L''(\lambda) \leq (b-a)^2/4 \) for any \( \lambda \geq 0 \).

Now by Taylor expansion, we have for some \( \lambda^* \in [0, \lambda] \)

\[
L(\lambda) = L(0) + \lambda L'(0) + \frac{\lambda^2}{2} L''(\lambda^*) \leq \frac{\lambda^2(b-a)^2}{8}.
\]

This implies that for any \( \lambda \geq 0 \), with \( S_n = \sum_{i=1}^n X_i \),

\[
Ee^{\lambda S_n} = \prod_{i=1}^n Ee^{\lambda X_i} \leq \exp(n\lambda^2(b-a)^2/8).
\]

So for any \( u \geq 0 \), and \( \lambda \geq 0 \),

\[
P(S_n \geq u) = P(e^{\lambda S_n} \geq e^{\lambda u}) \leq e^{-\lambda u} Ee^{\lambda S_n} \leq \exp \left( -\lambda u + \lambda^2 n(b-a)^2/8 \right).
\]

Choosing \( \lambda = 4u/n(b-a)^2 \) to minimize the right hand side of the above display so that we arrive at

\[
P(S_n \geq u) \leq \exp \left( -\frac{2u^2}{n(b-a)^2} \right).
\]

The proof is complete by letting \( u = nt \). \qed

**Proof of Lemma 9.11.** We consider separately the following two regimes:

1. \( M_n/\sqrt{n} \leq \|\theta - \theta_0\| \leq \varepsilon \).
Here $\epsilon > 0$ is a small enough constant to be chosen later. First we consider the regime (1). Let $\ell_{\theta_0}^L \equiv (\ell_{\theta_0} \vee (-L)) \wedge L$ (truncated at level $L$), where $L > 0$ is a large enough constant such that $P_{\theta_0} \ell_{\theta_0}^L (\ell_{\theta_0})^\top$ is non-singular (this is viable by dominated convergence theorem). For such $L$, define

$$\phi_n^{(1)} \equiv 1 \left( \left\| (P_n - P_{\theta_0}) \ell_{\theta_0}^L \right\| \geq \sqrt{M_n/n} \right).$$

Now for type I error, by CLT

$$P_{\theta_0} \phi_n^{(1)} = P_{\theta_0} \left( \left\| G_n \ell_{\theta_0}^L \right\| \geq \sqrt{M_n} \right) \to 0.$$

For type II error, using triangle inequality we have on the event $\phi_n^{(1)} = 0$,

$$\left\| (P_n - P_{\theta_0}) \ell_{\theta_0}^L \right\| \geq \left\| (P_{\theta_0} - P_{\theta_0}) \ell_{\theta_0}^L \right\| - \left\| (P_n - P_{\theta_0}) \ell_{\theta_0}^L \right\| \geq c\|\theta - \theta_0\| - \sqrt{M_n/n}$$

for $n$ large enough. The second inequality in the above display follows by the following reasons:

1. By quadratic differentiability, $(P_{\theta_0} - P_{\theta_0}) \ell_{\theta_0}^L = (P_{\theta_0} \ell_{\theta_0}^L (\ell_{\theta_0})^\top + o(1)) (\theta - \theta_0)$ for $\|\theta - \theta_0\| \leq \epsilon$ with $\epsilon > 0$ small enough. On the other hand, we have chosen $L$ so that $P_{\theta_0} \ell_{\theta_0}^L (\ell_{\theta_0})^\top$ is non-singular, so for $\epsilon > 0$ small enough, the smallest eigenvalue of $P_{\theta_0} \ell_{\theta_0}^L (\ell_{\theta_0})^\top + o(1)$, say $c$, is positively bounded away from 0.

2. On the event $\phi_n^{(1)} = 0$, by construction we have $\left\| (P_n - P_{\theta_0}) \ell_{\theta_0}^L \right\| < \sqrt{M_n/n}$.

Hence, for $n$ large enough and $M_n/\sqrt{n} \leq \|\theta - \theta_0\| \leq \epsilon$ with $\epsilon > 0$ small enough, by Hoeffding’s inequality

$$P_\theta^n (1 - \phi_n^{(1)}) \leq P_\theta \left( \left\| (P_n - P_{\theta_0}) \ell_{\theta_0}^L \right\| \geq \frac{1}{2} c \left\| \theta - \theta_0 \right\| \right)$$

$$= P_\theta \left( \left\| (P_n - P_{\theta_0}) \ell_{\theta_0}^L \right\|^2 \geq \frac{1}{4} c^2 \left\| \theta - \theta_0 \right\|^2 \right)$$

$$\leq \sum_{j=1}^{\dim(\Theta)} P_\theta \left( \left\| (P_n - P_{\theta_0}) (\ell_{\theta_0})_j \right\|^2 \geq \frac{1}{4\dim(\Theta)} c^2 \left\| \theta - \theta_0 \right\|^2 \right)$$

$$\leq \exp(-C(c, L)n\|\theta - \theta_0\|^2).$$

Next we consider the regime (2). By assumption, there exists a sequence of tests $\{\hat{\phi}_n\}$ such that for the $\epsilon > 0$ chosen in the above arguments,

$$P_{\theta_0} \hat{\phi}_n \to 0, \quad \sup_{\|\theta - \theta_0\| > \epsilon} P_\theta^n (1 - \hat{\phi}_n) \to 0.$$
Now we want to accelerate the type II error rates. We will use a classical ‘bootstrap trick’. Fix $k$ large enough such that

$$P_{\theta_0}^k \tilde{\phi}_k \lor \sup_{\|\theta - \theta_0\| > \varepsilon} P_{\theta}^k (1 - \tilde{\phi}_k) \leq \frac{1}{4}.$$  

For $n$ large enough, we may write $n = m \cdot k + r$ where $0 \leq r \leq k - 1$. Now let

$$Z_{n,j} \equiv \tilde{\phi}_k (Y_{(j-1)k+1}, \ldots, Y_{jk}), \quad j = 1, \ldots, m,$$

and

$$\phi_n^{(2)} = 1 \left( \frac{1}{m} \sum_{j=1}^{m} Z_{n,j} \geq \frac{1}{2} \right).$$

The idea here is that under $P_{\theta_0}$, $E_{\theta_0} Z_{n,j} \leq 1/4$ while under $P_{\theta}$ where $\|\theta - \theta_0\| > \varepsilon$, $E_{\theta} Z_{n,j} \geq 3/4$, and $m$ accelerates the concentration. Indeed, for type I error, by Hoeffding’s inequality,

$$P_{\theta_0}^n \phi_n^{(2)} = P_{\theta_0}^n \left( Z_{n,m} \geq \frac{1}{2} \right) \leq e^{-C_1 m} \leq e^{-C_2 n}.$$  

Similarly we can control the type II error:

$$\sup_{\|\theta - \theta_0\| > \varepsilon} P_{\theta}^n (1 - \phi_n^{(2)}) \leq e^{-C_3 n}.$$  

Finally we let $\phi_n = \phi_n^{(1)} \lor \phi_n^{(2)}$. \hfill \Box

**Lemma 9.13.** Let $P$ be a probability measure defined on $(\mathcal{X}, \mathcal{A})$. For any $C \in \mathcal{A}$ with $P(C) > 0$, let $P^C$ be the measure of $P$ restricted to $C$, i.e. $P^C(B) \equiv P(B \cap C)/P(C)$. Then

$$\sup_B |P(B) - P^C(B)| \leq 2P(C^c).$$

Here the supremum is taken over all $B \in \mathcal{A}$.

**Proof.** Since

$$P(B) - P^C(B) = P(B) - \frac{P(B \cap C)}{P(C)}$$  

$$= P(B \cap C) + P(B \cap C^c) - \frac{P(B \cap C)}{P(C)}$$  

$$= \frac{P(B \cap C)P(C) - P(B \cap C)}{P(C)} + P(B \cap C^c)$$  

$$= - \frac{P(B \cap C)}{P(C)} P(C^c) + P(B \cap C^c),$$
it follows that

\[ |P(B) - P^C(B)| \leq \frac{P(B \cap C)}{P(C)} P(C^c) + P(B \cap C^c) \leq 2P(C^c), \]

as desired. \(\square\)

**Lemma 9.14.** Let \(P, Q\) be two probability measures such that \(P \ll Q\). Then

\[
d_{TV}(P, Q) = 2 \int \left( 1 - \frac{dP}{dQ} \right)_+ dQ.
\]

**Proof.** Let \(\mu\) be the dominating measure (which always exists). Using the well-known representation for total variational distance,

\[
d_{TV}(P, Q) = 2 \int (q - p) d\mu
\]

as desired. \(\square\)

**Proof of Theorem 9.7.** Let \(\pi_n\) be the prior on \(h\), i.e. \(\Pi_n(B) = \Pi(\theta_0 + B/\sqrt{n})\), where \(\Pi\) is the prior on \(\Theta\). Let \(C_n \equiv B(0, M_n)\) for some \(M_n \nearrow \infty\), and \(\Pi_n^C\) be the measure of \(\Pi_n\) restricted to \(C_n\).

**Step 1** In this step, we control the difference between \(\Pi_n(B/Y)\) and \(\Pi_n^C(B/Y)\) in total variational norm. By Lemma 9.13,

\[
(9.7) \quad \sup_B |\Pi_n(B/Y) - \Pi_n^C(B/Y)| \leq 2\Pi_n(C_n^c|Y).
\]

Fix a ball \(U\) around 0, and we have

\[
\int P^n_{\theta_0 + h/\sqrt{n}}(\Pi_n(C_n^c|Y)) d\Pi_n^U(h)
\]

\[
= \int P^n_{\theta_0 + h/\sqrt{n}}(\Pi_n(C_n^c|Y)(\phi_n + (1 - \phi_n))(Y)) d\Pi_n^U(h)
\]

\[
\leq \int \left( P^n_{\theta_0 + h/\sqrt{n} \phi_n} \right) d\Pi_n^U(h) + \int P^n_{\theta_0 + h/\sqrt{n}}(\Pi_n(C_n^c|Y)(1 - \phi_n)(Y)) d\Pi_n^U(h)
\]

\[
\equiv (I) + (II).
\]

For (I), quadratic differentiability implies uniform LAN, which in turn implies that \(P^n_{\theta_0 + h/\sqrt{n} \phi_n} \triangleright P^n_{\theta_0}\), so the assumption on type I error implies that

\[
(I) = \pi(\theta_0) (1 + o(1)) \cdot P_{\theta_0 + h/\sqrt{n} \phi_n} \rightarrow 0.
\]

Now consider (II). Let \(\varepsilon_0 > 0\) be such that \(\inf_{||\theta - \theta_0|| \leq \varepsilon_0} \pi(\theta) > 0\). Then,

\[
(II) \leq \int_U P^n_{\theta_0 + h/\sqrt{n}} \left[ \frac{f^n_{C_n \theta + h/\sqrt{n}}(Y) \pi(\theta)}{f^n_{\theta + h/\sqrt{n}}(Y) \pi(\theta)} \cdot (1 - \phi_n(Y)) \right] \cdot \frac{1}{\Pi_n(U)} d\Pi_n(h)
\]
\[
\frac{1}{\Pi_n(U)} \int_{C_n^u} P^n_{\theta_0 + h/\sqrt{n}}(\Pi_n(U|Y)(1 - \phi_n)(Y)) \, d\Pi_n(h)
\]
(by Fubini’s theorem)
\[
\leq \frac{1}{\Pi_n(U)} \int_{|h| > M_n} P^n_{\theta_0 + h/\sqrt{n}}(1 - \phi_n) \, d\Pi_n(h) \quad \text{(using } \Pi_n(U|Y) \leq 1) \]
\[
\leq \frac{1}{\Pi_n(U)} \int_{|h| > M_n} e^{-c(\|h\|^2 \wedge n)} \, d\Pi_n(h) \quad \text{(by Lemma 9.11)}
\]
\[
= \frac{1}{\Pi_n(U)} \left( \int_{M_n < |h| \leq \varepsilon_0 \sqrt{n}} + \int_{|h| > \varepsilon_0 \sqrt{n}} \right) e^{-c(\|h\|^2 \wedge n)} \, d\Pi_n(h)
\]
\[
\leq \int_{M_n < |h| \leq \varepsilon_0 \sqrt{n}} e^{-c_1 \|h\|^2} \, d\Pi_n(h) + \frac{1}{\Pi_n(U)} e^{-c_2 \varepsilon_0^2 n} \int_{|h| > \varepsilon_0 \sqrt{n}} \, d\Pi_n(h)
\]
\[
\leq \int_{|h| > M_n} e^{-c_1 \|h\|^2} \, dh + (\sqrt{n})^k e^{-c_2 \varepsilon_0^2 n} \to 0
\]
as \(n \to \infty\). The last inequality in the above display used the definition of \(\varepsilon_0\) for the first term to replace the measure \(\Pi_n^U\) by the Lebesgue measure on the domain \(|h| \leq \varepsilon_0 \sqrt{n}\), while using the fact that \(\Pi_n(U) = \Pi(\theta_0 + U/\sqrt{n}) \geq \inf_{\|h - \theta_0\| \leq \varepsilon_0} \pi(\theta) \cdot \text{vol}(U/\sqrt{n}) \geq (\sqrt{n})^{-k}\).

Hence, we have proved that
\[
\int P^n_{\theta_0 + h/\sqrt{n}} \left( \sup_B |\Pi_n(B|Y) - \Pi_n^C(B|Y)| \right) \, d\Pi_n^U(h) \to 0,
\]
which by the contiguity argument above, shows that
\[
\sup_B |\Pi_n(B|Y) - \Pi_n^C(B|Y)| \to p_{\theta_0} 0.
\]

(Step 2) Let \(C = B(0, M)\), and \(N^C(\mu, \Sigma)\) be the normal distribution \(N(\mu, \Sigma)\) restricted on \(C\). Then

\[
\frac{1}{2} \sup_B \left| P(N^C(I_{\theta_0}^{-1} \Gamma_n \hat{\theta}_0, I_{\theta_0}^{-1}) \in B) - \Pi_n^C(B|Y) \right|
\]
\[
\leq \int \left( 1 - \frac{dN^C(I_{\theta_0}^{-1} \Gamma_n \hat{\theta}_0, I_{\theta_0}^{-1})}{1_C(h) \cdot \int_{I_{\theta_0}^{-1} \Gamma_n \hat{\theta}_0, I_{\theta_0}^{-1}} \, d\Pi_n(h|Y)} \right) \, d\Pi_n(h|Y)
\]
\[
\leq \int \left( 1 - \frac{\pi(\theta_0 + g/\sqrt{n}) f^n_{\theta_0 + g/\sqrt{n}}(Y) dN^C(I_{\theta_0}^{-1} \Gamma_n \hat{\theta}_0, I_{\theta_0}^{-1})}{\pi(\theta_0 + h/\sqrt{n}) f^n_{\theta_0 + h/\sqrt{n}}(Y) dN^C(I_{\theta_0}^{-1} \Gamma_n \hat{\theta}_0, I_{\theta_0}^{-1})} \right) \times dN^C(I_{\theta_0}^{-1} \Gamma_n \hat{\theta}_0, I_{\theta_0}^{-1})(g) \, d\Pi_n(h|Y)
\]
\[
= \int k_n(g, h)(Y) \, dN^C(I_{\theta_0}^{-1} \Gamma_n \hat{\theta}_0, I_{\theta_0}^{-1})(g) \, d\Pi_n(h|Y).
\]
Since $G_n \ell_{\theta_0}$ is bounded in probability, it is natural to expect that $dN^C(I_{\theta_0}^{-1} G_n \ell_{\theta_0}, I_{\theta_0}^{-1})$ behaves like the Lebesgue measure restricted to $C$, i.e. $d\lambda^C$, with high probability. To make this formal, fix any $\varepsilon > 0$, we may find some large $L > 0$ such that the event $\Omega \equiv \{|I_{\theta_0}^{-1} G_n \ell_{\theta_0}, I_{\theta_0}^{-1}| \leq L\}$ with $P_{\theta_0}$ probability at least $1 - \varepsilon$. This means

$$
\int P_{\theta_0 + x/\sqrt{n}}^n \sup_B \left| P(N^C(I_{\theta_0}^{-1} G_n \ell_{\theta_0}, I_{\theta_0}^{-1}) \in B) - \Pi_n^C(B|Y) \right| 1_\Omega \ d\Pi_n^C(x) 
\approx \int k_n(g, h)(y) \ d\lambda^C(g) \ d\Pi_n^C(h|y) \ dP_{\theta_0 + x/\sqrt{n}}^n(y) d\Pi_n^C(x)
$$

$$
= \int k_n(g, h)(y) \left[ \Pi_n^C(h|y) \left( \int dP_{\theta_0 + x/\sqrt{n}}^n(y) d\Pi_n^C(x) \right) \right] d\lambda^C(g)
$$

$$
= \int k_n(g, h)(y) \left( \Pi_n^C(h) dP_{\theta_0 + h/\sqrt{n}}^n(y) \right) d\lambda^C(g) \quad \text{(by Bayes' formula)}
$$

$$
\approx \int k_n(g, h)(y) dP_{\theta_0 + h/\sqrt{n}}^n(y) d\lambda^C(g) d\lambda^C(h)
$$

$$
= \int \left( P_{\theta_0 + h/\sqrt{n}}^n k_n(g, h) \right) d\lambda^C(g) d\lambda^C(h).
$$

For every fixed $g, h$, since $k_n(g, h) \to 0$ under $P_{\theta_0}^n$, and hence by contiguity $k_n(g, h) \to 0$ under $P_{\theta_0 + h/\sqrt{n}}^n$. Since $0 \leq k_n(g, h) \leq 1$, convergence in probability implies convergence in $L_1$, and therefore $P_{\theta_0 + h/\sqrt{n}}^n k_n(g, h) \to 0$ for every $g, h$. Now apply dominated convergence theorem to conclude that

$$
\int P_{\theta_0 + x/\sqrt{n}}^n \sup_B \left| P(N^C(I_{\theta_0}^{-1} G_n \ell_{\theta_0}, I_{\theta_0}^{-1}) \in B) - \Pi_n^C(B|Y) \right| 1_\Omega \ d\Pi_n^C(x) \to 0.
$$

On the other hand, using contiguity and dominated convergence theorem again,

$$
\int P_{\theta_0 + x/\sqrt{n}}^n \sup_B \left| P(N^C(I_{\theta_0}^{-1} G_n \ell_{\theta_0}, I_{\theta_0}^{-1}) \in B) - \Pi_n^C(B|Y) \right| 1_{\Omega^c} \ d\Pi_n^C(x) 
\leq 2 \int P_{\theta_0 + x/\sqrt{n}}^n 1_{\Omega^c} \ d\Pi_n^C(x) \to 0,
$$

as $\varepsilon \to 0$. So we have proved

$$
\int P_{\theta_0 + x/\sqrt{n}}^n \sup_B \left| P(N^C(I_{\theta_0}^{-1} G_n \ell_{\theta_0}, I_{\theta_0}^{-1}) \in B) - \Pi_n^C(B|Y) \right| d\Pi_n^C(x) \to 0
$$

and hence by contiguity for every fixed $C > 0$

$$
\sup_B \left| P(N^C(I_{\theta_0}^{-1} G_n \ell_{\theta_0}, I_{\theta_0}^{-1}) \in B) - \Pi_n^C(B|Y) \right| \to P_{\theta_0}^n 0,
$$

so it holds for some $C_n = B(0, M_n)$ with $M_n \not\to \infty$ that

$$
\sup_B \left| P(N^{C_n}(I_{\theta_0}^{-1} G_n \ell_{\theta_0}, I_{\theta_0}^{-1}) \in B) - \Pi_n^{C_n}(B|Y) \right| \to P_{\theta_0}^n 0.
$$
(Step 3) Finally, by Lemma 9.13, with the $C_n$ chosen in step 2,
\[
\sup_B \big| P(N^C_n(I^{-1}_{\hat{\theta}_0}G_n\hat{\ell}_{\hat{\theta}_0}, I^{-1}_{\hat{\theta}_0}) \in B) - P(N(I^{-1}_{\hat{\theta}_0}G_n\hat{\ell}_{\hat{\theta}_0}, I^{-1}_{\hat{\theta}_0}) \in B) \big| \\
\leq 2P(N(I^{-1}_{\hat{\theta}_0}G_n\hat{\ell}_{\hat{\theta}_0}, I^{-1}_{\hat{\theta}_0}) \in C_n^c) \to_p 0.
\]

The convergence in probability in the last step is left to the reader. \qed
10. Elementary Decision Theory

The statistician faces a choice between a number of estimators (or actions) given data. A natural question one may ask is what is a good action? To answer this question, we need to introduce the concept of a loss function $L(\theta, \delta(Y))$, which is a function of the truth $\theta$ and the action $\delta(Y)$. It represents the loss or cost due to making a specific decision. It is essentially the negative of utility. Once the loss function is chosen, the statistician’s job is to take the action to minimize the loss.


**Definition 10.1.** (1) A loss function is a map $L : \Theta \times \Theta \to \mathbb{R}_{\geq 0}$.
(2) The risk of an estimator (decision rule) $\delta$ is $R(\theta, \delta) \equiv E_{\theta}L(\theta, \delta(Y))$.
(3) Decision rule $\delta_1$ dominates $\delta_2$ with respect to $L$ iff for all $\theta \in \Theta$,
$$R(\theta, \delta_1) \leq R(\theta, \delta_2).$$
(4) Let $D$ be a collection of decision rules. Then $\delta$ is admissible in $D$ (with respect to $L$) iff there is no rule $\delta'$ that dominates $\delta$. $\delta$ is inadmissible in $D$ (with respect to $L$) iff there is a rule $\delta'$ that dominates $\delta$.

**Remark 10.2.** While admissibility is an important criterion, the following points should be kept in mind:
(a) Admissibility is typically a property of a decision rule with respect to a “class of decision rules,” and not a global property.
(b) Admissibility is a weak criterion of “goodness” of a decision rule. In fact, there are examples where a ridiculous decision rule can be admissible. Consider a single observation $Y \sim \text{Bin}(k, \theta)$ with $k$ known and consider the problem of estimating $\theta$ under a squared error loss from the single observation. Consider the decision rule $\delta(Y) = 1/2$ for $Y = 0, 1, \ldots, k$, i.e., irrespective of the outcome, $\theta$ is estimated as $1/2$. Then $\delta(Y)$ is an admissible estimator of $\theta$ (show this!).

To choose a rule $\delta$ from a class $D$, one can adopt the “play safe” strategy, and choose the rule that minimizes the maximum risk. Such a rule is called a minimax rule.

**Definition 10.3 (Minimax Rule).** The decision rule $\delta_M$ is minimax within a class of rules $D$ if
$$\delta_M = \arg \min_{\delta \in D} \max_{\theta \in \Theta} R(\theta, \delta).$$

Finding a minimax rule in a class of decision rules $D$ involves the following two steps:
(a) Fix a rule $\tilde{\delta} \in D$. Fine $\theta^*(\tilde{\delta})$ that maximizes $R(\theta, \tilde{\delta})$ with respect to $\theta$.
(b) Search over all $\delta \in D$ and find $\delta_M$ that maximizes $R(\theta^*(\delta), \delta)$. 

Example 10.4 (Finding a minimax rule). Consider a discrete parameter space \( \Theta = \{1, 2, 3\} \), a class of decision rules

\[
\mathcal{D} = \{\delta_1, \delta_2, \delta_3, \delta_4\},
\]

and the values of the risk function \( R(\theta, \delta) \) for each \( \theta \in \Theta \) and \( \delta \in \mathcal{D} \) shown in Table 1.

<table>
<thead>
<tr>
<th>Decision Rule</th>
<th>Parameter value</th>
<th>( \theta_1 )</th>
<th>( \theta_2 )</th>
<th>( \theta_3 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \delta_1 )</td>
<td>12</td>
<td>7</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>( \delta_2 )</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>( \delta_3 )</td>
<td>2</td>
<td>5</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>( \delta_4 )</td>
<td>12</td>
<td>7</td>
<td>2</td>
<td></td>
</tr>
</tbody>
</table>

- Which decision rules are admissible?
- Which is decision rule is minimax?

10.2. Bayes’ risk. Minimax strategy is a conservative strategy (protection against the worst possible scenario). To see this, consider a slightly different variant of Table 1 obtained by adding 1000 to each entry in the column under \( \theta_3 \), as shown in Table 2.

<table>
<thead>
<tr>
<th>Decision Rule</th>
<th>Parameter value</th>
<th>( \theta_1 )</th>
<th>( \theta_2 )</th>
<th>( \theta_3 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \delta_1 )</td>
<td>12</td>
<td>7</td>
<td>1001</td>
<td></td>
</tr>
<tr>
<td>( \delta_2 )</td>
<td>4</td>
<td>3</td>
<td>1002</td>
<td></td>
</tr>
<tr>
<td>( \delta_3 )</td>
<td>2</td>
<td>5</td>
<td>1002</td>
<td></td>
</tr>
<tr>
<td>( \delta_4 )</td>
<td>12</td>
<td>7</td>
<td>1002</td>
<td></td>
</tr>
</tbody>
</table>

Then \( \delta_1 \) in minimax in \( \{\delta_1, \ldots, \delta_4\} \). This is a strange choice, because clearly \( \delta_2 \) and \( \delta_3 \) are much better for \( \theta = 1 \) and \( \theta = 2 \), and \( \delta_1 \) is only marginally better for \( \theta = 3 \). A better decision can be reached if we know a priori that \( \theta \) has the following distribution:

\[
\theta = \begin{cases} 
1 & \text{with probability 0.4} \\
2 & \text{with probability 0.5} \\
3 & \text{with probability 0.1}
\end{cases}
\]

In this case, we can calculate the expected risk with respect to this prior distribution as shown in Table 3, and choose the rule \( \delta_5 \) as the decision rule with the smallest expected (Bayes’) risk.

Thus, an easy way to translate the frequentist idea of risk to a Bayesian framework is to simply incorporate a prior, and to consider the expected...
risk taken with respect to the prior distribution. This quantity is called the Bayes risk.

**Definition 10.5 (Bayes’ risk).** (1) Given a prior $\pi$ over the parameter space $\Theta$, the Bayes risk is

$$B(\pi, \delta) = E_\pi \{ R(\theta, \delta) \} = E_\pi \{ E_{\theta}(\theta, \delta(Y)) \}.$$  

(2) For a given prior $\pi$, the Bayes’ rule in a class $D$, is the decision rule $\delta^\pi \in D$ that minimizes the Bayes’ risk, i.e.,

$$\delta^\pi = \arg \min_{\delta \in D} B(\pi, \delta).$$

**Remark 10.6.** For a given prior $\pi(\theta)$, the Bayes’ rule in a class $D$, is the decision rule $\delta^\pi \in D$ that minimizes the expectation of the loss function with respect to the posterior distribution of $\theta$ given $Y$, i.e.,

$$\delta^\pi = \arg \min_{\delta \in D} r(Y, \delta(Y)),$$

where

$$r(Y, \delta(Y)) = \int L(\theta, \delta(Y)) \pi(\theta) d\theta.$$  

For example, under the squared error loss, the objective function

$$r(Y, \delta(Y)) = \int (\theta - \delta(Y))^2 \pi(\theta) d\theta,$$

which is minimized by

$$\delta^\pi = E(\theta | Y),$$

the posterior expectation.

10.3. **Bayes’ rule and admissibility.** Lehmann and Casella (1998) give a comprehensive discussion on this topic. Here we will discuss some of the main ideas and results on the connection between Bayes’ rules and admissibility. We shall see that under a wide array of conditions, Bayes’ estimators are admissible. Interestingly, a converse of this result is also true, and such a result is popularly known as the complete class theorem. First, we state a result on admissibility of unique Bayes’ rules.

**Theorem 10.7.** Any unique Bayes’ rule is admissible.
Proof. Let $\delta^\pi$ be Bayes for prior $\pi$. Consider a rule $\delta'$ for which

$$R(\theta, \delta') \leq R(\theta, \delta^\pi) \quad \forall \theta \in \Theta.$$ 

Then, the Bayes' risk associated with $\delta'$ and prior $\pi$ satisfies

$$B(\pi, \delta') = \int R(\theta, \delta') \pi(\theta) d\theta < \int R(\theta, \delta^\pi) \pi(\theta) d\theta = B(\pi, \delta^\pi).$$

Thus, $\delta'$ is also a Bayes rule. But by the condition of uniqueness of the Bayes' rule, we must have $\delta^\pi = \delta'$ almost surely. □

**Theorem 10.8.** Let $\Theta$ be an open subset of $\mathbb{R}$ and suppose

(i) The prior distribution $\pi(\theta)$ has support $\Theta$.
(ii) the risk function $R(\theta, \delta)$ is continuous in $\theta$ for all $\delta \in D$.

If $\delta^\pi$ is Bayes’ rule for $\pi$ with finite Bayes’ risk, then $\delta^\pi$ is admissible.

Proof. Assume that $\delta^\pi$ is inadmissible and let $\delta'$ dominate $\delta^\pi$. Then,$$
R(\theta, \delta') \leq R(\theta, \delta^\pi) \quad \forall \theta \in \Theta,
R(\theta, \delta') < R(\theta, \delta^\pi) \text{ for some } \theta_0 \in \Theta.
$$

By continuity, there exist some open set $U$ containing $\theta_0$ such that

$$\inf_{\theta \in U} (R(\theta, \delta^\pi) - R(\theta, \delta')) = \varepsilon > 0.$$ 

Then,

$$B(\pi, \delta^\pi) - B(\pi, \delta') \geq \int_U \{R(\theta, \delta^\pi) - R(\theta, \delta')\} \pi(\theta) d\theta, \quad \text{since } R(\theta, \delta^\pi) \geq R(\theta, \delta') \quad \forall \theta \in U^c,$$

$$\geq \varepsilon \int_U \pi(\theta) d\theta > 0.$$ 

This contradicts that $\delta^\pi$ is Bayes’ rule with respect to prior $\pi$. □

**Theorem 10.9.** Consider a discrete parameter space $\Theta = \{\theta_1, \ldots, \theta_K\}$, and suppose $\delta^\pi$ is Bayes for $\pi$, which is discrete with $\pi(\theta = \theta_j) = \pi_j$ for $j = 1, \ldots, K$, such that $\pi_j > 0$ for all $j \in \{1, \ldots, K\}$. Then $\delta^\pi$ is admissible.

Proof. Assume that $\delta^\pi$ is inadmissible and let there exist a $\delta'$ that dominates $\delta^\pi$. Then, $$R(\theta, \delta') \leq R(\theta, \delta^\pi) \quad \forall \theta \in \{1, \ldots, K\},$$

with strict inequality for at least one $\theta_0 \in \{\theta_1, \ldots, \theta_K\}$. Then

$$B(\pi, \delta') = \sum_{j=1}^K \pi_j R(\theta_j, \delta') < \sum_{j=1}^K \pi_j R(\theta_j, \delta^\pi) = B(\pi, \delta^\pi).$$

The inequality above occurs because $\pi_j > 0$ for all $j$. This contradicts the definition of Bayes’ rule. □
10.3.1. *A geometric approach to visualize admissibility and connect it to the Bayes’ rule.* Consider the situation where the parameter space is discrete and consists of only two possible values, i.e., $\Theta = \{\theta_1, \theta_2\}$. The risk of any decision rule $\delta$ is a two-dimensional vector $(R(\theta_1, \delta), R(\theta_2, \delta))$. Consider the class of decision rules $D = \{d_1, \ldots, d_8\}$, and their risk vectors plotted on a 2-D Cartesian plane as shown in Figure 2.

To determine if a given decision rule $\delta$ is admissible, draw a rectangle for the rule as in Figure 2 (green and red rectangles for rules $d_2$ and $d_5$ respectively); if other decision rules are within the area determined by the rectangle (excluding the boundary), $\delta$ is inadmissible. If no other decision rules are within the rectangle area (excluding the boundary), the rule is admissible. Thus, $d_2$ is admissible because nothing “beats” it in two dimensions, but $d_5$ is not admissible because it is dominated by $d_4$.

Another way to characterize admissible rules is to create a convex hull (smallest convex set) containing all the rules. Then all rules lying on the south-west edge are admissible rules. If we there are only finitely many rules that are allowed, then there are finitely many points on the graph, and we can form the convex hull of these points, which will be a polygon. The interpretation of this polygon is that it is the set of possible rules if we allow convex combination of the existing rules (i.e. randomized rules).

**Theorem 10.10.** Suppose that $D$ is a convex set with $|\Theta| < \infty$. If $\delta^* \in D$ is admissible, then $\delta^*$ is Bayes with respect to some prior $\pi$ on $\Theta$. 
Proof. Let $k \equiv |\Theta| < \infty$. Let $\mathcal{S} \equiv \{(R(\theta_1, \delta), \ldots, R(\theta_k, \delta)) : \delta \in \mathcal{D}\}$. Then $\mathcal{S}$ is a convex set in $\mathbb{R}^k$. Further for any $s \in \mathbb{R}^k$, let $Q(s) \equiv \{x \in \mathbb{R}^k : x_i \leq s_i, i = 1, \ldots, k\}$. Now let $s^* = (R(\theta_i, \delta^*))_{i=1}^k$. Then
\[ Q(s^*) \cap \mathcal{S} = \{s^*\}. \]

Let $\tilde{Q}(s^*) = Q(s^*) \setminus \{s^*\}$, the set $Q(s^*)$ without the point $s^*$. Then, $\tilde{Q}(s^*)$ is convex, as is $\mathcal{S}$, and $\tilde{Q}(s^*) \cap \mathcal{S} = \emptyset$. Thus, $\tilde{Q}(s^*)$ and $\mathcal{S}$ are two disjoint convex subsets of $\mathbb{R}^k$.

By the separating hyperplane theorem (= Hahn-Banach theorem in Hilbert spaces), there exists a hyperplane separating $\tilde{Q}(s^*)$ and $\mathcal{S}$, i.e., there exists a vector $w = (w_1, \ldots, w_k)^T$ such that
\[ w^T x \leq w^T s, \quad \forall x \in \tilde{Q}(s^*), \ s \in \mathcal{S}. \]

Note that no co-ordinate of $w$ can be negative. For example, if $w_1 < 0$, we can make $x_1$ arbitrarily negative within $\tilde{Q}(s^*)$, so that $w^T x$ is arbitrarily large and greater than $w'$ for $s \in \mathcal{S}$. Hence $w_j > 0$ for all $j = 1, \ldots, k$.

Define $\pi_j = w_j / \sum_{j=1}^k w_j$ for $j = 1, \ldots, k$. Then, from (10.1), it follows that
\[ \pi^T x \leq \pi^T s, \quad \forall x \in \tilde{Q}(s^*), \ s \in \mathcal{S}, \]
where $\pi = (\pi_1, \ldots, \pi_k)^T$ and $\pi_j > 0$ for all $j = 1, \ldots, k$. Thus, for any $s \in \mathcal{S}$, it follows from (10.2) that
\[ \pi^T s \geq \sup_{x \in \tilde{Q}(s^*)} \pi^T x = \pi^T s^*, \]
since $s^*$ is a limit point of $\tilde{Q}(s^*)$. Consequently,
\[ \pi^T s^* \geq \pi^T s \quad \forall s \in \mathcal{S}, \]
which means
\[ B(\pi, \delta^*) \leq B(\pi, \delta) \quad \forall \delta \in \mathcal{D}. \]

Noting that $B(\pi, \delta)$ is the Bayes’ risk for rule $\delta$ with respect to prior $\pi(\theta_j) = \pi_j$ for $j = 1, \ldots, k$, (10.3) implies that $\delta^*$ is the Bayes’ rule with respect to prior $\pi$. \hfill \Box

10.4. Admissibility and inadmissibility in Gaussian location models. Let us first consider the univariate case.

Theorem 10.11. For i.i.d. data $Y_1, \ldots, Y_n \sim \mathcal{N}(\mu, \sigma^2)$ with $\sigma^2$ known, the sample mean $\bar{Y}$ is an admissible estimator for $\mu$ with respect to squared-error loss.

Proof. We proceed by contradiction using the canonical form of the problem, where we assume $\frac{\sigma^2}{n} = 1$ without loss of generality. Suppose that $\bar{Y}$
is inadmissible. This implies that there exists some estimator \( t(\bar{Y}) \) which dominates \( \bar{Y} \), that is,
\[
E_\mu(t(Y) - \mu)^2 \leq E_\mu(\bar{Y} - \mu)^2 \quad \text{for all } \mu,
\]
with strict inequality holding for some \( \mu \).

Fix some \( \mu \) at which strict inequality holds, and without loss of generality, assume that this \( \mu = 0 \). This means that there exists some \( \varepsilon > 0 \) such that
\[
E_\mu(t(\bar{Y}) - \mu)^2 \leq 1 - \varepsilon I_{\mu \in (-\varepsilon, \varepsilon)}.
\]

Where \( I \) is an indicator function that is 1 when \( \mu \) is in the region where \( t(\bar{Y}) \) beats \( \bar{Y} \).

Now consider Bayes rules for \( \mu \) corresponding to priors of the form \( \mu \sim N(0, \tau^2). \) These estimators take the form \( c\bar{Y} \), where \( c \equiv \frac{\tau^2}{1 + \tau^2} \). Hence
\[
E_\pi \tau^2 E_\mu(c\bar{Y} - \mu)^2 \leq E_\pi \tau^2 E_\mu(t(\bar{Y}) - \mu)^2 \\
\leq E_\pi \tau^2 (1 - \varepsilon I_{\mu \in (-\varepsilon, \varepsilon)})
\]

where \( \pi_{\tau^2} \) is the prior defined with the \( \tau^2 \) corresponding to \( \mu \). By some calculation, we have
\[
\frac{\tau^2}{1 + \tau^2} \leq E_\pi \tau^2 (1 - \varepsilon I_{\mu \in (-\varepsilon, \varepsilon)}) \\
= 1 - \varepsilon P(-\varepsilon < \tau Z < \varepsilon) \\
= 1 - \varepsilon \left[ \Phi \left( \frac{\varepsilon}{\tau} \right) - \Phi \left( \frac{-\varepsilon}{\tau} \right) \right] = 1 - O \left( \frac{1}{\tau} \right).
\]

We now rewrite both sides to get:
\[
1 - \frac{1}{1 + \tau^2} \leq 1 - O \left( \frac{1}{\tau} \right).
\]

This is a contradiction as \( \tau \to \infty \). Thus, there exists no such estimator, and \( \bar{Y} \) is admissible.

When it comes to \( K \geq 3 \) independent Normal populations, however, James and Stein proved in 1961 that the sample averages are no longer admissible under squared error loss.

Consider \( K \) independent normal populations with mean \( \mu_i \) and known variances \( V_i \) for \( i = 1, \ldots, K \). Assume that \( V_i = 1 \) for \( i = 1, \ldots, K \). Let \( Y_1, \ldots, Y_K \) be observations such that \( Y_i \overset{iid}{\sim} N(\mu_i, 1) \) for \( i = 1, \ldots, K \). The problem is to estimate the vector \( \mu = (\mu_1, \ldots, \mu_K)^T \). Clearly the MLE of \( \mu \) is
\[
\hat{\mu}_{\text{MLE}} = Y = (Y_1, \ldots, Y_K)^T.
\]

For \( K = 1 \), \( \hat{\mu}_{\text{MLE}} \) is an admissible estimator of \( \mu \) under the squared error loss. For \( K \geq 3 \), consider the compound square error loss
\[
L(t, \hat{\mu}) = \|\mu - \hat{\mu}\|^2 = \sum_{i=1}^{K} (\mu_i - \hat{\mu}_i)^2.
\]
The risk function for $\hat{\mu}$ under loss 10.4 is

\begin{equation}
R(\mu, \hat{\mu}) = E_\mu \left\{ \sum_{i=1}^{K} (\mu_i - \hat{\mu}_i)^2 \right\}.
\end{equation}

Substituting $\hat{\mu} = \hat{\mu}_{\text{MLE}}$ in (10.5) yields the risk of the MLE as

\begin{equation}
R(\mu, \hat{\mu}) = E_\mu \left\{ \sum_{i=1}^{K} (\mu_i - Y_i)^2 \right\} = K.
\end{equation}

Is there any other estimator that can dominate the MLE? This seems unlikely, because observations are drawn from independent normal distributions. However, James and Stein proved the following striking result in 1961:

**Theorem 10.12.** The estimator

\begin{equation}
\hat{\mu}_{JS} = \left(1 - \hat{B}\right) Y,
\end{equation}

dominates the MLE under loss function (10.4), where

\[ \hat{B} = \frac{K - 2}{S}, \]

and $S = \sum_{i=1}^{K} Y_i^2 = \|Y\|^2$ is the squared norm of $Y$.

To prove the theorem, we need a fundamental lemma, also known as Stein’s lemma, which is interesting in its own right because it provides a characterization of the normal distribution. Below, we state the lemma.

**Lemma 10.13** (Stein’s lemma). If $Y \sim N(\mu, V)$ and $h(\cdot) : \mathbb{R} \rightarrow \mathbb{R}$ is a differentiable function satisfying $E |h'(Y)| < \infty$, then

\[ E \{h(Y)(Y - \mu)\} = VE \{h'(Y)\}. \]

The proof of the lemma is basically by partial integration so we omit the details.

**Proof of Theorem.** The risk of the J-S estimator with respect to loss (10.4) is :

\begin{align*}
R(\mu, \hat{\mu}_{JS}) &= E_\mu \left\{ \|\hat{\mu}_{JS} - \mu\|^2 \right\} \\
&= E_\mu \left\{ \|(1 - \hat{B})Y - \mu\|^2 \right\} \\
&= E_\mu \left\{ \sum_{i=1}^{K} (Y_i - \mu_i)^2 \right\} + E_\mu \left\{ \hat{B}^2 \sum_{i=1}^{K} Y_i^2 \right\} - 2E_\mu \left\{ \hat{B} \sum_{i=1}^{K} Y_i(Y_i - \mu_i) \right\}, \\
&= (I) + (II) + (III).
\end{align*}

Note that $(I) = K$, and

\begin{align*}
(II) &= E_\mu \left\{ \hat{B}^2 S \right\} = E_\mu \left\{ \left(\frac{K - 2}{S^2}\right) \times S \right\} = (K - 2)E_\mu \left( \frac{1}{S} \right).
\end{align*}
For (III), note that

\[(III) = -2E_\mu \left\{ \frac{K - 2}{S} \sum_{i=1}^{K} Y_i (Y_i - \mu_i) \right\} \]

\[= -2(K - 2) \sum_{i=1}^{K} E_\mu \left\{ \frac{Y_i}{S} (Y_i - \mu_i) \right\} \]

\[= -2(K - 2) \sum_{i=1}^{K} E_\mu \left\{ \frac{\partial}{\partial Y_i} \left( \frac{Y_i}{S} \right) \right\} \quad \text{(use Stein’s lemma)} \]

\[= -2(K - 2) \sum_{i=1}^{K} E_\mu \left\{ \frac{S - 2Y_i^2}{S^2} \right\} \]

\[= -2(K - 2)^2 E_\mu \left( \frac{1}{S} \right). \]

This implies that

\[\begin{align*}
R(\mu, \hat{\mu}^{JS}) &= K + (K - 2)^2 E_\mu \left( \frac{1}{S} \right) - 2(K - 2)^2 E_\mu \left( \frac{1}{S} \right) \\
&= K - (K - 2)^2 E_\mu \left( \frac{1}{S} \right) \\
&< K = R(\mu, \hat{\mu}^{MLE}),
\end{align*} \]

as desired. \(\square\)
11. Nonparametric estimation

11.1. Kernel density estimator. Let \( X_1, \ldots, X_n \) be i.i.d. samples from a probability density function \( p \) on \( \mathbb{R} \). The idea of a kernel density estimator is as follows. Let \( F(x) = \int_{-\infty}^{x} p(x) \, dx \) be the CDF of \( p \), and \( F_n(x) \equiv \frac{1}{n} \sum_{i=1}^{n} 1_{X_i \leq x} = \mathbb{P}_n 1 \leq x \) be the empirical CDF. By the strong law of large number, \( F_n(x) \to F(x) \) almost surely for all \( x \in \mathbb{R} \), so it is natural to expect that \( F_n \) is a ‘good estimator’ of \( F \). On the other hand,

\[
p(x) \approx \frac{F(x + h) - F(x - h)}{2h}, \quad \text{for } h \text{ small enough},
\]

so by replacing the CDF \( F \) with the empirical CDF \( F_n \), we have the following Rosenblatt estimator:

\[
\hat{p}_n^R(x) \equiv \frac{F_n(x + h) - F_n(x - h)}{2h} = \frac{1}{2nh} \sum_{i=1}^{n} 1_{x-h < X_i \leq x+h}
\]

\[
= \frac{1}{nh} \sum_{i=1}^{n} K_0 \left( \frac{X_i - x}{h} \right),
\]

where \( K_0(u) = \frac{1}{2} 1_{-1 < u \leq 1} \).

This motivates the following definitions.

Definition 11.1.  
1. \( K : \mathbb{R} \to \mathbb{R} \) is a kernel if \( \int K(u) \, du = 1 \).
2. A kernel density estimator (with kernel \( K \)) is given by

\[
\hat{p}_n(x) \equiv \frac{1}{nh} \sum_{i=1}^{n} K \left( \frac{X_i - x}{h} \right).
\]

Here \( h \) is usually called the ‘bandwidth’ and is omitted in the notation of \( \hat{p}_n \).

11.1.1. Mean squares error at a fixed point. The performance of the kernel density estimator \( \hat{p}_n \) can be evaluated from both local and global perspectives.

First we adopt a local perspective by considering the mean squared error at a fixed point:

\[
\text{MSE} = \text{MSE}(x_0) = E_p(\hat{p}_n(x_0) - p(x_0))^2
\]

\[
= \int \left( \hat{p}_n(x_0; x_1, \ldots, x_n) - p(x_0) \right)^2 p(x_0) \, dx. 
\]

It easily follows that

\[
\text{MSE} = (E_p \hat{p}_n(x_0) - p(x_0))^2 + E_p (\hat{p}_n(x_0) - E_p \hat{p}_n(x_0))^2
\]

\[
= b^2(x_0) + \sigma^2(x_0).
\]

The terms \( b^2(x_0), \sigma^2(x_0) \) are usually called the bias and variance of the kernel density estimator \( \hat{p}_n \).
Proposition 11.2. Let $K$ be a kernel such that $\|K\|_2^2 = \int K^2(u) \, du < \infty$. Then
\[ \sigma^2(x_0) \leq \frac{1}{nh} \|p\|_{\infty} \|K\|_2^2. \]

Proof. Note that
\[ \sigma^2(x_0) = \mathbb{E}_p \left[ \left( \frac{1}{nh} \sum_{i=1}^{n} K \left( \frac{X_i - x_0}{h} \right) - \mathbb{E}_p K \left( \frac{X - x_0}{h} \right) \right)^2 \right] \]
\[ = \frac{1}{nh^2} \text{Var}_p \left[ K \left( \frac{X - x_0}{h} \right) \right] \]
\[ \leq \frac{1}{nh^2} \mathbb{E}_p \left[ K^2 \left( \frac{X - x_0}{h} \right) \right] \]
\[ = \frac{1}{nh^2} \int K^2 \left( \frac{z - x_0}{h} \right) p(z) \, dz \leq \frac{1}{nh} \|p\|_{\infty} \|K\|_2^2, \]
as desired. \[ \square \]

To explicitly calculate the bias, we need to work with some class of densities.

Definition 11.3. Let $T \subset \mathbb{R}$ be an interval and $\beta, L > 0$. The Hölder class $\Sigma(\beta, L)$ on $T$ is defined by
\[ \Sigma(\beta, L) \equiv \{ f : T \to \mathbb{R} \text{ is } \lfloor \beta \rfloor \text{ times differentiable, and} \}
\[ |f^{(\ell)}(x) - f^{(\ell)}(x')| \leq L|x - x'|^{\beta - \ell}, \quad \forall x, x' \in T \}. \]

We also need the notion of kernel of order $\ell$ to calculate the bias.

Definition 11.4. A kernel $K$ is said of order $\ell$, iff $u \mapsto u^j K(u)$ is Lebesgue integrable for $j = 0, 1, \ldots, \ell$, and
\[ \int u^j K(u) \, du = \delta_{0,j}. \]

Here $\delta_{m,n} = 1_{m=n}$ is the Kronecker delta.

The existence of kernels of order $\ell$ can be seen as follows. Take $\{\phi_m\}_{m=0}^{\infty}$ to be the orthonormal basis of the Hilbert space $L_2([-1, 1], dx) = L_2[-1, 1]$ obtained by $\{x \mapsto x^j : j = 0, 1, \ldots\}$. $\{\phi_m\}_{m=0}^{\infty}$'s are known as the Legendre polynomial given by
\[ \phi_0(x) = \frac{1}{\sqrt{2}}, \quad \phi_m(x) = \sqrt{\frac{2m+1}{2}} \cdot \frac{1}{2^m m!} \frac{d^m}{dx^m}[(x^2 - 1)^m]. \]
Then we have the following.

Lemma 11.5.\[ (11.1) \quad K(u) \equiv \sum_{m=0}^{\ell} \phi_m(0) \phi_m(u) 1_{|u| \leq 1} \]
is a kernel of order $\ell$. 

Proof. Since $\phi_q$ is a polynomial of degree $q$, there exists $b_{qj}, j = 0, 1, \ldots, \ell$ such that

$$u^j = \sum_{q=0}^{\ell} b_{qj} \phi_q(u), \quad \forall u \in [-1, 1].$$

Then

$$\int u^j K(u) \, du = \sum_{q=0}^{\ell} \sum_{m=0}^{\ell} \int_{-1}^1 b_{qj} \phi_q(u) \phi_m(0) \phi_m(u) \, du$$

$$= \sum_{q=0}^{\ell} \sum_{m=0}^{\ell} b_{qj} \delta_{q,m} \phi_m(0) = \sum_{q=0}^{\ell} b_{qj} \phi_q(0) = \delta_{0,j},$$
as desired. \hfill \square

Now we can calculate the bias for the kernel density estimator.

**Proposition 11.6.** Let $p \in \Sigma(\beta, L)$ be a density and $K$ be a kernel of order $\ell \equiv \lceil \beta \rceil$ with $\int |u|^{\beta} |K(u)| \, du < \infty$. Then

$$b^2(x_0) \leq \left( \frac{L \int |u|^{\beta} |K(u)| \, du}{\ell!} \right)^2 \cdot h^{2\beta}. $$

Proof. Note that

$$b(x_0) = \frac{1}{h} \int K\left( \frac{z - x_0}{h} \right) p(z) \, dz - p(x_0)$$

$$= \int K(u) p(x_0 + uh) \, du - \int p(x_0) K(u) \, du$$

$$= \int K(u) (p(x_0 + uh) - p(x_0)) \, du.$$ 

On the other hand, by Taylor expansion,

$$p(x_0 + uh) = p(x_0) + p'(x_0) uh + \ldots + \frac{(uh)^\ell}{\ell!} p^{(\ell)}(x_0 + \tau uh),$$

for some $\tau \in [0, 1]$ (which depends on $u, h$). Since the kernel $K$ is assumed to be of order $\ell$, it follows that

$$|b(x_0)| = \left| \int K(u) \frac{(uh)^\ell}{\ell!} p^{(\ell)}(x_0 + \tau uh) \, du \right|$$

$$= \left| \int K(u) \frac{(uh)^\ell}{\ell!} \left[ p^{(\ell)}(x_0 + \tau uh) - p^{(\ell)}(x_0) \right] \, du \right|$$

$$\leq \int |K(u)| \frac{|uh|^\ell}{\ell!} |p^{(\ell)}(x_0 + \tau uh) - p^{(\ell)}(x_0)| \, du$$

$$\leq \int |K(u)| \frac{|uh|^\ell}{\ell!} L |uh|^{\beta - \ell} \, du \leq \frac{L \int |u|^{\beta} |K(u)| \, du}{\ell!} \cdot |h|^{\beta},$$
as desired. \hfill \square
**Theorem 11.7.** Suppose the conditions in Propositions 11.2 and 11.6 hold. Then with \( h = \alpha n^{-\frac{2\beta}{2\beta + 1}} \) for fixed number some \( \alpha > 0 \),

\[
\sup_{x_0 \in \mathbb{R}} \sup_{p \in \Sigma(\beta, L); \text{density}} E_p \left( \hat{p}_n(x_0) - p(x_0) \right)^2 \leq C \cdot n^{-\frac{2\beta}{2\beta + 1}} ,
\]

where the constant \( C > 0 \) depends only on \( \beta, L, \alpha \) and \( K \).

**Proof.** By Propositions 11.2 and 11.6, as long as

\[
\sup_{x \in \mathbb{R}} \sup_{p \in \Sigma(\beta, L)} p(x) < \infty ,
\]

we have that

\[
\text{MSE} \leq \text{const} \cdot \left( L^{2\beta} + \frac{1}{nh} \right) \lesssim n^{-\frac{2\beta}{2\beta + 1}} .
\]

To establish (11.2), for any bounded kernel \( K^* \) of order \( \ell \), the proof of Proposition 11.6 yields that

\[
\left| \int K^*(z - x)p(z) \, dz - p(x) \right| \leq L \ell! \int |u|^\beta K^*(u) \, du.
\]

This implies that

\[
p(x) \leq \frac{1}{\ell!} \int |u|^\beta K^*(u) \, du + \int |K^*(z - x)p(z) \, dz \leq \frac{1}{\ell!} \int |u|^\beta K^*(u) \, du + \|K^*\|_\infty .
\]

The right hand side of the above display is independent of \( p \) and \( x \) so we have proved (11.2). \( \square \)

11.1.2. **Mean integrated squared error.** Let us consider now the performance of the kernel density estimator from a global perspective, in particular we will measure its performance in \( L^2 \) distance as follows: define the mean integrated squared error by

\[
\text{MISE} = E_p \int ( \hat{p}_n(x) - p(x) )^2 \, dx
\]

\[
= \int \text{MSE}(x) \, dx
\]

\[
= \int b^2(x) \, dx + \int \sigma^2(x) \, dx.
\]

**Proposition 11.8.** Let \( K \) be a kernel such that \( \|K\|_2 < \infty \). Then

\[
\int \sigma^2(x) \, dx \leq \frac{1}{nh} \|K\|_2^2 .
\]

Note that we do not require \( \|p\|_\infty < \infty \) as in Proposition 11.2.
Proof. Similarly as in the proof of Proposition 11.2, we have

\[ \sigma^2(x) \leq \frac{1}{nh^2} E_p \left[ K^2 \left( \frac{X - x}{h} \right) \right]. \]  

(11.3)

Hence by Fubini’s theorem

\[
\int \sigma^2(x) \, dx \leq \frac{1}{nh^2} \int \left( \int K^2 \left( \frac{z - x}{h} \right) p(z) \, dz \right) \, dx
\]

\[
= \frac{1}{nh^2} \int p(z) \, dz \int K^2 \left( \frac{z - x}{h} \right) \, dx
\]

\[
= \frac{1}{nh} \int K^2(u) \, du,
\]

as desired. The last equality follows by change of variable in the integration.

\[ \square \]

To calculate the bias term, we also need to specify some function class. In the previous section we use the Hölder class that gives pointwise control of the differentiability of the density. Since now we are working with a global \( L^2 \) loss, we may use the following weaker (and hence more general) Sobolev class.

**Definition 11.9.** The Sobolev class \( W(\beta, L) \) is defined as follows: \( f \in W(\beta, L) \) if and only if \( f \) is \((\beta - 1)\) times differentiable, and \( f^{(\beta - 1)} \) is absolute continuous such that

\[ \int (f^{(\beta)}(x))^2 \, dx \leq L^2. \]

**Proposition 11.10.** Let \( \beta \geq 1 \) be an integer. Suppose the density \( p \in W(\beta, L) \) and \( K \) is a kernel of order \( \beta \) such that \( \int |u|^{\beta} |K(u)| \, du < \infty \). Then

\[ \int b^2(x) \, dx \leq \left( L \int |u|^{\beta} |K(u)| \, du \right)^2 \cdot h^{2\beta}. \]

The conclusion above is the same as in Proposition 11.6, but we need a somewhat different proof because now we are working under \( L_2 \) distance.

We need the following well-known generalized Minkowski inequality.

**Lemma 11.11 (Generalized integral Minkowski inequality).** For any Borel measurable function \( g : \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R} \), it holds that

\[ \int \left( \int g(u, x) \, du \right)^2 \, dx \leq \left[ \int \left( \int g^2(u, x) \, dx \right)^{1/2} \, du \right]^2. \]

**Proof.** Without loss of generality we assume that

\[ C_g \equiv \int \left( \int g^2(u, x) \, dx \right)^{1/2} \, du < \infty, \]
otherwise the claim is trivial. Let \( S(x) \equiv \int g(u, x) \, du \). Then for any \( f \in L_2 \equiv L_2(\mathbb{R}) \),
\[
|\langle S, f \rangle| = \left| \int f(x) \, dx \int g(u, x) \, du \right| \\
\leq \int |f(x)| \, dx \int |g(u, x)| \, du \\
= \int du \int |f(x)||g(u, x)| \, dx \quad \text{(Fubini)} \\
\leq \|f\|_2 \left( \int g^2(u, x) \, dx \right)^{1/2} \quad \text{(Cauchy-Schwarz)} \\
= C_g\|f\|_2.
\]
This shows that the linear operator
\[
S : L_2 \to \mathbb{R} \\
f \mapsto \langle S, f \rangle
\]
is continuous. By Riesz’s representation theorem, \( S \in L_2 \), and the operator norm of \( S \) is the \( L_2 \)-norm of the function \( S \):
\[
\|S\|_2 = \sup_{f \in L_2 : \|f\|_2 = 1} \langle S, f \rangle \leq C_g,
\]
as desired. \( \square \)

Proof of Proposition 11.10. Using the integral form of Taylor expansion, we get
\[
p(x + uh) = p(x) + p'(x)uh + \ldots + \frac{(uh)^\beta}{(\beta - 1)!} \int_0^1 (1 - \tau)^{\beta - 1} p^{(\beta)}(x + \tau uh) \, d\tau.
\]
Since the kernel \( K \) is of order \( \beta \), we have
\[
b(x) = \int K(u) \frac{(uh)^\beta}{(\beta - 1)!} \left( \int_0^1 (1 - \tau)^{\beta - 1} p^{(\beta)}(x + \tau uh) \, d\tau \right) \, du \\
\equiv \int g(u, x) \, du.
\]
This implies, by Minkowski inequality used twice,
\[
\int b^2(x) \, dx \\
= \int \left( \int g(u, x) \, du \right)^2 \, dx \\
\leq \left[ \int \left( \int g^2(u, x) \, dx \right)^{1/2} \, du \right]^2 \\
= \left[ \int |K(u)| \frac{|uh|^\beta}{(\beta - 1)!} \, du \left( \int \left( \int_0^1 (1 - \tau)^{\beta - 1} p^{(\beta)}(x + \tau uh) \right)^2 \, d\tau \right)^{1/2} \, dx \right]^2
\]
\[ \int |K(u)| \frac{|uh|^{\beta}}{(\beta - 1)!} du \left( \int \left( \int_0^1 h_u(\tau, x) d\tau \right)^{1/2} dx \right)^2 \]

\[ \leq \int |K(u)| \frac{|uh|^{\beta}}{(\beta - 1)!} du \left( \int \left( \int_0^1 h_u^2(\tau, x) d\tau \right)^{1/2} dx \right)^2 \]

\[ = \int |K(u)| \frac{|uh|^{\beta}}{(\beta - 1)!} du \left[ \int |1 - \tau|^{\beta - 1} \left( \int_0^1 (p^{(\beta)}(x + \tau uh))^2 dx \right)^{1/2} d\tau \right]^2 \]

\[ \leq \left[ \int |K(u)| \frac{|uh|^{\beta}}{(\beta - 1)!} du \int |1 - \tau|^{\beta - 1} L d\tau \right]^2 \]

\[ = \left( \frac{L \int |u|^{\beta} |K(u)| du}{\beta!} \right)^2 \cdot h^{2\beta}, \]

as desired. \( \square \)

**Theorem 11.12.** Suppose the conditions in Propositions 11.8 and 11.10 hold. Then with \( h = an^{-\frac{1}{2\beta + 1}} \) for fixed number some \( \alpha > 0 \),

\[ \sup_{p \in W(\beta, L)} E_p \int (\hat{p}_n(x) - p(x))^2 dx \leq C \cdot n^{-\frac{2\beta}{2\beta + 1}}, \]

where the constant \( C > 0 \) depends only on \( \beta, L, \alpha \) and \( K \).

**Proof.** Combining Propositions 11.8 and 11.10. \( \square \)

11.1.3. *Bandwidth selection: cross validation.* One major problem for the kernel density estimator is that there is no apriori optimal way to choose the bandwidth \( h \), since the optimal value of \( h^* = \arg \min_{h > 0} \text{MISE}(h) \)

in terms of mean integrated squared error would involve the unknown density \( p \). Cross-validation is a technique to choose the bandwidth \( h \) by estimating the mean integrated squared error. First note that

\[ \text{MISE}(h) = E_p \int (\hat{p}_n(x) - p(x))^2 dx = E_p \left[ \int \hat{p}_n^2 - 2 \int \hat{p}_n p \right] + \int p^2. \]

This shows that

\[ h^* = \arg \min_{h > 0} J(h) \equiv \arg \min_{h > 0} E_p \left[ \int \hat{p}_n^2 - 2 \int \hat{p}_n p \right]. \]

The idea now is to find an unbiased estimator for \( J(h) \). The leave-one-out cross validation technique proposes the following estimator:

\[ \text{CV}(h) \equiv \hat{J}(h) \equiv \int \hat{p}_n^2 - \frac{2}{n} \sum_{i=1}^n \hat{p}_{n,-i}(X_i), \]

where

\[ \hat{p}_{n,-i}(x) \equiv \frac{1}{(n-1)h} \sum_{j \neq i} K \left( \frac{X_j - x}{h} \right). \]
**Proposition 11.13.** Suppose the density $p$ and the kernel $K$ are such that $\int p^2(x) \, dx < \infty$, and for any $h > 0$,

$$\int p(x) \left| K \left( \frac{x - z}{h} \right) \right| p(z) \, dx \, dz < \infty.$$  

Then

$$E_p[CV(h)] = J(h) = MISE(h) - \int p^2.$$

**Proof.** We only need to show that

$$E_p \left[ \sum_{i=1}^{n} \hat{p}_{n,i}(X_i) \right] = E_p \int \hat{p}_n p.$$

To see this, note that the left hand side

$$E_p \left[ \sum_{i=1}^{n} \hat{p}_{n,i}(X_i) \right] = E X_1, \ldots, X_n \sim p \hat{p}_{n,-1}(X_1)$$

where the last equality follows by Fubini’s theorem and the assumption. On the other hand, the right hand side

$$E_p \int \hat{p}_n p = E_p \left[ \sum_{i=1}^{n} \int K \left( \frac{x - z}{h} \right) p(z) \, dz \right] = 1/h \int p(x) K \left( \frac{x - z}{h} \right) p(z) \, dx \, dz,$$

proving the claim. \(\Box\)

11.2. **Regression: projection estimators.** In this section, we consider the following regression model

$$Y_i = f(X_i) + \xi_i, \quad i = 1, \ldots, n.$$  

We assume that $f \in L_2([0,1]) \equiv L_2([0,1], dx)$, $X_i = i/n$, and the errors $\{\xi_i\}$’s are i.i.d. with $E\xi_1 = 0$ and $E\xi_1^2 = \sigma^2$.

Let $\{\phi_j\}_{j=1}^\infty$ be an orthonormal basis of $L_2([0,1])$, then the Fourier transform gives a canonical isometry between $L_2([0,1])$ and $\ell_2(\mathbb{N})$ in the following sense: the map

$$L_2([0,1]) \to \ell_2(\mathbb{N})$$

$$f \mapsto \left( \theta_j \equiv \langle f, \phi_j \rangle = \int f \phi_j \right)_{j=1}^\infty$$
is an isometry due to Parsevel’s identity: \( \|f\|_2 = \|\theta(f)\|_2 \), where \( \theta(f) = (\theta_j)_{j=1}^{\infty} \). We may also write

\[
f = \sum_{j=1}^{\infty} \theta_j \phi_j.
\]

Note here the convergence is in \( L_2 \). To avoid unnecessary technicality, we assume that the regression function \( f \) satisfies \( \|\theta(f)\|_1 < \infty \), i.e.

\[
\sum_{i=1}^{n} |\theta_i(f)| < \infty.
\]

This means that the expansion \( f = \sum_{j=1}^{\infty} \theta_j \phi_j \) is valid both in \( L_2 \) and pointwise.

The idea of projection estimator is simple: we want to choose some suitable truncation level \( N \) such that

\[
f \approx \sum_{j=1}^{N} \theta_j \phi_j,
\]

and we may estimate the unknown coefficient \( \theta_j \) by

\[
\hat{\theta}_j = \frac{1}{n} \sum_{i=1}^{n} Y_i \phi_j(X_i).
\]

From the above heuristics, we may formally define the projection estimator by

\[
(11.5) \quad \hat{f}_{nN}(x) = \sum_{j=1}^{N} \hat{\theta}_j \phi_j(x),
\]

where

\[
(11.6) \quad \hat{\theta}_j = \frac{1}{n} \sum_{i=1}^{n} Y_i \phi_j(X_i).
\]

In what follows, we will primarily be concerned with trigonometric basis given by

\[
\phi_1(x) = 1, \quad \phi_{2k}(x) = \sqrt{2} \cos(2\pi k x), \quad \phi_{2k+1}(x) = \sqrt{2} \sin(2\pi k x), \quad k = 1, 2, \ldots.
\]

The function class we work with in this section is largely similar to the one we have seen in the previous section, but with a slight modification due to the boundary effect. We define the periodic Sobolev class \( W^*(\beta, L) \) as follows: \( f \in W^*(\beta, L) \) if and only if \( f \in \mathcal{W}(\beta, L) \) and \( f^{(j)}(0) = f^{(j)}(1) \) for \( j = 1, \ldots, \beta - 1 \).

The isometry between \( L_2([0,1]) \) and \( \ell_2(\mathbb{N}) \) can also be implemented for the periodic Sobolev class. The following result is well-known.
Proposition 11.14. $f = \sum_{j=1}^{\infty} \theta_j \phi_j \in W^*(\beta, L)$ if and only if $\theta \in \Theta(\beta, L^2/\pi^{2\beta})$, where
\[
\Theta(\beta, Q) \equiv \left\{ \theta \in \ell_2(\mathbb{N}) : \sum_{j=1}^{\infty} a_j^2 \theta_j^2 \leq Q \right\},
\]
with
\[
a_j = \begin{cases} j^\beta, & j \text{ even}, \\ (j - 1)^\beta, & j \text{ odd}. \end{cases}
\]

For notational convenience, we often write $Q(L) = L^2/\pi^{2\beta}$, then $f \in W^*(\beta, L)$ iff $\theta(f) \in \Theta(\beta, Q(L))$.

Now let us consider the mean integrated square error of the projection estimator defined by
\[
\text{MISE} \equiv E_f \| \hat{f}_{nN} - f \|^2 = E_f \int_0^1 (\hat{f}_{nN}(x) - f(x))^2 \, dx.
\]
The following lemma will be useful in the calculation of bias and variance of the projection estimator $\hat{f}_{nN}$.

Lemma 11.15. With $\{\phi_j\}_{j=1}^{\infty}$ being the trigonometric basis,
\[
\frac{1}{n} \sum_{s=1}^{n} \phi_j \left( \frac{s}{n} \right) \phi_k \left( \frac{s}{n} \right) = \delta_{j,k}
\]
for $1 \leq j, k \leq n - 1$.

Proof. We only consider the case
\[
\phi_j(x) = \sqrt{2} \cos(2\pi mx), \quad j = 2m,
\]
\[
\phi_k(x) = \sqrt{2} \sin(2\pi lx), \quad k = 2l + 1.
\]
Define
\[
a = \exp \left( \sqrt{-1} \cdot \frac{2\pi m}{n} \right), \quad b = \exp \left( \sqrt{-1} \cdot \frac{2\pi \ell}{n} \right).
\]
Then the quantity of interest
\[
\frac{1}{n} \sum_{s=1}^{n} \phi_j \left( \frac{s}{n} \right) \phi_k \left( \frac{s}{n} \right) = \frac{1}{n} \sum_{s=1}^{n} \sqrt{2} \frac{a^s + a^{-s}}{2} \cdot \frac{\sqrt{2} (b^s - b^{-s})}{2\sqrt{-1}}
\]
\[
= \frac{1}{2\sqrt{-1}n} \sum_{s=1}^{n} \left( (ab)^s - (a/b)^s + (b/s)^s - (ab)^{-s} \right).
\]
Note that
\[
\sum_{s=1}^{n} (ab)^s = ab \cdot \frac{(ab)^n - 1}{ab - 1} = 0,
\]
\[
\sum_{s=1}^{n} (a/b)^s = a/b \cdot \frac{(a/b)^n - 1}{a/b - 1} = 0,
\]
\[
\sum_{s=1}^{n} (b/s)^s = b/s \cdot \frac{(b/s)^n - 1}{b/s - 1} = 0,
\]
\[
\sum_{s=1}^{n} (ab)^{-s} = ab^{-1} \cdot \frac{(ab)^{-n} - 1}{ab^{-1} - 1} = 0.
\]
where the last equality follows since \((ab)^n = 1\). It is also easy to verify that
\[
\sum_{s=1}^{n} (a/b)^s = \sum_{s=1}^{n} (b/a)^s = \sum_{s=1}^{n} (ab)^{-s} = 0,
\]
proving the claim. □

**Proposition 11.16.** Let \(\hat{\theta}_j\) be defined in (11.6). Then
\[
E_f \hat{\theta}_j = \theta_j + a_j,
\]
\[
E_f(\hat{\theta}_j - \theta_j)^2 = \frac{\sigma^2}{n} + \alpha_j^2, \quad 1 \leq j \leq n - 1.
\]
Hence,
\[
(11.7) \quad \text{MISE} = E_f \| \hat{f}_n - f \|^2 = \frac{\sigma^2 N}{n} + \sum_{j=N+1}^{\infty} \theta_j^2 + \sum_{j=1}^{N} \alpha_j^2.
\]
Here
\[
\alpha_j = \frac{1}{n} \sum_{i=1}^{n} f\left(\frac{i}{n}\right) \phi_j\left(\frac{i}{n}\right) - \int_{0}^{1} f(x) \phi_j(x) \, dx.
\]
The three terms on the right hand side of the mean integrated square error (11.7) can be understood as follows: the first term \(\sigma^2 N/n\) represents the variance term (estimation error) on the low frequency coordinates (first \(N\) coordinates that are used in the definition of the projection estimator); the second term \(\sum_{j=N+1}^{\infty} \theta_j^2\) is the bias term caused due to the truncation level \(N\); the third term \(\sum_{j=1}^{N} \alpha_j^2\) is the (extra) bias term due to the design points.

**Proof.** Note that
\[
\hat{\theta}_j = \frac{1}{n} \sum_{i=1}^{n} Y_i \phi_j\left(\frac{i}{n}\right)
\]
\[
= \frac{1}{n} \left( \sum_{i=1}^{n} f\left(\frac{i}{n}\right) \phi_j\left(\frac{i}{n}\right) + \sum_{i=1}^{n} \xi_i \phi_j\left(\frac{i}{n}\right) \right).
\]
This means
\[
E_f \hat{\theta}_j = \frac{1}{n} \sum_{i=1}^{n} f\left(\frac{i}{n}\right) \phi_j\left(\frac{i}{n}\right) = \theta_j + \alpha_j,
\]
and
\[
E_f(\hat{\theta}_j - \theta_j)^2 = E_f(\hat{\theta}_j - E_f \hat{\theta}_j)^2 + (E_f \hat{\theta}_j - \theta_j)^2
\]
\[
= E_f \left( \frac{1}{n} \sum_{i=1}^{n} \xi_i \phi_j\left(\frac{i}{n}\right) \right)^2 + \alpha_j^2
\]
\[
= \frac{1}{n^2} \sum_{i_1, i_2} E \xi_{i_1} \xi_{i_2} \phi_j\left(\frac{i_1}{n}\right) \phi_j\left(\frac{i_2}{n}\right) + \alpha_j^2.
\]
\[
\begin{align*}
\frac{1}{n^2} & \sum_{s=1}^{n} \sigma^2 \phi_j \left( \frac{s}{n} \right) \phi_j \left( \frac{s}{n} \right) + \alpha_j^2 \\
& = \frac{\sigma^2}{n} + \alpha_j^2,
\end{align*}
\]

where in the middle we used Lemma 11.15. Now consider the MISE:

\[
E_f \| \hat{f}_{nN} - f \|_2^2 = E_f \int_0^1 (\hat{f}_{nN}(x) - f(x))^2 \, dx
\]

\[
= E_f \int_0^1 \left( \sum_{j=1}^{N} (\hat{\theta}_j - \theta_j) \phi_j(x) - \sum_{j=N+1}^{\infty} \theta_j \phi_j(x) \right)^2 \, dx
\]

\[
= \sum_{j=1}^{N} E_f (\hat{\theta}_j - \theta_j)^2 + \sum_{j=N+1}^{\infty} \theta_j^2
\]

\[
= \frac{\sigma^2 N}{n} + \sum_{j=1}^{N} \alpha_j^2 + \sum_{j=N+1}^{\infty} \theta_j^2.
\]

The proof is complete. \(\square\)

We are now very close to get risk bounds for the projection estimator, except for the fact that we still need to handle the extra bias term (the third term) in (11.7).

**Lemma 11.17.** For any \(\theta \in \Theta(\beta, Q)\),

\[
\max_{1 \leq j \leq n-1} |\alpha_j| \leq C_{\beta, Q} \cdot n^{-\beta + 1/2},
\]

where the constant \(C_{\beta, Q}\) depends only on \(\beta, Q\).

*Proof.* Note that for any \(1 \leq j \leq n-1\),

\[
\alpha_j = \frac{1}{n} \sum_{i=1}^{n} f \left( \frac{i}{n} \right) \phi_j \left( \frac{i}{n} \right) - \theta_j
\]

\[
= \frac{1}{n} \sum_{i=1}^{n} \left( \sum_{m=1}^{\infty} \theta_m \phi_m \left( \frac{i}{n} \right) \right) \phi_j \left( \frac{i}{n} \right) - \theta_j
\]

\[
= \sum_{m=1}^{n-1} \theta_m \cdot \frac{1}{n} \sum_{i=1}^{n} \phi_m \left( \frac{i}{n} \right) \phi_j \left( \frac{i}{n} \right) - \theta_j
\]

\[
+ \frac{1}{n} \sum_{i=1}^{n} \sum_{m=n}^{\infty} \theta_m \phi_m \left( \frac{i}{n} \right) \phi_j \left( \frac{i}{n} \right)
\]

\[
= \frac{1}{n} \sum_{i=1}^{n} \sum_{m=n}^{\infty} \theta_m \phi_m \left( \frac{i}{n} \right) \phi_j \left( \frac{i}{n} \right).
\]
where the last inequality follows from Lemma 11.15. This shows that
\[
\max_{1 \leq j \leq n-1} |\alpha_j| \leq \max_{1 \leq j \leq n-1} \left| \sum_{m=n}^{\infty} \theta_m \frac{1}{n} \sum_{i=1}^{n} \phi_m \left( \frac{i}{n} \right) \phi_j \left( \frac{i}{n} \right) \right| \leq 2 \sum_{m=n}^{\infty} |\theta_m|.
\]

For any \( \theta \in \Theta(\beta, Q) \),
\[
\sum_{m=n}^{\infty} |\theta_m| = \sum_{m=1}^{\infty} a_m |\theta_m| \cdot \left( \frac{1}{a_m} |1_{m \geq n}| \right) \\
\leq \left( \sum_{m=1}^{\infty} a_m^2 \theta_m^2 \right)^{1/2} \left( \sum_{m=n}^{\infty} a_m^{-2} \right)^{1/2} \\
\leq Q^{1/2} \left[ \sum_{m=n}^{\infty} \frac{1}{(m-1)^{2\beta}} \right]^{1/2} \leq C_{\beta,Q} n^{-\beta+1/2}.
\]
The proof is complete. \( \square \)

**Theorem 11.18.** Suppose \( \beta \geq 1 \) and \( L > 0 \). Then with \( N = \lceil \alpha n^{-1/2\beta} \rceil \) for some fixed number \( \alpha > 0 \),
\[
\sup_{f \in W^*(\beta,L)} E_f \| \hat{f}_n - f \|_2^2 \leq C n^{-\frac{2\beta}{\beta+1}}.
\]

**Proof.** By (11.7) and the lemma above, we have that
\[
E_f \| \hat{f}_n - f \|_2^2 \leq \frac{\sigma^2 N}{n} + \sum_{j=N+1}^{\infty} \theta_j^2 + C_{\beta,L} n^{-2\beta+1} \\
\leq C_1 \frac{N}{n} + \sum_{j=N+1}^{\infty} \theta_j^2,
\]
where the last inequality follows from the assumption that \( \beta \geq 1 \). The claim follows by noting that
\[
\sum_{j=N+1}^{\infty} \theta_j^2 \leq \frac{1}{a_{N+1}^2} \sum_{j=1}^{\infty} a_j^2 \theta_j^2 \leq \frac{Q}{a_{N+1}^2} \leq \frac{Q}{N^{2\beta}}.
\]
The proof is complete. \( \square \)

### 11.3. Gaussian nonparametric models and their asymptotic equivalence

In this section we introduce several models that are closely tied to Gaussian structures. The main message is that, despite their apparent different appearance, their statistical properties measured in a decision-theoretic framework can be rather similar, due to the deep asymptotic equivalence theory.

**Model 1: Nonparametric Gaussian regression**
The non-parametric Gaussian regression model takes the following form: suppose we observe \( Y_1, \ldots, Y_n \) according to
\[
Y_i = f(X_i) + \xi_i, \quad i = 1, \ldots, n.
\]
Here \( f : [0, 1] \to \mathbb{R}, X_i = i/n \) and \( \xi_i \) are i.i.d \( N(0, \sigma^2) \).

Model 2: Gaussian white noise

The Gaussian white noise model takes the following form: suppose we observe \( \{Y(t) : t \in [0, 1]\} \) according to the stochastic differential equation:
\[
dY(t) = f(t) \, dt + \frac{\sigma}{\sqrt{n}} \, dW(t),
\]
where \( W \) is the standard Wiener process on \([0, 1]\). Observing \( \{Y(t) : t \in [0, 1]\} \) is equivalent to observing the whole trajectory along any \( g \in L^2([0, 1]) \):
\[
\left\{ g \mapsto \int_0^1 g(t) \, dY(t) \equiv Y_f(g) \sim N(\langle f, g \rangle, \|g\|_2^2/n) : g \in L^2([0, 1]) \right\}.
\]

Model 3: Gaussian sequence model

We may take \( g \) to be the orthonormal basis \( \{\phi_k\}_{k=1}^\infty \) in \( L^2([0, 1]) \) to get
\[
Y_k = \langle f, \phi_k \rangle + \frac{\sigma}{\sqrt{n}} g_k,
\]
where \( g_k \)'s are i.i.d \( N(0, 1) \).

**Definition 11.19.**

1. A statistical experiment \( \mathcal{E} \) is a tuple \( \mathcal{E} = (\mathcal{Y}, \mathcal{B}, \{P_f\}_{f \in \mathcal{F}}) \), where \( \mathcal{Y} \) is the sample space, and \( P_f \) the associated laws (=data generating mechanisms) indexed by \( \mathcal{F} \).
2. Given a decision rule \( \delta \in \mathcal{D} \), and a loss function \( L : \mathcal{F} \times \mathcal{D} \to \mathbb{R}_{\geq 0} \), the risk is defined by
\[
R_L(f, \delta) \equiv P_f L(f, \delta(Y)).
\]
3. The Le Cam distance between two experiments \( \mathcal{E}^{(i)} = (\mathcal{Y}^{(i)}, \mathcal{B}^{(i)}, \{P_f^{(i)}\}_{f \in \mathcal{F}})(i = 1, 2) \) is defined by
\[
\Delta_F(\mathcal{E}^{(1)}, \mathcal{E}^{(2)}) = \left( \sup_{\delta(2)} \inf_{\delta(1)} + \sup_{\delta(1)} \inf_{\delta(2)} \right) \sup_{\|L\|_\infty \leq 1} |R_L^{(1)}(f, \delta^{(1)}) - R_L^{(2)}(f, \delta^{(2)})|.
\]

If the Le Cam distance between two statistical experiments are small, for instance, \( \Delta_F(\mathcal{E}^{(1)}, \mathcal{E}^{(2)}) \leq \varepsilon \), then we know by definition that for any decision rule \( \delta^{(1)} \) in experiment \( \mathcal{E}^{(1)} \), there is a decision rule \( \delta^{(2)} \) in experiment \( \mathcal{E}^{(2)} \) such that the difference of the risk between the two estimators in two experiments under any bounded loss function is at most \( \varepsilon \). So in this sense, we may think of two experiments are close to each other if the Le Cam distance between them are small.
In what follows, we refer to $E^{(1)}$ as the Gaussian nonparametric regression model, $E^{(2)}$ as the Gaussian white noise model, and $E^{(3)}$ as the Gaussian sequence model. Brown and Low (1996) proved the following result.

**Theorem 11.20.** Suppose $F$ is a uniformly bounded Hölder class on $[0, 1]$ with $\beta > 1/2$. Then

$$\max_{i,j \in \{1,2,3\}} \Delta_F(E^{(i)}, E^{(j)}) \to 0,$$

as $n \to \infty$.

The density estimation model is shown to be asymptotically equivalent to the following Gaussian white model

$$dY(t) = 2\sqrt{f(t)} \, dt + \frac{1}{\sqrt{n}} \, dW(t),$$

provided $F$ consists of densities $f$ on $[0, 1]$ that are uniformly bounded away from zero and have Hölder smoothness larger than $1/2$ (the same threshold above), cf. Nussbaum (1996). The threshold $\beta > 1/2$ is known to be sharp, a result due to Brown and (C.-H.) Zhang (1998).

**11.4. Minimax lower bound.** For kernel density estimator and projection estimator, we have shown that the maximal risk over the whole parameter space satisfies

$$\sup_{\theta \in \Theta} E_{\theta} d^2(\hat{\theta}_n, \theta) \leq C\psi_n^2$$

for some sequence $\psi_n \to 0$ and some constant $C > 0$ that does not depend on the sample size $n$. The purpose of this section is to provide a lower bound for the above display in the minimax sense: we will show that

$$\inf_{\hat{\theta}_n} \sup_{\theta \in \Theta} E_{\theta} d^2(\hat{\theta}_n, \theta) \geq c\psi_n^2$$

for some constant $c > 0$ that does not depend on the the sample size $n$.

**Definition 11.21.** Let $(P_\theta : \theta \in \Theta)$ be a model (i.e. a collection of probability measures), where $(\Theta, d)$ is a semi-metric space. The **minimax risk over** $\Theta$ is defined by

$$R^*_n(\Theta) \equiv \inf_{\hat{\theta}_n} \sup_{\theta \in \Theta} E_{\theta} d^2(\hat{\theta}_n, \theta).$$

We say that $\{\psi_n\}_{n=1}^{\infty}$ is the minimax optimal rate of convergence over $\Theta$, iff

$$\limsup_n \psi_n^{-2} R^*_n(\Theta) < \infty, \quad \liminf_n \psi_n^{-2} R^*_n(\Theta) > 0.$$

Here is a general reduction scheme.

**Lemma 11.22.** Fix $s > 0$. Let $\theta_0, \ldots, \theta_M \in \Theta$ be such that $d(\theta_j, \theta_k) \geq 2s$ for $j \neq k$. Let

$$p_{\phi, M} \equiv \inf_{\phi} \max_{0 \leq j \leq M} P_{\theta_j}(\phi \neq j),$$
where the infimum is taken over all tests $\phi \in \{0, \ldots, M\}$. Then
\[
R_n^*(\Theta) \geq s^2 \cdot p_{e,M}.
\]

**Proof.** First note that by Markov inequality,
\[
R_n^*(\Theta) \geq s^2 \cdot \inf_{\theta_n} \sup_{\theta \in \Theta} P_\theta(d(\tilde{\theta}_n, \theta) \geq s).
\]
Now for any $\tilde{\theta}_n$, consider the following minimum distance test:
\[
\tilde{\phi} \equiv \arg\min_{0 \leq k \leq M} d(\tilde{\theta}_n, \theta_k).
\]
We claim that
\[
\tilde{\phi} \neq j \Rightarrow d(\tilde{\theta}_n, \theta_j) \geq s.
\]
To see this, note that if $d(\tilde{\theta}_n, \theta_j) < s$, since $\phi \neq j$, there exists some $j^*$ minimizing the distance $j \mapsto d(\tilde{\theta}_n, \theta_j)$, so in particular $d(\tilde{\theta}_n, \theta_{j^*}) \leq d(\tilde{\theta}_n, \theta_j) < s$, which necessarily entails $d(\theta_{j^*}, \theta_j) < 2s$, a contradiction to the construction of $\{\theta_0, \ldots, \theta_M\}$. Hence,
\[
P_{\theta_j}(d(\tilde{\theta}_n, \theta_j) \geq s) \geq P_{\theta_j}(\tilde{\phi} \neq j).
\]
This implies that
\[
\sup_{\theta \in \{\theta_0, \ldots, \theta_M\}} P_\theta(d(\tilde{\theta}_n, \theta) \geq s) \geq \max_{0 \leq j \leq M} P_{\theta_j}(\tilde{\phi} \neq j) \geq p_{e,M},
\]
as desired. \hfill $\square$

The above lemma reduces the problem of bounding from below the minimax risk by that of the minimax testing error $p_{e,M}$ over carefully chosen hypothesis $\{\theta_0, \ldots, \theta_M\}$. The next proposition gives some preliminary lower estimates on the quantity $p_{e,M}$.

**Proposition 11.23.** Let $\{P_j\}_{j=1}^M$ be probability measures absolutely continuous with respect to $P_0$. Then
\[
p_{e,M} = \inf_{\phi} \max_{0 \leq j \leq M} P_j(\phi \neq j) \geq \sup_{\tau > 0} \frac{\tau M}{1 + \tau M} \left[ \frac{1}{M} \sum_{j=1}^M P_j \left( \frac{dP_0}{dP_j} \geq \tau \right) \right].
\]

**Proof.** Let $A_j \equiv \{dP_0/dP_j \geq \tau\}$, then
\[
P_0(\phi \neq 0) = \sum_{j=1}^M P_0(\phi = j) \geq \sum_{j=1}^M \int_{A_j} 1_{\phi = j}\frac{dP_0}{dP_j} dP_j \geq \sum_{j=1}^M \tau P_j(\{\phi = j\} \cap A_j)
\]
\[
\geq \tau M \left[ \frac{1}{M} \sum_{j=1}^M P_j(\phi = j) \right] - \tau M \left[ \frac{1}{M} \sum_{i=1}^M P_j(A'_i) \right] \equiv \tau M(p_0 - \alpha).\]
Hence,
\[
\max_{1 \leq j \leq M} P_j(\phi \neq j) = \max \left\{ P_0(\phi \neq 0), \max_{1 \leq j \leq M} P_j(\phi \neq j) \right\} \\
\geq \max \left\{ \tau M(p_0 - \alpha), \frac{1}{M} \sum_{j=1}^{M} P_j(\phi \neq j) \right\} \\
= \max \left\{ \tau M(p_0 - \alpha), 1 - p_0 \right\} \\
\geq \inf_{p \in [0,1]} \max \{ \tau M(p - \alpha), 1 - p \} = \frac{\tau M(1 - \alpha)}{1 + \tau M},
\]
as desired. The last inequality follows by the observation that the infimum over \( p \in [0,1] \) is attained at the kink point for the function \( p \mapsto \max \{ \tau M(p - \alpha), 1 - p \} \).

The proposition shows that in order to bound from below \( p_{e, M} \), we only need to work with a good lower bound for \( P_j \left( \frac{dP_0}{dP_j} \geq \tau \right) \). To this end, we will need the notion of Kullback-Leibler divergence between two probability measures \( P, Q \).

**Definition 11.24.** The Kullback-Leibler divergence between two probability measures \( P, Q \), denoted \( D(P||Q) \), is defined by
\[
D(P||Q) \equiv \begin{cases} 
\int \log \frac{dP}{dQ} \, dP, & \text{if } P \ll Q; \\
+\infty, & \text{otherwise.}
\end{cases}
\]

**Lemma 11.25.** The following inequalities hold.

1. (Pinsker’s first inequality)
\[
d_{TV}(P, Q) \leq \sqrt{D(P||Q)/2}.
\]

2. (Pinsker’s second inequality)
\[
\int \left| \log \frac{dP}{dQ} \right| \, dP \leq D(P||Q) + \sqrt{2D(P||Q)}.
\]
\[
\int \left( \log \frac{dP}{dQ} \right)_+ \, dP \leq D(P||Q) + \sqrt{D(P||Q)/2}
\]

**Proof.** We assume without loss of generality that \( P \ll Q \) (otherwise the inequalities are trivial). For (1), define \( \psi(x) = x \log x - x + 1 \). By some tedious calculus, it can be shown that
\[
(4 + 2x)\psi(x) \geq 3(x - 1)^2, \quad \forall x \geq 0.
\]

Hence
\[
d_{TV}(P, Q) = \frac{1}{2} \int |p - q| = \frac{1}{2} \int_{q > 0} \left| \frac{p}{q} - 1 \right| q \\
\leq \frac{1}{2} \int_{q > 0} \sqrt{\left( \frac{4}{3} + \frac{2p}{3q} \right) \psi \left( \frac{p}{q} \right) q}.
\]
\[ \frac{1}{2} \sqrt{\int \left( \frac{4q}{3} + \frac{2p}{3} \right) \cdot \sqrt{\int_{q>0} q^2(p/q)}} \]

\[ = \sqrt{\frac{1}{2} \int_{q>0} p \log(p/q) = \sqrt{D(P||Q)/2}}. \]

The last inequality in the above display uses Cauchy-Schwarz.

For (2), we only prove the first inequality. The second is completely analogous. To see this,

\[ \int \left| \log \frac{dP}{dQ} \right| dP = \int \left( \log \frac{dP}{dQ} \right)_+ dP + \int \left( \log \frac{dP}{dQ} \right)_- dP \]

\[ = D(P||Q) + 2 \int \left( \log \frac{dP}{dQ} \right)_- dP \]

\[ \leq D(P||Q) + 2 d_{TV}(P, Q) \]

\[ \leq D(P||Q) + \sqrt{2D(P||Q)}, \]

as desired. \[ \square \]

**Proposition 11.26.** Let \( P_0, \ldots, P_M \) be probability measures such that

\[ \frac{1}{M} \sum_{j=1}^{M} D(P_j||P_0) \leq \alpha_*. \]

Then

\[ p_{e,M} \geq \sup_{\tau>0} \left[ \frac{\tau M}{1 + \tau M} \left( 1 + \alpha_* + \sqrt{\alpha_*/2} \right) \right]. \]

**Proof.** Note that

\[ \frac{1}{M} \sum_{j=1}^{M} P_j \left( \frac{dP_0}{dP_j} \geq \tau \right) \]

\[ = \frac{1}{M} \sum_{j=1}^{M} \left[ 1 - P_j \left( \log \frac{dP_j}{dP_0} > \log(1/\tau) \right) \right] \]

\[ \geq \frac{1}{M} \sum_{j=1}^{M} \left[ 1 - \frac{1}{\log(1/\tau)} \int \left( \log \frac{dP_j}{dP_0} \right)_+ dP_j \right] \]

(by Markov’s inequality)

\[ \geq \frac{1}{M} \sum_{j=1}^{M} \left[ 1 - \frac{1}{\log(1/\tau)} \left( D(P_j||P_0) + \sqrt{D(P_j||P_0)/2} \right) \right] \]

(by Pinsker’s second inequality)

\[ = 1 - \frac{1}{\log(1/\tau)} \left[ \frac{1}{M} \sum_{j=1}^{M} D(P_j||P_0) + \frac{1}{\sqrt{2}} \left( \frac{1}{M} \sum_{j=1}^{M} \sqrt{D(P_j||P_0)} \right) \right] \]
\[ \geq 1 - \frac{1}{\log(1/\tau)} (\alpha_* + \sqrt{\alpha_*/2}). \]

The claim follows from Proposition 11.23. \qed

Now we are in position to state the main minimax theorem using Kullback-Leibler divergence. This is usually known as Le Cam’s method.

**Theorem 11.27.** Let \( M \geq 2 \). Suppose \( \{\theta_0, \ldots, \theta_M\} \) are chosen such that

- \( d(\theta_j, \theta_k) \geq 2s \) for all \( j \neq k \);
- \( P_j \ll P_0 \) for all \( j \) and for some \( \alpha \in (0, 1/8) \),

\[ \frac{1}{M} \sum_{j=1}^{M} K(P_j, P_0) \leq \alpha \log M. \]

Then

\[ s^{-2} R_n^*(\Theta) \geq \frac{\sqrt{M}}{1 + \sqrt{M}} \left( 1 - 2\alpha - \sqrt{2\alpha/\log M} \right) > 0. \]

**Proof.** Apply Proposition 11.26 with \( \alpha_* \equiv \alpha \log M \) and let \( \tau = 1/\sqrt{M} \) in the supremum therein to see that

\[ p_{e,M} \geq \frac{\sqrt{M}}{1 + \sqrt{M}} \left( 1 - 2\alpha - \sqrt{2\alpha/\log M} \right) \geq \frac{\sqrt{M}}{1 + \sqrt{M}} \left( 1 - 2\alpha - \sqrt{2\alpha/\log 2} \right) > 0 \]

for \( \alpha \in (0, 1/8) \). \qed

Now we will use Theorem 11.27 to prove a lower bound for estimation in Hölder class in the nonparametric regression setting. Before this, we still need one more technical tool.

**Lemma 11.28 (Varshamov-Gilbert bound).** Let \( \Omega = \{0, 1\}^m \) with \( m \geq 8 \).

Then there exists \( \{\omega^{(0)}, \ldots, \omega^{(M)}\} \subset \Omega \) such that

\[ d_H(\omega^{(j)}, \omega^{(k)}) = \sum_{\ell=1}^{m} 1(\omega^{(j)}_\ell \neq \omega^{(k)}_\ell) \geq m/8, \]

and \( M \geq 2^{m/8} \). In other words, the packing number for the space \( \Omega \) satisfies

\[ \log \mathcal{D}(m/8, \Omega, d_H) \geq m/8. \]

The distance \( d_H \) is usually referred to the Hamming distance on the unit cube \( \Omega \).

**Proof.** Let \( \omega^{(0)} = (0, \ldots, 0) \) and \( \Omega_0 = \Omega \). For defined \( \omega^{(j-1)} \), let \( \Omega_j \equiv \{\omega \in \Omega_{j-1} : d_H(\omega, \omega^{(j-1)}) > [m/8]\} \), and let \( \omega^{(j)} \in \Omega \) be any element. This recursive definition terminates at step \( M \) in the sense that \( \Omega_M \neq \emptyset \) while...
$\Omega_{M+1} = \emptyset$. Let $A_j \equiv \{\omega \in \Omega_j : d_H(\omega, \omega^{(j)}) \leq \lfloor m/8 \rfloor \}$ be the excluded set at step $j$, and $n_j = |A_j|$ be its cardinality. By definition,

$$\max_{1 \leq j \leq M} n_j \leq \sum_{i=1}^{\lfloor m/8 \rfloor} \binom{m}{i}.$$

On the other hand, since $\{A_j\}$’s are disjoint, it follows that

$$2^m = |\Omega| = n_0 + \ldots + n_M \leq (M + 1) \sum_{i=1}^{\lfloor m/8 \rfloor} \binom{m}{i},$$

which is equivalent to

$$M + 1 \geq \frac{1}{\sum_{i=1}^{\lfloor m/8 \rfloor} 2^{-m} \binom{m}{i}} = \frac{1}{P(\text{Binom}(m, 1/2) \leq \lfloor m/8 \rfloor)} = \frac{1}{P(\sum_{i=1}^{n} (Z_i - 1/2) \leq -3m/8)}.$$

Here $Z_i$’s are i.i.d. Bernoulli random variables taking values $\{0, 1\}$ with equal probability. By Hoeffding’s inequality, it follows that

$$P\left(\sum_{i=1}^{n} (Z_i - 1/2) \leq -3m/8\right) \leq \exp(-2(9m^2/64)/m) = \exp(-9m/32) \leq 2^{-m/4}.$$

Hence

$$M \geq 2^{m/4} - 1 \geq 2^{m/8}$$

for $m \geq 8$.

Now back to the regression setting (11.4), for each $n$, we need to construct good hypothesis $\{f_1, \ldots, f_M\}$ ($M$ depends on $n$) to get the lower bound $s^2 \equiv n^{-\frac{2\beta}{2\beta + 1}}$.

To this end, choose a kernel $K$ such that $K \in \Sigma(\beta, 1/2) \cap C^\infty(\mathbb{R})$ and supp$(K) \subset (-1/2, 1/2)$. Let $m$ be an integer to be chosen later on. Let the bandwidth $h = 1/m$, $x_k = (k - 1/2)/m$ and

$$\phi_k(x) \equiv L \cdot h^\beta K\left(\frac{x - x_k}{h}\right), \quad k = 1, \ldots, m, \quad x \in [0, 1].$$

Some easy to check properties of $\phi_k$:

- $\{\phi_k\}$’s have disjoint supports.
- $\phi_k \in \Sigma(\beta, L/2)$.

Now for any $\omega \in \{0, 1\}^m$, define

$$f_\omega(x) \equiv \sum_{k=1}^{m} \omega_k \phi_k(x).$$
Hence for any $\omega, \omega' \in \{0, 1\}^m$,
\[
d(f_\omega, f_{\omega'}) = \left[ \int_0^1 (f_\omega(x) - f_{\omega'}(x))^2 \, dx \right]^{1/2} = \left[ \sum_{k=1}^m (\omega_k - \omega'_k)^2 \int_{(k-1)/m}^{k/m} \phi_k^2(x) \, dx \right]^{1/2} = L h^{\beta+1/2} \|K\|_2 \left[ \sum_{k=1}^m (\omega_k - \omega'_k)^2 \right]^{1/2} = L h^{\beta+1/2} \|K\|_2 \sqrt{d_H(\omega, \omega')}.
\]

By Varshamov-Gilbert bound, there exists a subset $\{\omega^{(0)}, \ldots, \omega^{(M)}\} \subset \Omega$ such that each pair of elements are $m/8$ away from each other in the Hamming distance $d_H$, and the cardinality $M \geq 2^{m/8}$. The hypothesis is constructed via
\[
\{f_j \equiv f_{\omega^{(j)}} : 0 \leq j \leq M\}
\]
with
\[
m = \left\lfloor \rho \cdot n^{\frac{1}{2\beta+1}} \right\rfloor
\]
for some large enough $\rho > 0$ constant that will be clarified later on.

To apply Theorem 11.27, we need to verify the following conditions:
(1) $f_j \in \Sigma(\beta, L)$.
(2) $\|f_j - f_k\|_2 \geq 2s$ for $j \neq k$.
(3) $\frac{1}{M} \sum_{j=1}^M D(P_{f_j}||P_0) \leq \alpha \log M$ for some $\alpha \in (0, 1/8)$.

(1) is easy to check by the fact that $\phi_k$’s have disjoint supports and $\phi_k \in \Sigma(\beta, L/2)$. For (2), if $j \neq k$,
\[
\|f_j - f_k\|_2 = L h^{\beta+1/2} \|K\|_2 \sqrt{d_H(\omega^{(j)}, \omega^{(k)})} \geq C_{L,K} m^{-\beta} = C_{L,K} \rho^{-\beta} n^{-\frac{\beta}{2\beta+1}}.
\]

Now assume Gaussian errors $\xi_i$’s and the design points are $X_i = i/n$, we have that
\[
D(P_{f_j}||P_0) \approx \sum_{i=1}^n f_j^2(X_i) = \sum_{k=1}^m \sum_{X_i \in ((k-1)/m, k/m]} \phi_k^2(X_i)
\]
\[
\leq L^2 h^{2\beta} \|K\|_2^2 \sum_{k=1}^m \sum_{i=1}^n \{i : X_i \in ((k-1)/m, k/m]\}
\]
\[
\leq C_{K,L'} nh^{2\beta} = C_{K,L'} \rho^{-2\beta} n^{\frac{1}{2\beta+1}} \leq \rho n^{\frac{1}{2\beta+1}}/64 \leq m/64 \leq \frac{1}{8\log 2} \log M.
\]

In the above sequence of inequalities we choose $\rho > 0$ large enough to ensure that $C_{K,L'} \rho^{-2\beta} n^{\frac{1}{2\beta+1}} \leq \rho n^{\frac{1}{2\beta+1}}/64$. Now conditions (1)-(3) are verified so we may apply Theorem 11.27 to obtain the following:
Theorem 11.29. Consider the nonparametric regression model (11.4) with Gaussian errors \( \xi_i \sim N(0,1) \) and the design points are \( X_i = i/n \) for \( i = 1, \ldots, n \). Then there exists some \( c > 0 \) such that
\[
\inf_{\tilde{f}_n} \sup_{f \in \Sigma(\beta,L)} E_f [\|\tilde{f}_n - f\|_2^2] \geq c \cdot n^{-\frac{2\beta}{2\beta+1}}.
\]

Similar construction can also be used for the class \( W^*(\beta,L) \), so combined with Theorem 11.18 we see that the projection estimator achieves the minimax optimal rate of convergence.
References


