Asymptotics Under Nonstandard Conditions

A Dissertation
Presented to the Faculty of the Graduate School
of
Yale University
in Candidacy for the Degree of
Doctor of Philosophy

by
Peter Radchenko

Dissertation Director: David Pollard

May 2004

Contents

List of Figures

Acknowledgements

I am indebted to my advisor, Professor David Pollard, for the enormous amount of time, energy, and knowledge, that he has given to me so generously. Without his guidance and support, this work would not have been completed.

I am very grateful to Professor John Hartigan for his help with the k-means part of the dissertation. He spent a lot of his time discussing this topic with me, and shared many valuable ideas and suggestions.

I would like to thank Professor Marten Wegkamp for helpful discussions of the non-linear least squares part of the dissertation.

I am grateful to Professors Andrew Barron, Joseph Chang, and Nicholas Hengartner, for their guidance in and out of the classroom.

I very much appreciate the support and help of Kathryn Young and Susan Jackson-Mack, Yale Statistics Department's Business Manager and Administrative Assistant.

I would like to thank Professors Regis Serinko and Jogesh Babu for drawing my attention to their results on univariate k-means clustering.

Preface

The first three chapters of this thesis are devoted to asymptotics in k-means clustering. Chapter 1 reviews existing asymptotic results and summarizes my contribution.

Chapter 2 considers two examples in k-means where the asymptotics is nonstandard. The first example is the two means problem on the real line, when the underlying distribution is double exponential. The second derivative of the population criterion function is singular, which causes a slower convergence rate of $n^{-1/4}$ for the optimal sample centers. I have recently discovered that Serinko and Babu (1992) were the first to recognize the unusual asymptotic behavior in this example. They used one-dimensional techniques to derive the $n^{-1/4}$ asymptotics for the sample centers and the split probability. I obtain similar results using empirical process techniques developed in Pollard (1981) and Pollard (1982a) for the general multidimensional case. In addition I derive the $n^{-1/2}$ rate of convergence and a limit theorem for the distance between the optimal sample centers. I also consider a delicate two dimensional example where several rates of convergence are present. The irregularity of the asymptotics in this example comes from two sources: two (rather than one) optimal configurations of population centers and a singular second derivative of the population criterion function for one of these optimal configurations. I derive the limiting behavior of the optimal sample centers.

Chapter 3 deals with the case of non-unique population minima in the general setting. The consistency result that I give is an extension of the result in Pollard (1981). I also

present conditions on the population criterion function that guarantee an $n^{-1/2}$ rate of convergence of the optimal sample centers to the set of all population minima. I prove a central limit theorem that handles the case of non-unique population minima.

To the best of my knowledge, there are no asymptotic results in the literature that handle the multidimensional k-means problem either in the case of non-unique population solution or in the case of a singular second derivative of the population criterion function.

Chapter 4 is a joint work on asymptotics of nonlinear least squares with my adviser, Professor David Pollard, copied verbatim from Pollard and Radchenko (2003). The motivation for this research was an example of Wu (1981), which his limit theorems could not handle. We established the asymptotics for the example by means of new limit theorems, extending ideas of Wu (1981) and Van de Geer (1990). My contribution included deriving the consistency result and the central limit theorem for Wu's example.

Chapter 1

Introduction

1.1 Review of Standard Asymptotic Results

The k-means procedure divides observations $x_1, \ldots, x_n$ in $\mathbb{R}^d$ into k sets by locating the cluster centers and then assigning each observation to the closest center. The set of cluster centers $b_n = \{b_{1n}, \ldots, b_{kn}\}$ is chosen to minimize

$$W_n(a) = n^{-1} \sum_{i \le n} \min_{1 \le j \le k} |x_i - a_j|^2 \tag{1.1}$$

as a function of sets $a = \{a_1, \ldots, a_k\}$ of k not necessarily distinct points in $\mathbb{R}^d$. I write $|\cdot|$ for the usual Euclidean norm in $\mathbb{R}^d$.

The observations are independent and are coming from a population distribution P. For every set $b = \{b_1, \ldots, b_k\}$ of k points in $\mathbb{R}^d$, define $\phi(x, b)$ to be the squared distance from x to the closest point in b,

$$\phi(x, b) = \min_{1 \le j \le k} |x - b_j|^2.$$

The empirical measure P_n places mass n^{-1} at each $x_1, \ldots, x_n$. I will use the abbreviation

$Qf = \int f dQ$ for a given measurable function f and a signed measure Q. Note that that $W_n(b)$ is the P_n-expected squared distance to the closest point in b,

$$W_n(b) = P_n\phi(\cdot, b).$$

Let $W(b)$ be the population counterpart,

$$W(b) = P\phi(\cdot, b).$$

I will refer to W as the **population criterion function**, and I will call W_n the **empirical criterion function**. Note that for each fixed set b, the sample value $W_n(b)$ converges to its expectation $W(b)$ by the law of large numbers. This suggest that the empirical and the population criterion functions get close to each other as the sample size increases, and hence their minima should be close. Suppose a set b_n minimizes the function W_n and a set a minimizes the function W. Call b_n a set of **optimal sample centers**, and call a a set of **optimal population centers**.

Note that if P has a finite second moment and is not concentrated on fewer than k points, each set of optimal population centers has to contain exactly k points. Under these conditions, and given that the set a of optimal population centers is unique, Pollard (1981) showed that the sets b_n of optimal empirical centers are strongly consistent with respect to the Hausdorff metric $H(\cdot, \cdot)$. The Hausdorff metric is defined for compact subsets E, F of $\mathbb{R}^d$ by

$$H(E, F) = \left(\max_{e \in E} \min_{f \in F} |e - f| \right) \wedge \left(\max_{f \in F} \min_{e \in E} |e - f| \right).$$

The idea of Pollard's consistency proof is the following. First invoke a compactification argument to get all the optimal empirical centers almost surely within a compact region in $\mathbb{R}^d$. Then establish a uniform strong law of large numbers for W_n and prove continuity of

the limit function W. Almost sure convergence of the minimum of W_n to the minimum of W follows directly.

Note that if we take a δ less than half of the smallest distance between points of a, then the sets b that satisfy $H(b, a) < \delta$ have to contain exactly k points, each lying within a δ Euclidean distance of a uniquely determined point in a. Hence, convergence of b_n in the Hausdorff metric can be translated into convergence of individual points. Write the set $a = \{a_1, \ldots, a_k\}$ as an $\mathbb{R}^{kd}$ vector $(a_1, \ldots, a_k)$. Translate convergence of b_n to a into convergence of vectors in $\mathbb{R}^{kd}$ with respect to the Euclidean metric. Note that when $H(b_n, a)$ is small enough, the **$\mathbf{R^{kd}}$-representation** of b_n is uniquely determined by the labeling of points in a. Throughout this text, when working in a small Hausdorff neighborhood of the set a, I will treat sets in this neighborhood as vectors in $\mathbb{R}^{kd}$ assuming that the vector representation is induced by a particular labeling of points in a. I will abuse the notation and use the same symbol for sets and vectors.

Under some regularity conditions, Pollard (1982a) derived a central limit theorem for the vector of optimal empirical centers. Let Γ be the second derivative matrix of W evaluated at the vector a of optimal population centers. Under some smoothness assumptions about the underlying distribution Pollard established the following quadratic approximation to the empirical criterion function uniformly over h in shrinking neighborhoods of zero,

$$W_n(a + h) - W_n(a) = (1/2)h'\Gamma h - n^{-1/2}h'Z_n + o_p(n^{-1/2}|h|) + o(|h|^2). \qquad (1.2)$$

I write that a sequence of random functions $g_n(h)$ is of order $o_p(n^{-1/2}|h|)$ **uniformly in shrinking neighborhoods** of zero, if for every sequence $\{r_n\}$ of positive numbers converging to zero there exists a $o_p(1)$ sequence of random variables ϵ_n, such that the upper bound $|g_n(h)| \leq \epsilon_n n^{-1/2}|h|$ holds whenever $|h| \leq r_n$. Note that this definition can be

generalized to include uniform approximations over a class of sequences of random neighborhoods, for example the class of $O_p(n^{-1/2})$ neighborhoods. By the stochastic order of a sequence of random neighborhoods I understand the stochastic order of the corresponding sequence of diameters.

For nonsingular Γ, Pollard applied comparison arguments to approximation (1.2), and derived that

$$b_n = a + n^{-1/2}\Gamma^{-1}Z_n + o_p(n^{-1/2}), \tag{1.3}$$

which implies a central limit theorem for b_n because vectors Z_n are asymptotically normal with mean zero.

Below are some ideas behind Pollard's derivation of approximation (1.2). Let ν_n denote the empirical process,

$$\nu_n(\cdot) = n^{1/2}\left(P_n(\cdot) - P(\cdot)\right).$$

Observe that

$$W_n(a+h) - W_n(a) = W(a+h) - W(a) - n^{-1/2}\nu_n^x\left(\phi(x, a+h) - \phi(x,a)\right).$$

Write a Taylor expansion for the function W near a. The linear term in the expansion vanishes because a is the minimum of W. Finally, extract a linear term in h from the stochastic part and use empirical process techniques to control the remainder terms. For more detail see Pollard (1982a) or read the proofs of my Lemma 3 and Lemma 4 of Section 2.1, which follow in Pollard's footsteps in deriving a similar approximation.

Pollard's results generalize the one dimensional consistency and central limit theorems of Hartigan (1978). They also extend the results of MacQueen (1967), who obtained consistency for the within group sum of squares of a k-means algorithm that distributes

the observations sequentially among k clusters. For a discussion of k-means algorithms and practical applications see Hartigan (1975).

1.2 Motivation

The theory developed in the statistical literature for the method of k-means has been applied to the study of k-level quantizers (see, for example, Pollard 1982b). A k-level **quantizer** is a map q from $\mathbb{R}^d$ into a subset $\{b_1, \ldots, b_k\}$ of itself. Such a map is used to convert an input signal in $\mathbb{R}^d$ into an output that takes on at most k values. An optimal quantizer for the signal coming from a probability distribution P minimizes the **distortion**, which is measured by the mean-squared error $P^x|x - q(x)|^2$. For an exposition of general quantization theory see, for example, Graf and Luschgy (2000).

Note that the set of values of an optimal quantizer minimizes the corresponding criterion function W discussed earlier. Also note, that if $\{a_1, \ldots, a_k\}$ is a set of optimal population centers then the quantizer that maps each x into its nearest center a_i is optimal. Hence, the k-level quantization problem is equivalent to the k-means problem.

The motivation for my research on k-means came from a question posed by Bartlett, Linder, and Lugosi (1998). Temporarily write $W(P, \cdot)$ instead of $W(\cdot)$, to indicate the dependence on the underlying distribution, define $W^*(P) = \inf_a W(P, a)$, and write $\mathcal{P}_{\sqrt{d}}$ for the set of all probability measures P that concentrate on the closed ball of radius $\sqrt{d}$ centered at the origin. In this notation, their question becomes: If we require

$$\mathbb{P}[W(P, a_n) - W^*(P)] \leq \alpha_n C(P) \quad \text{for all } P \in \mathcal{P}_{\sqrt{d}}, \tag{1.4}$$

where $C(P)$ is a constant that depends only on P, how fast can the sequence $\{\alpha_n\}$ go to zero? The expectation on the left-hand side is taken over samples from P. More formally,

$\mathbb{P}$ should be understood as the n-fold product measure P^n. It is important that the set a_n depends only on the sample, for otherwise the left-hand side could be made equal to zero by taking a_n to be the set of optimal population centers for P. Under a further restriction that $\sup_{P \in \mathcal{P}_{\sqrt{d}}} C(P) < \infty$, Bartlett et al. showed that the uniform rate α_n cannot be better than $n^{-1/2}$. Citing results from Linder, Lugosi, and Zeger 1994 and Devroye, Györfi, and Lugosi 1996, they noted that the n^{-1} uniform rate is achieved by the set b_n of k-means optimal sample centers.

In this dissertation I study the closely related problem of how fast the left-hand side of (1.4) can tend to zero if we take a_n as the set b_n of optimal sample centers, that is the set that minimizes W_n. I also establish the corresponding asymptotic distribution theory for $W(P, b_n) - W^*(P)$, the **distortion redundancy**.

Bartlett et al. pointed out that the distortion redundancy goes to zero at the n^{-1} rate for those P where the regularity conditions of Pollard (1982a) are satisfied, as a consequence of approximation (1.2). To establish this approximation for the empirical criterion function near the population minimum a, Pollard first showed that

$$W_n(a+h) - W_n(a) = W(a+h) - W(a) - n^{-1/2} h' Z_n + o_p(n^{-1/2}|h|), \tag{1.5}$$

then combined it with a Taylor expansion of W around a,

$$W(a+h) - W(a) = (1/2) h' \Gamma h + o(|h|^2). \tag{1.6}$$

I have reverted to the convention of writing $W(\cdot)$ instead of $W(P, \cdot)$. If the matrix Γ is nonsingular, approximation (1.3) and the central limit theorem hold for the set b_n of optimal sample centers. Together with approximation (1.6) they imply

$$W(b_n) - W^* = (1/2) n^{-1} Z_n' \Gamma^{-1} Z_n + o(n^{-1}).$$

As noted earlier, I assume that the underlying distribution has a finite second moment and is not concentrated on less than k points. Under these conditions Z_n has a nonzero limiting distribution, and hence $W(b_n)$ indeed converges to the minimum of W at the rate n^{-1}.

To summarize, if P has a finite second moment and is not concentrated on less than k points and if Pollard's regularity conditions are satisfied, then the central limit theorem holds for the optimal sample centers and the distortion redundancy settles down at the rate n^{-1}. The regularity conditions are as follows:

A: The set a of optimal population centers is unique.

B: Matrix Γ in expansion (1.6) of the population criterion function is nonsingular.

1.3 Summary of the Main Results

In Chapter 2 I consider two examples where the regularity conditions fail. I investigate how a lack of regularity can influence the asymptotic behavior of optimal sample centers and distortion redundancy.

The first example is the two means problem on the real line when the underlying distribution is double exponential. There is a unique optimal pair $\{-1, 1\}$ of population centers, however the second derivative matrix Γ of population criterion function W is singular at $(-1, 1)$. To derive the asymptotics of the optimal sample centers I expand the approximation in (1.6) to include the cubic terms so that (1.5) becomes a cubic approximation to the empirical criterion function. I use comparison arguments to show that optimal centers centers converge at the rate $n^{-1/4}$ and find the limiting distribution. In addition, I refine the error terms in approximation (1.5) and show that the distance between the optimal sample centers converges at the rate $n^{1/2}$. I find the limiting distribution for this distance.

Serinko and Babu (1992) also considered the example described above. They used one-dimensional techniques to derive the $n^{-1/4}$ level asymptotics. They also proved a

general limit theorem for k-means on the real line, tying the rate of convergence of the optimal sample centers to the behavior of a certain population criterion function.

The second example I consider is the two means problem when the underlying distribution is concentrated on two parallel lines in the plane. Each line contains half of the probability, and the conditional distribution on each line is double exponential. The lines are spread apart at the exact distance that insures that there are two configurations of optimal population centers (see figure 2.2.) With positive probabilities the optimal sample centers settle down to one or the other optimal population configuration. The sample centers settle down to one of the configurations at the rate $n^{-1/4}$, and they settle down to the other at the rate $n^{-1/2}$. The distortion redundancy goes to zero at the rate $n^{-3/4}$ in the first case, and at the rate n^{-1} in the second. Moreover, in the first case some linear combinations of the sample centers settle down at the rate $n^{-1/2}$. I derive the joint limiting distribution for all the random quantities contributing to the asymptotics of the sample centers.

Chapter 3 establishes some general results that handle the non-uniqueness of the population minimizer. The lack of a unique set a about which to develop approximations like and (1.5) and (1.6) causes technical difficulties, which, to the best of my knowledge, have not been addressed in the literature. Under some assumptions on the behavior of W near the population minimizers, I establish the $n^{-1/2}$ rate of convergence and a limit theorem for the set of optimal sample centers.

Chapter 2

Nonstandard Rates of Convergence in k-Means

In this chapter I discuss two examples in which Pollard's regularity conditions are not satisfied. The first example is a two means problem on the real line with double exponential underlying distribution. In Section 2.2 I analyze the asymptotics in this example using two methods to parametrize the problem: parametrization by the centers and parametrization by the split point. Parametrization by the centers is the general approach of Pollard that was described in the Introduction. Parametrization by the split point is a traditional way to handle the two means problem on the real line; it is defined and discussed in Section 2.1.

2.1 Two Clusters in One Dimension

2.1.1 Some Definitions

Consider a distribution P on the real line. Assume it has a finite mean μ and a finite variance σ^2. I will not consider distributions that are concentrated on less than two points.

Definition 1 *For each $t \in \mathbb{R}$ denote the probabilities that P assigns to $(-\infty, t]$ and (t, ∞)*

by π_{t-} and π_{t+} respectively, and let $\overline{X}_{t-}$ and $\overline{X}_{t+}$ stand for the corresponding conditional means:

$$\pi_{t+} = P^x\{x > t\} \qquad \pi_{t-} = P^x\{x \le t\}$$
$$\overline{X}_{t+} = P^x x\{x > t\}/\pi_{t+} \qquad \overline{X}_{t-} = P^x x\{x \le t\}/\pi_{t-}.$$

For the corresponding sample quantities use the superscript n. For example,

$$\pi^n_{t+} = P^x_n\{x > t\} \quad \text{and} \quad \overline{X}^n_{t+} = P^x_n x\{x > t\}/\pi^n_{t+}.$$

For any distinct pair of centers on the real line the partition that they define consists of two half-lines $(\infty, s]$ and (s, ∞) where s is the midpoint between the centers. It is convenient to view the 2-means clustering in $\mathbb{R}$ as an optimization over all such partitions. Note that this optimization problem can be parametrized by a single parameter, the point that splits $\mathbb{R}$ into two half-lines.

Definition 2 *Each point s on the real line splits it into two parts, $(-\infty, s]$ and (s, ∞). Denote the **within cluster sum of squares** that corresponds to this partition by $V(s)$:*

$$V(s) = P^x\{x \le s\}(x - \overline{X}_{s-})^2 + P^x\{x > s\}(x - \overline{X}_{s+})^2.$$

The corresponding sample quantity is

$$V_n(s) = P^x_n\{x \le s\}(x - \overline{X}^n_{s-})^2 + P^x_n\{x > s\}(x - \overline{X}^n_{s+})^2.$$

*Denote the **between cluster sum of squares** by $G(s)$:*

$$G(s) = \pi_{s-}\overline{X}^2_{s-} + \pi_{s+}\overline{X}^2_{s+} - \mu^2. \tag{2.1}$$

*Call s the **split point**. A split point is **optimal** if it minimizes the within cluster sum of*

squares function.

Note that

$$V(s) + G(s) = \sigma^2,$$

hence minimizing the within cluster sum of squares is equivalent to maximizing the between cluster sum of squares. In the case of two clusters in one dimension the between cluster sum of squares has been traditionally used as the criterion function (see, for example, Hartigan 1978).

I will use the following natural $\mathbb{R}^2$-representation for two-point sets in $\mathbb{R}$. If a set a consists of two points a_1 and a_2, and $a_1 \leq a_2$, the vector representation for a is (a_1, a_2).

Definition 3 *For a split point s and a set $a = (a_1, a_2)$ define the **generalized sum of squares** function,*

$$\mathcal{W}(s, a) = \mathcal{W}(s, a_1, a_2) = P^x\{x \leq s\}(x - a_1)^2 + P^x\{x > s\}(x - a_2)^2,$$

*and the **bias-squared** function,*

$$B^2(s, a) = B^2(s, a_1, a_2) = \pi_{s-}(\overline{X}_{s-} - a_1)^2 + \pi_{s+}(\overline{X}_{s+} - a_2)^2. \tag{2.2}$$

The corresponding sample quantities are

$$\mathcal{W}_n(s, a) = P_n^x\{x \leq s\}(x - a_1)^2 + P_n^x\{x > s\}(x - a_2)^2,$$

and

$$B_n^2(s, a) = \pi_{s-}^n(\overline{X}_{s-}^n - a_1)^2 + \pi_{s+}^n(\overline{X}_{s+}^n - a_2)^2.$$

Note that when $s = (a_1 + a_2)/2$, the generalized sum of squares $\mathcal{W}(s, a)$ becomes $W(a)$, the population criterion function for the parametrization by the centers. Note the following

equality:

$$\mathcal{W}(s,a) = V(s) + B^2(s,a). \tag{2.3}$$

Observe that for all real s, u, and v, such that $u \leq s \leq v$ and $0 < \pi_{s-} < 1$,

$$\begin{aligned} \mathcal{W}(s,u,v) &\geq V(s) = \mathcal{W}(s, \overline{X}_{s-}, \overline{X}_{s+}) \\ \mathcal{W}(s,u,v) &\geq W(u,v) = \mathcal{W}(u/2 + v/2, u, v). \end{aligned} \tag{2.4}$$

Pollard (1981) showed that function W is continuous when the collection of two-point sets is equipped with the Hausdorff metric. It follows that the minimum value of W is achieved. Apply inequalities (2.4) to deduce that the minimum values of $\mathcal{W}$ and V are also achieved and that these values are equal to the minimum value of W. Observe that if a split point s is optimal, i.e. it minimizes the within cluster sum of squares function V, then the triplets $(s, \overline{X}_{s-}, \overline{X}_{s+})$ and $(\frac{1}{2}(\overline{X}_{s-} + \overline{X}_{s+}), \overline{X}_{s-}, \overline{X}_{s+})$ minimize function $\mathcal{W}$, and the pair $(\overline{X}_{s-}, \overline{X}_{s+})$ minimizes function W. Also note that if an optimal split point s does not coincide with $\frac{1}{2}(\overline{X}_{s-} + \overline{X}_{s+})$, then the open interval with endpoints at s and $\frac{1}{2}(\overline{X}_{s-} + \overline{X}_{s+})$ contains zero probability.

The reasoning in the above paragraph is true for any underlying probability distribution. The conclusions apply to the sample functions $\mathcal{W}_n$, W_n, and V_n.

2.1.2 Approximation to Sample Within Sum of Squares Function

Approximation (1.5) of Pollard (1982a) is a crucial step to deriving asymptotics of the optimal sample centers. Hartigan (1978) gives a similar approximation to the sample between sum of squares function. Here, I derive an approximation to the sample within sum of squares function following the general approach of Pollard (1982a).

Without loss of generality, suppose that the optimal population split point is at zero. Assume, as always, that the population distribution is not concentrated on just one point,

hence probabilities π_{0-} and π_{0+} are both positive. The following lemma approximates empirical cluster means by the corresponding population conditional means.

Lemma 1 *The following approximations hold uniformly over s in shrinking neighborhoods of zero:*

$$\begin{aligned} \overline{X}^n_{s-} - \overline{X}_{s-} &= n^{-1/2}\nu^x_n(x - \overline{X}_{0-})\{x \leq 0\}/\pi_{0-} + o_p(n^{-1/2}) \\ \overline{X}^n_{s+} - \overline{X}_{s+} &= n^{-1/2}\nu^x_n(x - \overline{X}_{0+})\{x \leq 0\}/\pi_{0+} + o_p(n^{-1/2}). \end{aligned} \tag{2.5}$$

Proof: Note that $\pi^n_{s-} = \pi_{s-} + n^{-1/2}\nu^x_n\{x \leq s\}$. Observe that

$$\begin{aligned} \overline{X}^n_{s-} - \overline{X}_{s-} &= \frac{P^x_n x\{x \leq s\} - P^x x\{x \leq s\}}{\pi^n_{s-}} - P^x x\{x \leq s\}\frac{\pi^n_{s-} - \pi_{s-}}{\pi_{s-}\pi^n_{s-}} \\ &= \frac{n^{-1/2}\nu^x_n x\{x \leq s\}}{\pi^n_{s-}} - \frac{n^{-1/2}\nu^x_n\{x \leq s\}P^x x\{x \leq s\}}{\pi_{s-}\pi^n_{s-}}. \end{aligned} \tag{2.6}$$

As s goes to zero, functions $f_s(x) = x\{0 < x \leq s\}$ and $g_s(x) = \{0 < x \leq s\}$ converge to zero in the L_2 sense by dominated convergence. Note that as s ranges over a neighborhood of zero, classes of functions $\{f_s\}$ and $\{g_s\}$ satisfy the conditions of Theorem 7 in the appendix. Hence, processes $\nu^x_n f_s(x)$ and $\nu^x_n g_s(x)$ are stochastically equicontinuous at zero, and thus $\nu^x_n f_s(x) = o_p(1)$ and $\nu^x_n g_s(x) = o_p(1)$ uniformly over s in shrinking neighborhoods of zero. Deduce the following approximations,

$$\begin{aligned} \pi_{s-} &= \pi_{0-} + o(1) \\ P^x x\{x \leq s\} &= P^x x\{x \leq 0\} + o_p(1) \\ \nu^x_n\{x \leq s\} &= \nu^x_n\{x \leq 0\} + o_p(1) \\ \pi^n_{s-} &= \pi_{0-} + o_p(1) \\ \nu^x_n x\{x \leq s\} &= \nu^x_n x\{x \leq 0\} + o_p(1), \end{aligned} \tag{2.7}$$

which hold uniformly over s in shrinking neighborhoods of zero. Plug these approximations into expression (2.6) and deduce that

$$\overline{X}^n_{s-} - \overline{X}_{s-} = n^{-1/2}\nu^x_n(x - \overline{X}_{0-})\{x \leq 0\}/\pi_{0-} + o_p(n^{-1/2}).$$

Argue analogously for $\overline{X}^n_{s+} - \overline{X}_{s+}$. $\square$

Apply equality (2.3) to the empirical probability P_n and derive

$$V_n(s) = \mathcal{W}_n(s, \overline{X}_{s-}, \overline{X}_{s+}) - B^2_n(s, \overline{X}_{s-}, \overline{X}_{s+}).$$

Use approximations (2.5) and (2.7) to handle the bias-squared function uniformly over s in shrinking neighborhoods of zero:

$$\begin{aligned} B^2_n(s, \overline{X}_{s-}, \overline{X}_{s+}) &= \pi^n_{s-}(\overline{X}^n_{s-} - \overline{X}_{s-})^2 + \pi^n_{s+}(\overline{X}^n_{s+} - \overline{X}_{s+})^2 \\ &= n^{-1}\pi_{0-}(\nu^x_n(x - \overline{X}_{0-})\{x \leq 0\})^2 + n^{-1}\pi_{0+}(\nu^x_n(x - \overline{X}_{0+})\{x > 0\})^2 \\ &\quad + o_p(n^{-1}). \end{aligned}$$

Conclude that

$$V_n(s) - V_n(0) = \mathcal{W}_n(s, \overline{X}_{s-}, \overline{X}_{s+}) - \mathcal{W}_n(0, \overline{X}_{0-}, \overline{X}_{0+}) + o_p(n^{-1}),$$

uniformly in shrinking neighborhoods of zero. Note that $\mathcal{W}(s, \overline{X}_{s-}, \overline{X}_{s+})$ is exactly the within cluster sum of squares $V(s)$. Deduce the following

Lemma 2 *Let $\psi(x, s)$ stand for $\{x \leq s\}(x - \overline{X}_{s-})^2 + \{x > s\}(x - \overline{X}_{s+})^2$. The following approximation holds uniformly over s in shrinking neighborhoods of zero:*

$$V_n(s) - V_n(0) = V(s) - V(0) + n^{-1/2}\nu^x_n\Big(\psi(x, s) - \psi(x, 0)\Big) + o_p(n^{-1}).$$

The following lemma helps extract a linear term in s from the random part of the above approximation.

Lemma 3 *Suppose that there exists a neighborhood of zero, such that a restriction of P on this neighborhood has a continuous density function f with respect to Lebesgue measure. The function*

$$\Delta(x) = \frac{2f(0)\overline{X}_{0-}}{\pi_{0-}}(x - \overline{X}_{0-})\{x \leq 0\} - \frac{2f(0)\overline{X}_{0+}}{\pi_{0+}}(x - \overline{X}_{0+})\{x > 0\}$$

is the $L_2(P)$ derivative of $\psi(x, s)$ at $s = 0$, in the sense that the remainder functions $R_s(x)$, defined by

$$\psi(x, s) = \psi(x, 0) + s\Delta(x) + sR_s(x),$$

converge to zero with respect to the $L_2(P)$ pseudo-metric.

Proof: Differentiate $\overline{X}_{s-}$ and $\overline{X}_{s+}$ to derive the following approximations for s tending to zero:

$$\overline{X}_{s-} = \overline{X}_{0-} - sf(0)\overline{X}_{0-}/\pi_{0-} + o(s)$$
$$\overline{X}_{s+} = \overline{X}_{0+} + sf(0)\overline{X}_{0+}/\pi_{0+} + o(s).$$

Fix a negative x. For all s small enough,

$$\begin{aligned} sR_s(x) &= (x - \overline{X}_{s-})^2 - (x - \overline{X}_{0-})^2 - \frac{2f(0)\overline{X}_{0-}}{\pi_{0-}}(x - \overline{X}_{0-}) \\ &= o(s). \end{aligned}$$

Argue analogously for a positive x and conclude that $R_s(x)$ go to zero pointwise at all non-zero x. Since f is continuous, P places no mass at zero, hence functions $R_s(x)$ converge to zero at P-almost all x.

Observe that for all s in a small enough neighborhood of zero,

$$\begin{aligned} |R_s(x)| &\leq |s|^{-1}\left(|s\Delta(x)| + |(x-\overline{X}_{s-})^2 - (x-\overline{X}_{0-})^2| \vee |(x-\overline{X}_{s+})^2 - (x-\overline{X}_{0+})^2|\right) \\ &\leq |\Delta(x)| + |s|^{-1}\left(|(x-\overline{X}_{s-})^2 - (x-\overline{X}_{0-})^2| + |(x-\overline{X}_{s+})^2 - (x-\overline{X}_{0+})^2|\right) \\ &\leq C(1+|x|), \end{aligned} \tag{2.8}$$

where C depends only on the neighborhood. By dominated convergence, functions $R_s(x)$ converge to zero in the $L_2(P)$ sense as s goes to zero. □

Note that for any s, function $R_s(x)$ is a sum of at most three linear functions with disjoint supports. The support of each of these functions in an interval on the real line. Observe also that a class of functions $\{R_s(x)\}$ for s in a small enough neighborhood of zero has a square integrable envelope by inequality (2.8). Hence this class of functions satisfies the conditions of Theorem 7 in the appendix, and thus the empirical process $\{\nu_n^x R_s(x)\}$ is equicontinuous at $s = 0$. Use the fact that functions $R_s(x)$ converge to zero in the $L_2(P)$ sense to conclude that

$$\nu_n^x R_s(x) = o_p(1),$$

uniformly over s in shrinking neighborhoods of zero. Combine this result with the statements of Lemma 2 and Lemma 3 to deduce

Lemma 4 *Suppose that there exist a neighborhood of zero, such that a restriction of P onto this neighborhood has a continuous density function f with respect to Lebesgue measure. Define random variables Z_{1n} and Z_{2n} by*

$$\begin{aligned} Z_{1n} &= [-2f(0)\overline{X}_{0-}/\pi_{0-}]\nu_n^x(x-\overline{X}_{0-})\{x \leq 0\} \\ Z_{2n} &= [2f(0)\overline{X}_{0+}/\pi_{0+}]\nu_n^x(x-\overline{X}_{0+})\{x > 0\}. \end{aligned}$$

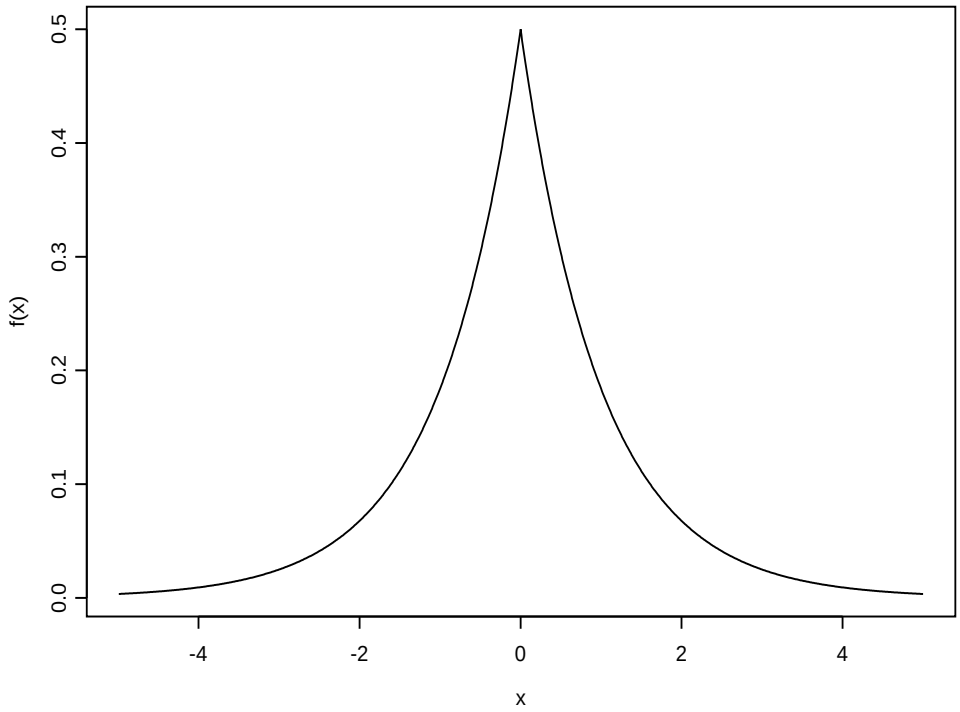

Figure 2.1: Double exponential density

The following approximation holds uniformly over s in shrinking neighborhoods of zero:

$$V_n(s) - V_n(0) = V(s) - V(0) - n^{-1/2}s(Z_{1n} + Z_{2n}) + o_p(n^{-1/2}|s|) + o_p(n^{-1}). \quad (2.9)$$

2.2 Double Exponential Example

2.2.1 Population Criterion Functions

Let P correspond to double exponential distribution on the real line (see figure 2.1). For every t write out the expressions for probabilities π_{t+} and π_{t-} and conditional means X_{t+}

and X_{t-}:

$$\begin{aligned}
\pi_{t+} &= \{t > 0\}e^{-t}/2 + \{t \leq 0\}(1 - e^t/2) \\
\pi_{t-} &= \{t > 0\}(1 - e^{-t}/2) + \{t \leq 0\}e^t/2 \\
\overline{X}_{t+} &= \{t > 0\}(1 + t) + \{t \leq 0\}(1 - t)\pi_{t-}/\pi_{t+} \\
\overline{X}_{t-} &= \{t > 0\}(-1 - t)\pi_{t+}/\pi_{t-} + \{t \leq 0\}(-1 + t).
\end{aligned} \tag{2.10}$$

The between cluster sum of squares function $G(s)$ is symmetric about zero. Consider a nonnegative s. Plug in the above expressions for conditional means $\overline{X}_{s-}$ and $\overline{X}_{s+}$ into formula (2.1) for G to derive $G(s) = (1 + s)^2\pi_{s+}/\pi_{s+}$. Rewrite this expression using the above formulas for π_{s+} and π_{s+} to get

$$G(s) = \frac{(1 + s)^2}{2e^s - 1} = \frac{1 + 2s + s^2}{1 + 2s + s^2 + s^3/3 + \ldots}, \quad \text{for } s \geq 0.$$

Deduce $G(s) < G(0)$ for all positive s. Thus, zero is the unique optimal split point for the double exponential distribution.

Apply Taylor expansion to approximate $G(s)$ for positive s approaching zero:

$$\begin{aligned}
G(s) &= (1 + s)^2[1 - 2(e^s - 1) + 4(e^s - 1)^2 - 8(e^s - 1)^3 + O(s^4)] \\
&= (1 + s)^2[1 - 2(s + s^2/2 + s^3/6) + 4(s^2 + s^3) - 8s^3 + O(s^4)] \\
&= (1 + 2s + s^2)[1 - 2s + 3s^2 - 13s^3/2 + O(s^4)] \\
&= 1 - s^3/3 + O(s^4).
\end{aligned}$$

Use the symmetry of G to deduce $G(s) - G(0) = -|s|^3/3 + O(s^4)$. Conclude that

$$V(s) - V(0) = |s|^3/3 + O(s^4), \tag{2.11}$$

as s goes to zero.

Now I can derive the approximation to the population criterion function W, which is parametrized by the centers. Note that $(-1, 1)$ is the set of optimal population centers. Indeed, -1 and 1 are conditional means of the corresponding half-planes defined by the optimal split at zero. Denote by $h = (h_1, h_2)'$ the vector of increments to the set of optimal population centers. Define $\widetilde{h}$ to be $(h_s, h_d)'$, where

$$h_s = (h_1 + h_2)/2, \qquad h_d = (h_1 - h_2)/2.$$

Note that $\widetilde{h}$ is a linear function of h and vice versa.

Let c be $(-1, 1)$, the vector of optimal population centers, and let $\widetilde{c}$ stand for the vector of new centers $c + h$. Note that $\widetilde{c}$ splits the real line at the point h_s. Use equality (2.3) to split the criterion function W into a sum of the within cluster sum of squares part and the bias squared part:

$$W(\widetilde{c}) - W(c) = V(h_s) - V(0) + B^2(h_s, \widetilde{c}). \tag{2.12}$$

Apply Taylor expansions to rewrite expressions (2.10) for t near zero:

$$\begin{aligned}
\pi_{t+} &= \{t > 0\}(1/2 - t/2 + t^2/4) + \{t \le 0\}(1/2 - t/2 - t^2/4) + O(t^3) \\
\pi_{t-} &= \{t > 0\}(1/2 + t/2 - t^2/4) + \{t \le 0\}(1/2 + t/2 + t^2/4) + O(t^3) \\
\overline{X}_{t+} &= \{t > 0\}(1 + t) + \{t \le 0\}(1 + t + t^2) + O(t^3) \\
\overline{X}_{t-} &= \{t > 0\}(-1 + t - t^2) + \{t \le 0\}(-1 + t) + O(t^3).
\end{aligned} \tag{2.13}$$

The expressions for π_{t+} and π_{t+} will only be used in the next section. Derive the approximations for the squared distances from the elements of $\widetilde{c}$ to the corresponding conditional

means defined by the split at h_s, as h goes to zero:

$$(-1+h_1-\overline{X}_{h_s-})^2=(h_s-h_1-h_s^2\{h_s>0\}+O(|h|^3))^2=h_d^2+2h_s^2h_d\{h_s>0\}+O(|h|^4)$$
$$(1+h_2-\overline{X}_{h_s+})^2=(h_s-h_2+h_s^2\{h_s<0\}+O(|h|^3))^2=h_d^2+2h_s^2h_d\{h_s<0\}+O(|h|^4).$$

Plug the above approximations into expression (2.2), which defines function B^2, and use $\pi_{h_s-}=1/2+O(|h|)$ to get

$$\begin{aligned} B^2(h_s,\widetilde{c}) &= h_d^2+2\pi_{h_s-}h_s^2h_d\{h_s>0\}+2\pi_{h_st+}h_s^2h_d\{h_s<0\}+O(|h|^4) \\ &= h_d^2+h_s^2h_d+O(|h|^4). \end{aligned} \tag{2.14}$$

Combine equality (2.12) with approximations (2.11) and (2.14) to derive the approximation to population criterion function W at its minimum as h tends to zero:

$$W(-1+h_1,1+h_2)-W(-1,1)=h_d^2+|h_s|^3/3+h_s^2h_d+O(|h|^4). \tag{2.15}$$

2.2.2 Asymptotics: Parametrization by the Split Point

Combine approximations (2.9) and (2.11) to derive

$$V_n(s_n)-V_n(0)=|s_n|^3/3+O_p(n^{-1/2}|s_n|)+o_p(|s_n|^3)+o_p(n^{-1}).$$

Note that the left hand side is non-positive and deduce $|s_n|^3=O_p(n^{-1/2}|s_n|)+o_p(n^{-1})$. Fix a positive C. The above equality implies $|s_n|^2\{|s_n|>Cn^{-1/4}\}=O_p(n^{-1/2})$, which guarantees

$$|s_n|=O_p(n^{-1/4}).$$

Fix a random sequence of $O_p(n^{-1/4})$ neighborhoods $\mathcal{G}_n$, such that $s_n \in \mathcal{G}_n$ holds with probability tending to one. Approximations (2.9) and (2.11) yield

$$V_n(s) - V_n(0) = |s|^3/3 - n^{-1/2}s(Z_{1n} + Z_{2n}) + o_p(n^{-3/4}), \tag{2.16}$$

uniformly over $\mathcal{G}_n$.

Let $t = n^{1/4}s$ and $\mathcal{G}'_n = n^{1/4}\mathcal{G}_n$. The above approximation becomes

$$V_n(s) - V_n(0) = n^{-3/4}\left(|t|^3/3 - t(Z_{1n} + Z_{2n}) + o_p(1)\right), \tag{2.17}$$

uniformly over $\mathcal{G}'_n$. Let t^* minimize the function $f(t) = |t|^3/3 - t(Z_{1n} + Z_{2n})$. Suppose $Z_{1n} + Z_{2n} \geq 0$. Note that in this case $f(|t|) \leq f(-|t|)$ for all t, and the equality is achieved only if either t or $Z_{1n} + Z_{2n}$ is zero. Hence, t^* is nonnegative. Analyze $f(t)$ for nonnegative t to deduce $t^* = \sqrt{Z_{1n} + Z_{2n}}$. Analogously, $t^* = \sqrt{-Z_{1n} - Z_{2n}}$ when $Z_{1n} + Z_{2n} \leq 0$. Conclude that

$$t^* = \mathcal{S}(Z_{1n} + Z_{2n})\sqrt{|Z_{1n} + Z_{2n}|},$$

where $\mathcal{S}(t)$ is the sign function $\mathcal{S}(t) = \{t > 0\} - \{t < 0\}$.

Suppose $Z_{1n} + Z_{2n} \geq 0$. Observe that when t is nonnegative,

$$f(t) - f(t^*) \geq (t - t^*)^2\left(t + (t - t^*)/3\right),$$

and the right hand side is bounded below by $\frac{2}{3}(t - t^*)^2\sqrt{Z_{1n} + Z_{2n}}$. Note that when t is negative $f(t)$ is positive, and hence the difference $f(t) - f(t^*)$ is bounded below by $\frac{2}{3}(Z_{1n} + Z_{2n})^{3/2}$.

Argue analogously for the case $Z_{1n} + Z_{2n} \leq 0$, and conclude that

$$f(t) - f(t^*) \geq \tfrac{2}{3}|Z_{1n} + Z_{2n}|^{1/2} \min\left((t - t^*)^2, |Z_{1n} + Z_{2n}|\right), \tag{2.18}$$

for all t and all sample points.

Let t_n stand for $n^{1/4}s_n$. Note that t_n minimizes the criterion function C_n, which is defined by $n^{-3/4}C_n(t) = V_n(n^{-1/4}t) - V_n(0)$. Write approximation (2.17) in the form

$$C_n(t) = f(t) + o_p(1).$$

Compare $C_n(t_n)$ to $C_n(t^*)$ to deduce that $f(t_n) - f(t^*)$ is of order $o_p(1)$. Apply inequality (2.18) to derive

$$|Z_{1n} + Z_{2n}|^{1/2} \min\left((t - t^*)^2, |Z_{1n} + Z_{2n}|\right) = o_p(1).$$

The random quantity $|Z_{1n} + Z_{2n}|$ is positive with probability tending to one. Deduce that $|t - t^*|$ is of order $o_p(1)$. Go back to the original parametrization to conclude

$$s_n = n^{-1/4}\mathcal{S}(Z_{1n} + Z_{2n})\sqrt{|Z_{1n} + Z_{2n}|} + o_p(n^{-1/4}).$$

Apply the central limit theorem to (Z_{1n}, Z_{2n}) to deduce the weak convergence of s_n,

$$n^{1/4}s_n \rightsquigarrow \mathcal{S}(Z)\sqrt{|Z|},$$

where Z has an $N(0, 4)$ distribution. This convergence result was also proved by Serinko and Babu (1992) using techniques that apply only in one dimension.

2.2.3 Asymptotics: Parametrization by the Centers

Use approximation (2.15) to bound the population criterion function below by a cubic. It follows from Lemma 5 in Section 2.3 that there exists a positive c_0, such that for all h small enough,

$$W(c+h) - W(c) > c_0(h_d^2 + |h_s|^3). \tag{2.19}$$

Let $\widehat{h}_n = (\widehat{h}_{1n}, \widehat{h}_{2n})$ minimize the function $W_n(-1+h_1, 1+h_2)$. To simplify the notation I will omit the subscript n for $\widehat{h}_n$. Define $\widehat{h}_s$ and $\widehat{h}_d$ analogously to h_s and h_d. Recall that $c = (-1, 1)$ is the vector of optimal population centers. Apply approximation (1.5) to write

$$W_n(c+\widehat{h}) - W_n(c) = W(c+\widehat{h}) - W(c) + O_p(n^{-1/2}|\widehat{h}|).$$

Apply inequality $W_n(c+\widehat{h}) \leq W_n(c)$ together with inequality (2.19) and consistency of $c+\widehat{h}$ to derive

$$\begin{aligned} |\widehat{h}|^3 &= O_p(n^{-1/2}|\widehat{h}|) \\ \widehat{h}_d^2 + |\widehat{h}_s|^3 &= O_p(n^{-1/2}|\widehat{h}|). \end{aligned}$$

Use the first of the equalities above to derive $\widehat{h} = O_p(n^{-1/4})$, which implies $\widehat{h}_s = O_p(n^{-1/4})$. Now use the second equality to deduce $\widehat{h}_d = O_p(n^{-3/8})$.

Recall that there is a unique linear correspondence between (h_s, h_d) and (h_1, h_2). Let $C_n(h_s, h_d)$ stand for $W_n(c+h) - W_n(c)$, the empirical criterion function evaluated at $c+h$. Note that C_n is minimized by $(\widehat{h}_s, \widehat{h}_d)$. Fix a sequence of random neighborhoods $\mathcal{G}_n$ that satisfy

$$\sup_{h \in \mathcal{G}_n} |h_s| = O_p(n^{-1/4}) \quad \text{and} \quad \sup_{h \in \mathcal{G}_n} |h_d| = O_p(n^{-3/8}),$$

such that $\widehat{h} \in \mathcal{G}_n$ holds with probability tending to one. The following approximation fol-

lows from Lemma 6 in Section 2.3. There exist a sequence of random univariate functions a_n, such that uniformly in $\mathcal{G}_n$,

$$C_n(h_s, h_d) = h_d^2 + O_p(n^{-1/2}|h_d|) + a_n(h_s) + o_p(n^{-1}).$$

Use this approximation to compare $C_n(\widehat{h}_s, \widehat{h}_d)$ with $C_n(\widehat{h}_s, 0)$. Derive

$$\widehat{h}_d^2 = O_p(n^{-1/2}|\widehat{h}_d|) + o_p(n^{-1}).$$

Fix a positive C. The above equality implies $|\widehat{h}_d|\{|\widehat{h}_d| > Cn^{-1/2}\} = O_p(n^{-1/2})$, which in turn forces

$$|\widehat{h}_d| = O_p(n^{-1/2}).$$

Fix a sequence of random neighborhoods $\mathcal{K}_n$ that satisfy

$$\sup_{h \in \mathcal{K}_n} |h_s| = O_p(n^{-1/4}) \quad \text{and} \quad \sup_{h \in \mathcal{K}_n} |h_d| = O_p(n^{-1/2}),$$

such that $\widehat{h}$ falls in $\mathcal{K}_n$ with probability tending to one. Combine approximations (1.5) and (2.15) to deduce

$$W_n(c + h) - W_n(c) = |h_s|^3/3 - n^{-1/2}h_s(Z_{1n} + Z_{2n}) + o_p(n^{-3/4}).$$

Observe that once h_s is substituted with s, the above approximation becomes exactly the same as approximation (2.16) of the sample within sum of squares function V_n. This is not surprising since h_s is the split point defined by the vector $c + h$ of centers. Repeat the argument in Subsection 2.2.2 to derive

$$\widehat{h}_s = n^{-1/4}\mathcal{S}(Z_{1n} + Z_{2n})\sqrt{|Z_{1n} + Z_{2n}|} + o_p(n^{-1/4}).$$

It follows from Lemma 6 in Section 2.3 that

$$C_n(h_s, h_d) = h_d^2 + h_s^2 h_d - n^{-1/2} h_d (Z_{1n} - Z_{2n}) + a_n(h_s) + o_p(n^{-1}),$$

uniformly in $\mathcal{K}_n$. Let S_n stand for $Z_{1n} - Z_{2n} - |Z_{1n} + Z_{2n}|$. Observe that

$$C_n(\widehat{h}_s, h_d) = h_d^2 - n^{-1/2} h_d S_n + a_n(\widehat{h}_s) + o_p(n^{-1}).$$

Rewrite the above approximation by completing the square:

$$C_n(\widehat{h}_s, h_d) = (h_d^2 - n^{-1/2} S_n/2)^2 - n^{-1} S_n^2/4 + a_n(\widehat{h}_s) + o_p(n^{-1}). \tag{2.20}$$

Denote $n^{-1/2} S_n/2$ by h_d^*. Because S_n is of order $O_p(1)$, neighborhoods $\mathcal{K}_n$ can be assumed, without loss of generality, to contain vectors that correspond to the pairs $(\widehat{h}_s, h_d^*)$. Use approximation (2.20) to compare $C_n(\widehat{h}_s, \widehat{h}_d)$ with $C_n(\widehat{h}_s, \widehat{h}_d^*)$ and conclude that $\widehat{h}_d = \widehat{h}_d^* + o_p(n^{-1/2})$. Thus,

$$\begin{aligned} \widehat{h}_s &= n^{-1/4} \mathcal{S}(Z_{1n} + Z_{2n}) \sqrt{|Z_{1n} + Z_{2n}|} + o_p(n^{-1/4}) \\ \widehat{h}_d &= n^{-1/2} (Z_{1n} - Z_{2n} - |Z_{1n} + Z_{2n}|)/2 + o_p(n^{-1/2}), \end{aligned}$$

where

$$(Z_{1n}, Z_{2n}) = 2 \left(\nu_n^x (x+1)\{x \le 0\}, \;\; \nu_n^x (x-1)\{x > 0\} \right) \rightsquigarrow N(0, 2I).$$

Recall that W^* stands for the minimum value of population criterion function W. Combine the above approximations to $(\widehat{h}_s, \widehat{h}_d)$ with approximation (2.15) to the function W and

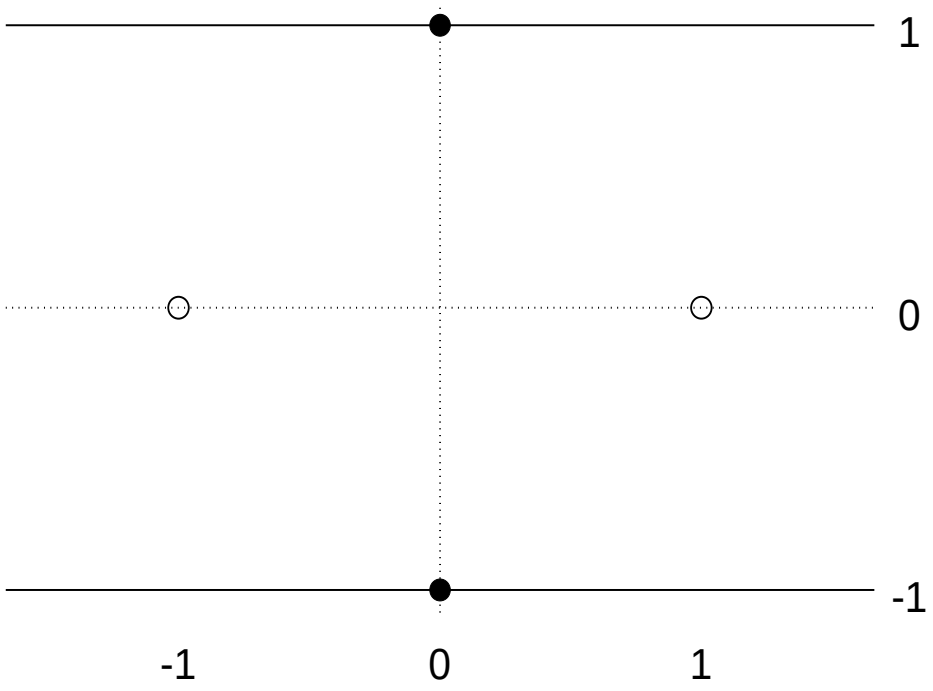

Figure 2.2: Two optimal configurations of centers

deduce the following approximation to the distortion redundancy:

$$W(-1+\widehat{h}_1, 1+\widehat{h}_2) - W^* = n^{-3/4}|Z_{1n} + Z_{2n}|^{3/2} + o_p(n^{-3/4}).$$

2.3 A Two-Dimensional Example

Let Q be a distribution on the plane that concentrates on two parallel lines with distance 2 apart. Let Q put probability one half on each line, and let the conditional distribution on each line be double exponential. For convenience, introduce an (x, y) coordinate system

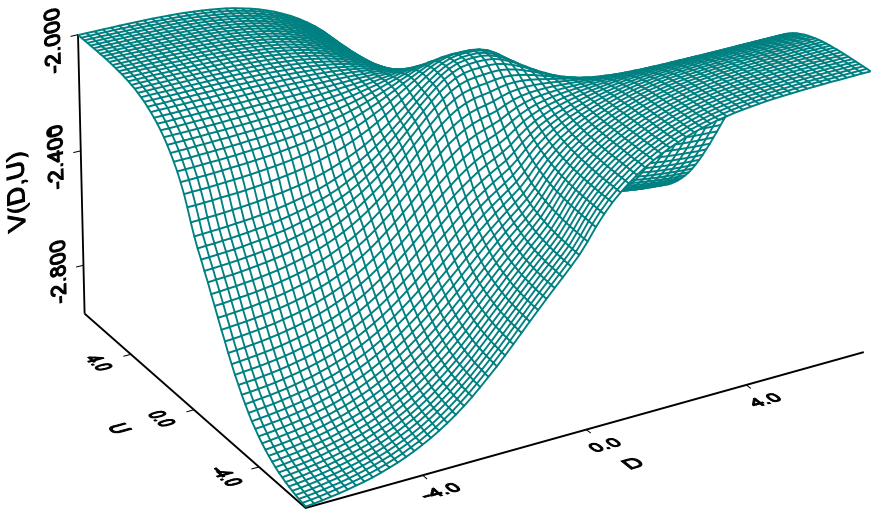

Figure 2.3: Within sum of squares as a function of the split line

on the plane. Population distribution Q is a product measure on the plane (x, y):

$$Q^{x,y} = P^x \times \mu^y,$$

where $\mu\{-1\} = \mu\{1\} = 1/2$ and P is the usual double exponential distribution.

There are two pairs of optimal population centers (see figure 2.2). The optimal pairs are $\{(-1, 0), (1, 0)\}$ and $\{(0, -1), (0, 1)\}$; denote them by $c^V = (c_1^V, c_2^V)'$ and $c^H = (c_1^H, c_2^H)'$ respectively, signifying that the corresponding split line is either vertical or horizontal. Figure 2.3 illustrates the dependence of the population within sum of squares on the split line, which is specified by the coordinates D and U that correspond to the two points of intersection of the split line with the lines $y = -1$ and $y = 1$, respectively. The negative

of the criterion function is actually plotted to enhance the visual presentation.

2.3.1 Population Criterion Function: Horizontal Split

Temporarily omit the superscript for the set c^H of optimal population centers that split the plane by a horizontal line. Think of c as a four-dimensional vector $(0, -1, 0, 1)$. Consider a vector of new centers $\widetilde{c} = c + h$ in a small neighborhood of c. Suppose the points of intersection of the new split line with the lines $y = -1$ and $y = 1$ are D and U. The cut points D and U are very sensitive to small changes in h. In fact there exists a positive constant K such that

$$|D| \wedge |U| > K\frac{1}{|h|}.$$

For any partition $\mathcal{P} = A \cup A^c$ of the plane and any vector $b = (b_1, b_2)$ in $\mathbb{R}^2 \times \mathbb{R}^2$ define

$$\mathcal{W}(\mathcal{P}, b) = Q^z\{z \in A\}|z - b_1|^2 + \{z \in A^c\}|z - b_2|^2,$$

the generalized within sum of squares (compare with the one-dimensional definition). Define the biased squared function analogously. Let $\mathcal{P}_{\widetilde{c}}$ stand for the partition of the plane that $\widetilde{c}$ defines. Observe that

$$\begin{aligned} W(\widetilde{c}) - W(c) &= \mathcal{W}(\mathcal{P}_{\widetilde{c}}, \widetilde{c}) - \mathcal{W}(\mathcal{P}_c, c) \\ &= \left[\mathcal{W}(\mathcal{P}_{\widetilde{c}}, \widetilde{c}) - \mathcal{W}(\mathcal{P}_c, \widetilde{c})\right] + \left[\mathcal{W}(\mathcal{P}_c, \widetilde{c}) - \mathcal{W}(\mathcal{P}_c, c)\right]. \end{aligned}$$

The second difference is just

$$B^2(\mathcal{P}_c, \widetilde{c}) = |h|^2/2.$$

Bound the first difference by

$$P^x\{x > |D| \wedge |U|\}(2|x|^2 + 4) < P^x\{x > K/|h|\}(|x|^2 + 2).$$

The last probability goes to zero exponentially fast, hence it can safely be considered $O(|h|^3)$. Conclude that

$$W(c^H + h) - W(c^H) = |h|^2/2 + O(|h|^3). \tag{2.21}$$

2.3.2 Population Criterion Function: Vertical Split

Think of c^V as a four-dimensional vector $(-1, 0, 1, 0)$. Denote by $h = (\delta_1, \epsilon_1, \delta_2, \epsilon_2)'$ the 4-dimensional vector of increments to c^V. Let $\widetilde{c}$ stand for $c^V + h$. Denote the pair of x-coordinates of $\widetilde{c}$ by $\widetilde{c}_x$ and the pair of y-coordinates by $\widetilde{c}_y$. Denote the intersection points of the new split line with the lines $y = -1$ and $y = 1$ by D and U respectively. Note that these points depend on h. Denote the sets $\{D \geq 0\}$ and $\{U \geq 0\}$ by $\mathcal{D}_+$ and $\mathcal{U}_+$ respectively. Denote the closures of the corresponding complements by $\mathcal{D}_-$ and $\mathcal{U}_-$. Because the squared distance between any two points in the plane can be split into the sum of the x-distance squared and the y-distance squared, the difference $W(\widetilde{c}) - W(c^V)$ can be split into two parts, the x and the y contributions. The x-contribution is

$$\left[\mathcal{W}(D, \widetilde{c}_x) + \mathcal{W}(U, \widetilde{c}_x)\right]/2 - V(0).$$

Subtract and add $(V(D) + V(U)/2$ to rewrite the x-contribution as

$$\left[B^2(D, \widetilde{c}_x) + B^2(U, \widetilde{c}_x) + H(D) + H(U)\right]/2. \tag{2.22}$$

Let $\widetilde{h}$ stand for $(\delta_s, \delta_d, \epsilon_s, \epsilon_d)'$, where

$$\delta_s = (\delta_1 + d_2)/2, \quad \delta_d = (\delta_1 - \delta_2)/2, \quad \epsilon_s = (\epsilon_1 + \epsilon_2)/2, \quad \epsilon_d = (\epsilon_1 - \epsilon_2)/2.$$

Note that $\widetilde{h}$ is a linear function of h and vice versa. For $s = O(|h|)$ use approximations (2.13) to π_{s-} and $\overline{X}_{s-}$ to derive

$$\begin{aligned} B^2(s, \widetilde{c}_x) &= \pi_{s-}(\overline{X}_{s-} + 1 - \delta_1)^2 + \pi_{s+}(\overline{X}_{s+} - 1 - \delta_2)^2 \\ &= \Big[\{s > 0\}(1/2 + s/2) + \{s \le 0\}(1/2 + s/2)\Big] \\ &\quad \times \Big[\{s > 0\}(-1 + s - s^2) + \{s \le 0\}(-1 + s) + 1 - \delta_1\Big]^2 \\ &\quad + \Big[\{s > 0\}(1/2 - s/2) + \{s \le 0\}(1/2 - s/2)\Big] \\ &\quad \times \Big[\{s > 0\}(1 + s) + \{s \le 0\}(1 + s + s^2) - 1 - \delta_2\Big]^2 + O(|h|^4). \end{aligned}$$

Collect the $\{s > 0\}$ terms:

$$(1/2 + s/2)(s - \delta_1 - s^2)^2 + (1/2 - s/2)(s - \delta_2)^2 + O(|h|^4)$$

and the $\{s \le 0\}$ terms:

$$(1/2 + s/2)(s - \delta_1)^2 + (1/2 - s/2)(s - \delta_2 + s^2)^2 + O(|h|^4).$$

Note that the difference between the $\{s \le 0\}$ and the $\{s > 0\}$ terms is

$$(2s - \delta_1 - \delta_2)s^2 + O(|h|^4). \tag{2.23}$$

Rewrite the $\{s > 0\}$ terms as

$$(s-\delta_1)^2/2 - (s-\delta_1)s^2 + s(s-\delta_1)^2/2 + (s-\delta_2)^2/2 - s(s-\delta_2)^2/2 + O(|h|^4).$$

Use approximations

$$\begin{aligned} U &= \delta_s + \epsilon_d + \delta_d\epsilon_d - \epsilon_s\epsilon_d + O(|h|^3) \\ D &= \delta_s - \epsilon_d - \delta_d\epsilon_d - \epsilon_s\epsilon_d + O(|h|^3) \end{aligned} \tag{2.24}$$

to derive that on $\mathcal{D}_+\mathcal{U}_+$,

$$\begin{aligned} B^2(D,\widetilde{c}_x) + B^2(U,\widetilde{c}_x) = \\ &(\delta_d + \epsilon_d + \epsilon_d\delta_d + \epsilon_s\epsilon_d)^2/2 + (\delta_d+\epsilon_d)(\delta_s-\epsilon_d)^2 + (\delta_s-\epsilon_d)(\delta_d+\epsilon_d)^2/2 \\ +&(\delta_d - \epsilon_d - \epsilon_d\delta_d - \epsilon_s\epsilon_d)^2/2 - (\delta_s-\epsilon_d)(\delta_d-\epsilon_d)^2/2 \\ +&(\epsilon_d - \delta_d + \epsilon_d\delta_d - \epsilon_s\epsilon_d)^2/2 - (\epsilon_d-\delta_d)(\delta_s+\epsilon_d)^2 + (\delta_s+\epsilon_d)(\epsilon_d-\delta_d)^2/2 \\ +&(\delta_d + \epsilon_d + \epsilon_d\delta_d - \epsilon_s\epsilon_d)^2/2 - (\delta_s+\epsilon_d)(\delta_d+\epsilon_d)^2/2 + O(|h|^4). \end{aligned}$$

This expression simplifies to

$$2(\delta_d^2 + \epsilon_d^2 + \delta_s^2\delta_d + \delta_d\epsilon_d^2) - 4\delta_s\epsilon_d^2 + O(|h|^4). \tag{2.25}$$

The remaining terms of the x-contribution to $W(\widetilde{c}) - W(c^V)$ on $\mathcal{D}_+\mathcal{U}_+$ are

$$\begin{aligned} H(D) + H(U) &= (D^3 + U^3)/3 \\ &= \left((\delta_s-\epsilon_d)^3 + (\delta_s+\epsilon_d)^3\right)/3 + O(|h|^4) \\ &= 2\delta_s^3/3 + 2\delta_s\epsilon_d^2 + O(|h|^4). \end{aligned} \tag{2.26}$$

Combine approximations (2.25) and (2.26) using (2.22) to derive the x-contribution to $W(\widetilde{c}) - W(c^V)$ on $\mathcal{D}_+\mathcal{U}_+$:

$$\delta_d^2 + \epsilon_d^2 + \delta_s^3/3 + \delta_s^2\delta_d - \delta_s\epsilon_d^2 + \delta_d\epsilon_d^2 + O(|h|^4). \tag{2.27}$$

The y-contribution to $W(\widetilde{c}) - W(c^V)$ is

$$\begin{aligned}&\left[(1-\epsilon_1)^2\pi_{U-} + (1+\epsilon_1)^2\pi_{D-} + (1-\epsilon_2)^2\pi_{U+} + (1+\epsilon_2)^2\pi_{D+}\right]/2 - 1\\ &\quad = \epsilon_1^2(\pi_{U-} + \pi_{D-})/2 + \epsilon_2^2(\pi_{U+} + \pi_{D+})/2 + \epsilon_1(\pi_{D-} - \pi_{U-}) + \epsilon_2(\pi_{D+} - \pi_{U+}).\end{aligned}$$

On $\mathcal{D}_+\mathcal{U}_+$ the y-contribution becomes

$$\begin{aligned}&\epsilon_1^2(1 + U/2 + D/2)/2 + \epsilon_2^2(1 - U/2 - D/2)/2\\ &+\epsilon_1(D/2 - D^2/4 - U/2 + U^2/4) + \epsilon_2(U/2 - U^2/4 - D/2 + D^2/4) + O(|h|^4).\end{aligned}$$

Use approximation (2.24) to write it as

$$\epsilon_1^2(1+\delta_s)/2 + \epsilon_2^2(1-\delta_s)/2 + \epsilon_1(-\epsilon_d - \delta_d\epsilon_d + \delta_s\epsilon_d) + \epsilon_2(\epsilon_d + \delta_d\epsilon_d - \delta_s\epsilon_d) + O(|h|^4).$$

Simplify this expression to derive the y-contribution to $W(\widetilde{c}) - W(c^V)$ on $\mathcal{D}_+\mathcal{U}_+$:

$$\epsilon_s^2 - \epsilon_d^2 + 2\delta_s\epsilon_s\epsilon_d + 2\delta_s\epsilon_d^2 - 2\delta_d\epsilon_d^2 + O(|h|^4). \tag{2.28}$$

Combine approximations (2.27) and (2.28) to conclude that on $\mathcal{D}_+\mathcal{U}_+$,

$$W(c^V + h) - W(c^V) = \delta_d^2 + \epsilon_s^2 + \delta_s^3/3 + \delta_s^2\delta_d + 2\delta_s\epsilon_s\epsilon_d + \delta_s\epsilon_d^2 - \delta_d\epsilon_d^2 + O(|h|^4). \tag{2.29}$$

Use approximations (2.23) and (2.25) to derive that on $\mathcal{D}_+\mathcal{U}_-$,

$$B^2(D,\widetilde{c}_x) + B^2(U,\widetilde{c}_x) = 2(\delta_d^2 + \epsilon_d^2 + \delta_s^2\delta_d + \delta_d\epsilon_d^2) - 4\delta_s\epsilon_d^2 + (2U - \delta_1 - \delta_2)U^2 + O(|h|^4).$$

Apply expansion (2.24) to rewrite this expression as

$$2(\delta_d^2 + \epsilon_d^2 + \delta_s^2\delta_d + \delta_d\epsilon_d^2) + 2\delta_s^2\epsilon_d + 2\epsilon_d^3 + O(|h|^4). \tag{2.30}$$

The remaining terms of the x-contribution to $W(\widetilde{c}) - W(c^V)$ on $\mathcal{D}_+\mathcal{U}_-$ are

$$\begin{aligned} H(D) + H(U) &= (D^3 - U^3)/3 \\ &= \left((\delta_s - \epsilon_d)^3 - (\delta_s + \epsilon_d)^3\right)/3 + O(|h|^4) \\ &= -2\delta_s^2\epsilon_d - 2\epsilon_d^3/3 + O(|h|^4). \end{aligned} \tag{2.31}$$

Combine approximations (2.30) and (2.31) using (2.22) to derive the x-contribution to $W(\widetilde{c}) - W(c^V)$ on $\mathcal{D}_+\mathcal{U}_-$:

$$\delta_d^2 + \epsilon_d^2 + \delta_s^2\delta_d + \delta_d\epsilon_d^2 + 2\epsilon_d^3/3 + O(|h|^4). \tag{2.32}$$

The y-contribution on $\mathcal{D}_+\mathcal{U}_-$ is

$$\begin{aligned} &\epsilon_1^2(1 + U/2 + D/2)/2 + \epsilon_2^2(1 - U/2 - D/2)/2 \\ &+\epsilon_1(D/2 - D^2/4 - U/2 - U^2/4) + \epsilon_2(U/2 + U^2/4 - D/2 + D^2/4) + O(|h|^4). \end{aligned}$$

Note that this is the same expression as the one for the y-contribution on $\mathcal{D}_+\mathcal{U}_+$ minus a $(U^2\epsilon_d)$ term, which can be written as

$$\delta_s^2\epsilon_d + 2\delta_s\epsilon_d^2 + \epsilon_d^3 + O(|h|^4),$$

using expansion (2.24). It follows from approximation (2.28) that the y-contribution to $W(\widetilde{c}) - W(c^V)$ on $\mathcal{D}_+\mathcal{U}_-$ is

$$\epsilon_s^2 - \epsilon_d^2 - \delta_s^2\epsilon_d + 2\delta_s\epsilon_s\epsilon_d - 2\delta_d\epsilon_d^2 - \epsilon_d^3 + O(|h|^4). \tag{2.33}$$

Combine approximations (2.32) and (2.33) to conclude that on $\mathcal{D}_+\mathcal{U}_-$,

$$W(c^V + h) - W(c^V) = \delta_d^2 + \epsilon_s^2 + \delta_s^2\delta_d - \delta_s^2\epsilon_d + 2\delta_s\epsilon_s\epsilon_d - \delta_d\epsilon_d^2 - \epsilon_d^3/3 + O(|h|^4). \tag{2.34}$$

Function W is invariant to reflection about the line $x = 0$ by the symmetry of the distribution $Q^{x,y}$. This reflection is given by

$$\begin{cases} \tilde{\delta}_1 = -\delta_2 \\ \tilde{\epsilon}_1 = \quad \epsilon_2 \\ \tilde{\delta}_2 = -\delta_1 \\ \tilde{\epsilon}_2 = \quad \epsilon_1 \ , \end{cases}$$

which corresponds to

$$\begin{cases} \tilde{\delta}_s = -\delta_s \\ \tilde{\delta}_d = \quad \delta_d \\ \tilde{\epsilon}_s = \quad \epsilon_s \\ \tilde{\epsilon}_d = -\epsilon_d \ . \end{cases} \tag{2.35}$$

Note that if a set of centers lies in $\mathcal{D}_-\mathcal{U}_-$ then its reflection about $x = 0$ lies in $\mathcal{D}_+\mathcal{U}_+$. Apply the change of variables given by (2.35) to the expression in (2.29) and deduce that on $\mathcal{D}_-\mathcal{U}_-$

$$W(c^V + h) - W(c^V) = \delta_d^2 + \epsilon_s^2 - \delta_s^3/3 + \delta_s^2\delta_d + 2\delta_s\epsilon_s\epsilon_d - \delta_s\epsilon_d^2 - \delta_d\epsilon_d^2 + O(|h|^4). \tag{2.36}$$

If a set of centers lies in $\mathcal{D}_-\mathcal{U}_+$ then its reflection about $x = 0$ lies in $\mathcal{D}_+\mathcal{U}_-$. Use (2.34) to deduce that on $\mathcal{D}_-\mathcal{U}_+$

$$W(c^V + h) - W(c^V) = \delta_d^2 + \epsilon_s^2 + \delta_s^2\delta_d + \delta_s^2\epsilon_d + 2\delta_s\epsilon_s\epsilon_d - \delta_d\epsilon_d^2 + \epsilon_d^3/3 + O(|h|^4). \quad (2.37)$$

Observe that δ_s is non-negative on $\mathcal{D}_+\mathcal{U}_+$ and non-positive on $\mathcal{D}_-\mathcal{U}_-$. Also note that ϵ_d is non-negative on $\mathcal{D}_-\mathcal{U}_+$ and non-positive on $\mathcal{D}_+\mathcal{U}_-$. Combine (2.29) with (2.36), and (2.34) with (2.37), to conclude that for small h,

$$W(c^V + h) - W(c^V) =$$
$$\begin{cases} \delta_d^2 + \epsilon_s^2 + |\delta_s|^3/3 + \delta_s^2\delta_d + 2\delta_s\epsilon_s\epsilon_d + |\delta_s|\epsilon_d^2 - \delta_d\epsilon_d^2 + O(|h|^4), \text{ on } \mathcal{D}_+\mathcal{U}_+ \cup \mathcal{D}_-\mathcal{U}_- \\ \delta_d^2 + \epsilon_s^2 + \delta_s^2\delta_d + \delta_s^2|\epsilon_d| + 2\delta_s\epsilon_s\epsilon_d - \delta_d\epsilon_d^2 + |\epsilon_d|^3/3 + O(|h|^4), \text{ on } \mathcal{D}_+\mathcal{U}_- \cup \mathcal{D}_-\mathcal{U}_+. \end{cases}$$

Note that the union in the top line of this expression can be safely replaced by the set $\{|d_s| > |\epsilon_d|\}$, and the union in the bottom line by the set $\{|d_s| \leq |\epsilon_d|\}$. Indeed, the difference between the terms in the top line and the terms in the bottom line is $(|\delta_s| - |\epsilon_d|)^3/3 + O(|h|^4)$, which on the set

$$\Big[(\mathcal{D}_+\mathcal{U}_+ \cup \mathcal{D}_-\mathcal{U}_-) \triangle \{|d_s| > |\epsilon_d|\}\Big] \cup \Big[(\mathcal{D}_+\mathcal{U}_- \cup \mathcal{D}_-\mathcal{U}_+) \triangle \{|d_s| \leq |\epsilon_d|\}\Big]$$

is $O(|h|^4)$ by an application of formula (2.24). The approximation becomes

$$W(c^V+h)-W(c^V) = \delta_d^2+\epsilon_s^2+\delta_s^2\delta_d+2\delta_s\epsilon_s\epsilon_d-\delta_d\epsilon_d^2+\tfrac{1}{6}((|\delta_s|+|\epsilon_d|)^3+||\delta_s|-|\epsilon_d||^3)+O(|h|^4). \quad (2.38)$$

2.3.3 Asymptotics: Horizontal Split

Let b_n be the set of optimal sample centers. Consistency results (for example, Proposition 1 in Chapter 3) force b_n to converge to the pair of sets c^H, c^V with respect to the Hausdorff metric. There exists a sequence of positive numbers r_n going to zero, such that with probability tending to one, b_n is within r_n of either c^H or c^V in Hausdorff metric. Let b_n^H and b_n^V minimize W_n over the Hausdorff balls of radius r_n around c^H and c^V respectively. Then with probability tending to one,

$$b_n = \underset{b \in \{b_n^H, b_n^V\}}{\operatorname{argmin}} W_n(b).$$

Function W is approximated by a nonsingular quadratic in (2.21), thus b_n^H converges to c^H at the usual $n^{-1/2}$ rate:

$$n^{1/2}(b_n^H - c^H) = Z_n^H + o_p(1), \tag{2.39}$$

where

$$Z_n^H = 2\nu_n^{x,y} x \begin{pmatrix} \{y \le 0\} \\ 0 \\ \{y > 0\} \\ 0 \end{pmatrix}.$$

In fact, the y-coordinates of the centers in the set c_n^H converge to the y-coordinates of the centers in the set c^H exponentially fast.

2.3.4 Rates of Convergence for the Vertical Split

Asymptotic behavior of b_n^V is less trivial because, as seen from approximation (2.38), a nonsingular quadratic lower bound of the population criterion function can not be established. Instead, the following bound holds.

Lemma 5 *There exists a positive c_0, such that for all h small enough,*

$$W(c^V + h) - W(c^V) > c_0(\delta_d^2 + \epsilon_s^2 + |\delta_s|^3 + |\epsilon_d|^3). \tag{2.40}$$

Proof: It suffices to show that the sum

$$\delta_d^2/4 + \epsilon_s^2/4 + |\delta_s|^3/7 + |\epsilon_d|^3/7 + \delta_s^2\delta_d - \delta_d\epsilon_d^2 + 2\delta_s\epsilon_s\epsilon_d \tag{2.41}$$

is nonnegative for h small enough. Use equalities

$$\begin{aligned} \delta_d^2/4 + \delta_s^2\delta_d - \delta_d\epsilon_d^2 &= (\delta_d/2 + \delta_s^2 - \epsilon_d^2)^2 - (\delta_s^2 - \epsilon_d^2)^2, \text{ and} \\ \epsilon_s^2/4 + 2\delta_s\epsilon_s\epsilon_d &= (\epsilon_s/2 + 2\delta_s\epsilon_d)^2 - 4\delta_s^2\epsilon_d^2, \end{aligned}$$

to bound the sum (2.41) below by

$$|\delta_s|^3/7 + |\epsilon_d|^3/7 - (\delta_s^2 - \epsilon_d^2)^2 - 4\delta_s^2\epsilon_d^2.$$

The last expression is nonnegative for all h small enough. □

Search for the minimizer of W_n among the sets $\widetilde{c}$ that are within the Hausdorff distance r_n of the set c^V. As before, write $\widetilde{c} = c^V + h$. The 4-dimensional vector of increments h corresponds to the vector $\widetilde{h} = (\delta_s, \delta_d, \epsilon_s, \epsilon_d)'$. Each element of h and $\widetilde{h}$ is bounded above in absolute value by r_n. Denote by $\mathcal{N}_n$ the 4-dimensional square that is centered at zero and has side length $r_n/2$. Denote by $C_n(\widetilde{h})$ the difference $W_n(c^V + h) - W_n(c^V)$. Write it in the following convenient form:

$$C_n(\widetilde{h}) = W(c^V + h) - W(c^V) + n^{-1/2}\nu_n^z\left(\phi(z, c^V + h) - \phi(z, c^V)\right). \tag{2.42}$$

It follows from Pollard (1982a) that

$$\nu_n^z\left(\phi(z,c^V+h)-\phi(z,c^V)\right)=|h|K_n(h),$$

where $\sup_{\mathcal{N}_n} K_n(h) = O_p(r_n)$. I use stochastic order symbols rather than explicit functions of h and $\widetilde{h}$. Approximation (2.42) becomes

$$C_n(\widetilde{h}) = W(c^V+h) - W(c^V) + O_p(n^{-1/2}|h|), \tag{2.43}$$

uniformly in $\mathcal{N}_n$.

Suppose $\widehat{h}_n = (\widehat{\delta}_s, \widehat{\delta}_d, \widehat{\epsilon}_s, \widehat{\epsilon}_d)$ minimizes $C_n(\widetilde{h})$. Let n be large enough for the lower bound (2.40) to hold in $\mathcal{N}_n$. Use this lower bound and approximation (2.43) to compare $C_n(\widehat{h})$ with $C_n(0)$ and deduce that

$$c_0(\widehat{\delta}_d{}^2 + \widehat{\epsilon}_s{}^2 + |\widehat{\delta}_s|^3 + |\widehat{\epsilon}_d|^3) \le O_p(n^{-1/2}|\widehat{h}|),$$

uniformly in $\mathcal{N}_n$. Conclude $\widetilde{h} = O_p(n^{-1/4})$ and hence, applying the above inequality again,

$$(\widehat{\delta}_s, \widehat{\epsilon}_d) = O_p(n^{-1/4}) \quad \text{and} \quad (\widehat{\delta}_d, \widehat{\epsilon}_s) = O_p(n^{-3/8}).$$

Fix a sequence of random neighborhoods $\mathcal{G}_n$ that satisfy

$$\sup_{\widetilde{h}\in\mathcal{G}_n} |\delta_s| \vee |\epsilon_d| = O_p(n^{-1/4}) \text{ and } \sup_{\widetilde{h}\in\mathcal{G}_n} |\delta_d| \vee |\epsilon_s| = O_p(n^{-3/8}),$$

such that $\mathcal{G}_n \subseteq \mathcal{N}_n$ and $\widehat{h} \in \mathcal{G}_n$ with probability tending to one. Localize the search for $\widehat{h}$ and consider vectors $\widetilde{h}$ in $\mathcal{G}_n$.

Pollard (1982a) provides a more precise form of approximation (2.43):

$$C_n(\widetilde{h}) = W(c^V + h) - W(c^V) - n^{-1/2} h' Z_n + o_p(n^{-1/2}|h|), \tag{2.44}$$

where

$$Z_n = 2\nu_n^{x,y} \begin{pmatrix} \begin{pmatrix} x+1 \\ y \end{pmatrix} \{x \leq 0\} \\ \begin{pmatrix} x-1 \\ y \end{pmatrix} \{x > 0\} \end{pmatrix},$$

and the approximation is uniform in $\mathcal{N}_n$. The error of the approximation in $\mathcal{G}_n$ is $o_p(n^{-3/4})$, which is not good enough to derive asymptotics of (δ_d, ϵ_s). The following lemma refines the error terms in (2.44) for the specific problem at hand.

Lemma 6 *The random part of approximation (2.44) can be written as a sum of some random function of (δ_s, ϵ_d) and a $O_p(n^{-1/2}|(\delta_d, \epsilon_s)|)$ quantity. As a result, uniformly in $\mathcal{G}_n$,*

$$C_n(\widetilde{h}) = \delta_d^2 + \epsilon_s^2 + O_p(n^{-1/2}|(\delta_d, \epsilon_s)|) + a_n(\delta_s, \epsilon_d) + o_p(n^{-1}), \tag{2.45}$$

where $a_n(\delta_s, \epsilon_d)$ is a random function of δ_s and ϵ_d.

Proof: Write approximation (2.44) as

$$C_n(\widetilde{h}) = W(c^V + h) - W(c^V) - n^{-1/2} h' Z_n + n^{-1/2} \nu_n^z R(z, c^c, h). \tag{2.46}$$

Split the $n^{-1/2} h' Z_n$ term into a random function of (δ_s, ϵ_d) and the uniform $O_p(n^{-1/2}|(\delta_d, \epsilon_s)|)$ quantity.

The $O(|h|^4)$ term in the formula (2.38) for the increment of the function W may contain factors of δ_s^4 and ϵ_d^4. The rest of the $O(|h|^4)$ terms are $o(n^{-1})$ uniformly in $\mathcal{G}_n$. The

non-random contribution to the approximation (2.46) becomes

$$W(c^V + h) - W(c^V) = \delta_d^2 + \epsilon_s^2 + a_1(\delta_s, \epsilon_d) + o(n^{-1}),$$

uniformly in $\mathcal{G}_n$, where $a_1(\delta_s, \epsilon_d)$ is a deterministic function of just δ_s and ϵ_d.

Let '$(t_1, t_2]$' stand for the interval (or the indicator of the interval) $(t_1 \wedge t_2, t_1 \vee t_2]$. Denote the closed half-plane $\{x \leq 0\}$ by H_-, the open half-plane $\{x > 0\}$ by H_+, and let

$$A = (0, D] \times \{-1\} \cup (0, U] \times \{1\}.$$

Observe that

$$\begin{aligned} R(z, c^c, h) &= |h_1|^2 H_- \setminus A + |h_2|^2 H_+ \setminus A \\ &\quad + \left(|z - c_2^V - h_2|^2 - |z - c_1^V|^2 + 2h_1'(z - c_1^V) \right) AH_- \\ &\quad + \left(|z - c_1^V - h_1|^2 - |z - c_2^V|^2 + 2h_2'(z - c_2^V) \right) AH_+. \end{aligned}$$

Rewrite this expression in terms of $\widetilde{h}$:

$$\begin{aligned} R(z, c^c, h) &= \delta_s^2 + \delta_d^2 + \epsilon_s^2 + \epsilon_d^2 + 2(\delta_s\delta_d + \epsilon_s\epsilon_d)(H_- - H_+) \\ &\quad -4(x - x\delta_d - \delta_s - y\epsilon_d + \delta_s\delta_d + \epsilon_s\epsilon_d)(AH_- - AH_+). \end{aligned}$$

Note that $\nu_n(AH_- - AH_+) = o_p(1)$ uniformly over h in $\mathcal{N}_n$. This follows from approximation (2.24) and either the strong approximation to the empirical process (Komlós, Major, and Tusnády 1975), or an application of Theorem 7 in the appendix . It follows from the usual central limit theorem that $\nu_n(H_- - H_+) = O_p(1)$. Use the partition

$$A = (0, \delta_s - \epsilon_d] \times \{-1\} \cup (0, \delta_s + \epsilon_d] \times \{1\} \cup (\delta_s - \epsilon_d, D] \times \{-1\} \cup (\delta_s + \epsilon_d, U] \times \{1\}$$

along with approximation (2.24), to deduce that uniformly in $\mathcal{G}_n$,

$$
\begin{aligned}
n^{-1/2}\nu^{x,y}R(z, c^c, h) = \\
& b_n(\delta_s, \epsilon_d) + o_p(n^{-1}) \\
& + O_p(n^{-3/4})\nu_n(H_- - H_+)\left[(\delta_s - \epsilon_d, D] \times \{-1\} + (\delta_s + \epsilon_d, U] \times \{1\}\right],
\end{aligned}
$$

where $b_n(\delta_s, \epsilon_d)$ is a random function of just δ_s and ϵ_d. Approximation (2.24) for the crossing points of the split line implies that the length of intervals $(\delta_s - \epsilon_d, D]$ and $(\delta_s + \epsilon_d, U]$ is of order $O_p(n^{-5/8})$. Use oscillation properties of the empirical process from Shorack and Wellner (1986, page 765) to conclude that the last term in the equation above is $o_p(n^{-1})$ uniformly in $\mathcal{G}_n$. □

Use approximation (2.45) of Lemma 6 to compare $C_n(\widehat{\delta}_s, \widehat{\delta}_d, \widehat{\epsilon}_s, \widehat{\epsilon}_d)$ with $C_n(\widehat{\delta}_s, 0, \widehat{\epsilon}_d, 0)$ and derive

$$\widehat{\delta}_d{}^2 + \widehat{\epsilon}_s{}^2 = O_p(n^{-1/2}|(\widehat{\delta}_d, \widehat{\epsilon}_s)|) + o_p(n^{-1}).$$

Fix a positive C. The above inequality yields

$$|(\widehat{\delta}_d, \widehat{\epsilon}_s)|\{|(\widehat{\delta}_d, \widehat{\epsilon}_s)| > Cn^{-1/2}\} = O_p(n^{-1/2}),$$

which guarantees

$$|(\widehat{\delta}_d, \widehat{\epsilon}_s)| = O_p(n^{-1/2}).$$

Confine the search for $(\widehat{\delta}_s, \widehat{\delta}_d, \widehat{\epsilon}_s, \widehat{\epsilon}_d)$ to the sequence of $O_p(n^{-1/4}) \times O_p(n^{-1/2}) \times O_p(n^{-1/2}) \times O_p(n^{-1/4})$ neighborhoods $\{\mathcal{K}_n\}$ with $\mathcal{K}_n \subseteq \mathcal{G}_n$, such that $\widehat{h} \in \mathcal{K}_n$ with probability tending to one.

2.3.5 Limiting Distribution of $(\widehat{\delta}_s, \widehat{\epsilon}_d)$

Combine approximations (2.38) and (2.44) to deduce that uniformly in $\mathcal{K}_n$,

$$C_n(\widetilde{h}) = \tfrac{1}{6}\left(|\delta_s| + |\epsilon_d|\right)^3 + \tfrac{1}{6}\left||\delta_s| - |\epsilon_d|\right|^3 - n^{-1/2}(\delta_s, \epsilon_d)Y_n + o_p(n^{-3/4}), \qquad (2.47)$$

where

$$Y_n = (Y_{n1}, Y_{n2})' = 2\nu_n^{x,y}\begin{pmatrix} x + \{x \le 0\} - \{x > 0\} \\ y\{x \le 0\} - y\{x > 0\} \end{pmatrix}. \qquad (2.48)$$

Let $u = n^{-1/4}\delta_s$ and $v = n^{-1/4}\epsilon_d$. As $\widetilde{h}$ ranges over $\mathcal{K}_n$, vector $(u, v)'$ ranges over a neighborhood $\mathcal{B}_n$ in $\mathbb{R}^2$. Note that the sequence of the diameters of $\{\mathcal{B}_n\}$ is $O_p(1)$. Define

$$\begin{aligned} g(u,v) &= \tfrac{1}{6}\left(|u| + |v|\right)^3 + \tfrac{1}{6}\left||u| - |v|\right|^3, \text{ and} \\ f(u,v) &= g(u,v) - uY_{n1} - vY_{n2}. \end{aligned}$$

Note that function f is random, because it depends on the sample. The bivariate function g is symmetric and invariant to sign changes of its coordinates. Thus, if one knows the values of g on an octant, for example, $u \ge v \ge 0$, one knows the values of g on the whole plane.

Define the function $\mathcal{S}(x)$ by $\mathcal{S}(x) = \{x > 0\} - \{x < 0\}$. The following lemma locates the minimum of the function f.

Lemma 7 *For each fixed sample the minimum of f is unique. The random point (u^*, v^*) that minimizes f is given by*

$$\begin{aligned} u^* &= \tfrac{1}{2}\mathcal{S}(Y_{n1})[(|Y_{n1}| + |Y_{n2}|)^{1/2} + \mathcal{S}(|Y_{n1}| - |Y_{n2}|)\left||Y_{n1}| - |Y_{n2}|\right|^{1/2}] \\ v^* &= \tfrac{1}{2}\mathcal{S}(Y_{n2})[(|Y_{n1}| + |Y_{n2}|)^{1/2} - \mathcal{S}(|Y_{n1}| - |Y_{n2}|)\left||Y_{n1}| - |Y_{n2}|\right|^{1/2}], \end{aligned} \qquad (2.49)$$

where Y_{n1}, Y_{n2} are defined by formula (2.48).

There exist nonnegative random variables ξ_{1n} and ξ_{2n} whose limiting distributions concentrate on the positive half-line, such that for all u and v,

$$f(u,v) - f(u^*,v^*) \geq \left[(|u-u^*|^3 + |v-v^*|^3)\xi_{1n}\right] \wedge \xi_{2n}. \tag{2.50}$$

Remark. To simplify the notation I omit the subscript n for the variables u^* and v^*.

Proof: Fix the values of Y_{n1} and Y_{n2} to make f a function of just u and v. Since

$$|f| \geq |g| - |uY_{n1}| - |vY_{n2}| \geq \tfrac{1}{6}\left(|u|+|v|\right)^3 - |uY_{n1}| - |vY_{n2}|,$$

the large values of u and v do not matter in the minimization of f, and it would not be a loss of generality to assume that the minimization occurs over a large ball centered at the origin.

First, suppose $Y_{n1} \geq Y_{n2} \geq 0$. Because $f = g - uY_{n1} - vY_{n2}$ and g is invariant to sign changes of its coordinates, all the minima of f lie in $\{u \geq 0, v \geq 0\}$. Symmetry of g further confines the minima to the set $\mathcal{O}_1 = \{u \geq v \geq 0\}$, the first octant of the plane (u,v). The second derivative of f is given by the matrix

$$\frac{\partial f(u,v)}{\partial u \partial v} = 2\begin{pmatrix} u & v \\ v & u \end{pmatrix},$$

which is positive definite in the interior of the first octant. Equate the first derivative of f

to zero, to conclude that the only local minimum in the interior of the set $\mathcal{O}_1$ is given by

$$\begin{aligned} u^* &= \left[(Y_{n1} + Y_{n2})^{1/2} + (Y_{n1} - Y_{n2})^{1/2} \right] /2 \\ v^* &= \left[(Y_{n1} + Y_{n2})^{1/2} - (Y_{n1} - Y_{n2})^{1/2} \right] /2. \end{aligned} \tag{2.51}$$

Compare $f(u^*, v^*)$ the values of f on the boundary of the first octant. Observe that the unique minimum over the set $\{u \geq v = 0\}$ is achieved at $u = Y_{n1}^{1/2}$, the unique minimum over the set $\{u = v \geq 0\}$ is achieved at $u = v = \frac{1}{2}(Y_{n1} + Y_{n2})^{1/2}$, and the corresponding values of function f are

$$\begin{aligned} \min_{\{u \geq v = 0\}} f(u, v) &= -\tfrac{2}{3} Y_{n1}^{3/2} \\ \min_{\{u = v \geq 0\}} f(u, v) &= -\tfrac{1}{3}(Y_{n1} + Y_{n2})^{3/2}. \end{aligned} \tag{2.52}$$

Invoke convexity of the power function with positive exponent to conclude that

$$f(u^*, v^*) = -\tfrac{1}{3}((Y_{n1} + Y_{n2})^{3/2} + (Y_{n1} - Y_{n2})^{3/2}) \tag{2.53}$$

is no greater than either of these values. The equality only occurs if Y_{n1} equals Y_{n2} or if Y_{n2} equals zero. In both cases the point (u^*, v^*) coincides with the unique minimum over the boundary of the set $\mathcal{O}_1$. Conclude that the formula (2.51) gives the unique minimum of f in the case $Y_{n1} \geq Y_{n2} \geq 0$.

Consider a more general case $|Y_{n1}| \geq |Y_{n2}|$. The equality

$$g(u, v) - uY_{n1} - vY_{n2} = g\left(\mathcal{S}(Y_{n1})u, \mathcal{S}(Y_{n2})v \right) - \mathcal{S}(Y_{n1})u|Y_{n1}| - \mathcal{S}(Y_{n2})u|Y_{n2}|$$

implies that the unique minimum of the function f is given by

$$\begin{aligned}\mathcal{S}(Y_{n1})u^* &= \left[(|Y_{n1}|+|Y_{n2}|)^{1/2}+(|Y_{n1}|-|Y_{n2}|)^{1/2}\right]/2\\ \mathcal{S}(Y_{n2})v^* &= \left[(|Y_{n1}|+|Y_{n2}|)^{1/2}-(|Y_{n1}|-|Y_{n2}|)^{1/2}\right]/2.\end{aligned}$$

In the case $|Y_{n1}| \le |Y_{n2}|$ use the symmetry of the function g to derive

$$\begin{aligned}\mathcal{S}(Y_{n1})u^* &= \left[(|Y_{n2}|+|Y_{n1}|)^{1/2}-(|Y_{n2}|-|Y_{n1}|)^{1/2}\right]/2\\ \mathcal{S}(Y_{n2})v^* &= \left[(|Y_{n2}|+|Y_{n1}|)^{1/2}+(|Y_{n2}|-|Y_{n1}|)^{1/2}\right]/2.\end{aligned}$$

It is only left to establish lower bound (2.50). Fix the sample, and suppose that $Y_{n1} \ge Y_{n2} \ge 0$, and thus, $u^* \ge v^* \ge 0$. Establish the bound for points in $\mathcal{O}_1$, that is for $u \ge v \ge 0$.

Denote $u-u^*$ by h_u, and $v-v^*$ by h_v. Observe that

$$\begin{aligned}f(u,v)-f(u^*,v^*) &= u^*h_u^2+2v^*h_uh_v+u^*h_v^2+\tfrac{1}{3}h_u^3+h_uh_v^2\\ &= (u^*-v^*)(h_u^2+h_v^2)+v^*(h_u+h_v)^2+\tfrac{1}{3}h_u^3+h_uh_v^2.\end{aligned}$$

In the case $u^* > v^*$ bound the difference $f(u,v)-f(u^*,v^*)$ below by

$$\tfrac{1}{2}(u^*-v^*)(h_u^2+h_v^2) = \tfrac{1}{2}(Y_{n1}-Y_{n2})^{1/2}(h_u^2+h_v^2) \qquad (2.54)$$

for all (h_u,h_v) small enough. In the case $u^*=v^*$ the difference becomes

$$f(u,v)-f(u^*,v^*) = v^*(h_u+h_v)^2+\tfrac{1}{3}h_u^3+h_uh_v^2.$$

If h_u and h_v are the same sign, bound the difference below by

$$\tfrac{1}{2}v^*(h_u^2+h_v^2) = (\tfrac{1}{2}Y_{n1})^{1/2}(h_u^2+h_v^2) \tag{2.55}$$

for (h_u, h_v) small enough. If $h_u h_v \leq 0$, use $u \geq v$ and $u^* = v^*$ to deduce $h_u \geq 0$, and bound the difference below by

$$(|h_u|^3 + |h_v|^3)/6. \tag{2.56}$$

Let B_* denote the open ball of radius $(Y_{n1} - Y_{n2})^{1/2}/50$ centered at (u^*, v^*). Combine bounds (2.54), (2.55), and (2.56), to conclude that the following lower bound holds on $B_*\mathcal{O}_1$:

$$f(u,v) - f(u^*,v^*) \geq \tfrac{1}{6}[(Y_{n1}-Y_{n2})^{1/2} \wedge 1](|h_u|^3 + |h_v|^3). \tag{2.57}$$

Define $\xi_{1n} = \frac{1}{6}[(Y_{n1}-Y_{n2})^{1/2} \wedge 1]$.

Since (u^*, v^*) is the only local minimum of f in the interior of $\mathcal{O}_1$, and the large values of u and v do not matter in the minimization, the minimum of f over $(B_*\mathcal{O}_1)^c$ must be achieved either on the boundary of $B_*\mathcal{O}_1$ or on the boundary of $\mathcal{O}_1$. Apply formulas (2.52), (2.53), and (2.57) to deduce the following lower bound on $(B_*\mathcal{O}_1)^c$:

$$\begin{aligned} f(u,v) - f(u^*,v^*) \quad \geq \quad & c\left[(Y_{n1}+Y_{n2})^{3/2} + (Y_{n1}-Y_{n2})^{3/2} - 2Y_{n1}^{3/2}\right] \\ & \wedge (Y_{n1}-Y_{n2})^{3/2} \wedge (Y_{n1}-Y_{n2})^2, \end{aligned}$$

where c is a positive constant. Denote the random quantity on the right hand side of the last inequality by ξ_{2n}.

Combine the lower bound on $B_*\mathcal{O}_1$ with the bound on $(B_*\mathcal{O}_1)^c$ to deduce that for

$u^* \geq v^* \geq 0$ and $u \geq v \geq 0$

$$f(u,v) - f(u^*,v^*) \geq \left[(|u-u^*|^3 + |v-v^*|^3)\xi_{1n}\right] \wedge \xi_{2n}. \tag{2.58}$$

Use the properties of f to extend this lower bound to the general case. Random variables ξ_{1n} and ξ_{2n} are nonnegative. Since (Y_{n1}, Y_{n2}) converges to a non-degenerate bivariate gaussian distribution, the limiting distributions of ξ_{1n} and ξ_{2n} assign probability one to the positive half-line. □

Write approximation (2.47) in terms of u and v:

$$n^{-3/4}C_n(\widetilde{h}) = f(u,v) + o_p(1),$$

uniformly in $\mathcal{K}_n$. Let $(\widehat{u},\widehat{v})$ and $(\delta_s^*, \epsilon_d^*)$ stand for $n^{1/4}(\delta_s, \epsilon_d^*)$ and $n^{-1/4}(u^*, v^*)$ respectively. Note that δ_s^* and ϵ_d^* are of order $O_p(n^{-1/4})$. Go back and change the neighborhoods $\mathcal{N}_n, \mathcal{G}_n$, and $\mathcal{K}_n$, to make sure that $\widehat{h}^* = (\delta_s^*, \widehat{\delta}_d, \widehat{\epsilon}_s, \epsilon_d^*)'$ lies in $\mathcal{K}_n$ with probability tending to one. Compare $C_n(\widehat{h})$ with $C_n(\widehat{h}^*)$ to deduce that

$$f(\widehat{u},\widehat{v}) \leq f(u^*,v^*) + o_p(1).$$

Combine the last inequality with lower bound (2.50) from Lemma 7 to conclude

$$\left[(|\widehat{u}-u^*|^3 + |\widehat{v}-v^*|^3)\xi_{1n}\right] \wedge \xi_{2n} = o_p(1).$$

Denote the expression on the left-hand side by α_n. Use the bound

$$P\{|\widehat{u}-u^*|^3 + |\widehat{v}-v^*|^3 > \epsilon\} \leq P\{\alpha_n > \epsilon\xi_{1n}\} + P\{\alpha_n > \xi_{2n}\}$$

for every positive ϵ, and the limiting properties of ξ_{1n}, ξ_{2n} to derive

$$|\widehat{u} - u^*| + |\widehat{v} - v^*| = o_p(1).$$

It follows that

$$\begin{aligned} \widehat{\delta}_s &= n^{-1/4}u^* + o_p(1) \\ \widehat{\epsilon}_d &= n^{-1/4}v^* + o_p(1), \end{aligned} \tag{2.59}$$

where the expressions for u^*, v^* are given by formula (2.49).

2.3.6 Limiting Distribution of $(\widehat{\delta}_d, \widehat{\epsilon}_s)$

Recall approximation (2.44), which holds uniformly in $\mathcal{K}_n$:

$$C_n(\widetilde{h}) = W(c^V + h) - W(c^V) - n^{-1/2}h'Z_n + o_p(n^{-1/2}|h|).$$

Lemma 6 splits the $o_p(n^{-1/2}|h|)$ term into a random function of (δ_s, ϵ_d) and a $o_p(n^{-1})$ term uniformly in $\mathcal{G}_n$. Apply expansion (2.38) to rewrite the above approximation as

$$C_n(\widetilde{h}) = \delta_d^2 + \epsilon_s^2 + \delta_s^2\delta_d + 2\delta_s\epsilon_s\epsilon_d - \delta_d\epsilon_d^2 - n^{-1/2}(\delta_d, \epsilon_s)'T_n + b_n(\delta_s, \epsilon_d) + o_p(n^{-1}),$$

uniformly in $\mathcal{K}_n$. Here $b_n(\delta_s, \epsilon_d)$ is some random function of (δ_s, ϵ_d), and

$$T_n = (T_{n1}, T_{n2})' = 2\nu_n^{x,y} \begin{pmatrix} 1 + x\{x \leq 0\} - x\{x > 0\} \\ y \end{pmatrix}. \tag{2.60}$$

Let $S = (S_1, S_2)'$ stand for $\frac{1}{2}(T_{n1} - \widehat{u}^2 + \widehat{v}^2, T_{n2} - 2\widehat{u}\widehat{v})'$. Apply formula (2.59) to derive

$$C_n(\widehat{\delta}_s, \delta_d, \epsilon_s, \widehat{\epsilon}_d) = \delta_d^2 + \epsilon_s^2 - 2n^{-1/2}(\delta_d, \epsilon_s)'S + b_n(\widehat{\delta}_s, \widehat{\epsilon}_d) + o_p(n^{-1}),$$

uniformly in $\mathcal{K}_n$. Complete the square and write the above approximation as

$$C_n(\widehat{\delta}_s, \delta_d, \epsilon_s, \widehat{\epsilon}_d) = |(\delta_d, \epsilon_s)' - n^{-1/2}S|^2 - n^{-1}|S|^2 + b_n(\widehat{\delta}_s, \widehat{\epsilon}_d) + o_p(n^{-1}).$$

Note that S_1 and S_2 are of order $O_p(1)$. Go back and change the neighborhoods $\mathcal{N}_n, \mathcal{G}_n$, and $\mathcal{K}_n$, to make sure that the vector $\widehat{h}^s = (\widehat{\delta}_s, n^{-1/2}S_1, n^{-1/2}S_2, \widehat{\epsilon}_d)'$ lies in $\mathcal{K}_n$ with probability tending to one. Compare $C_n(\widehat{h})$ with $C_n(\widehat{h}^s)$ to conclude that

$$\begin{aligned} \widehat{\delta}_d &= n^{-1/2}[T_{n1} - (u^*)^2 + (v^*)^2]/2 + o_p(1) \\ \widehat{\epsilon}_s &= n^{-1/2}(T_{n2} - 2u^*v^*)/2 + o_p(1), \end{aligned} \tag{2.61}$$

where the expressions for $u^*, v^*, T_{n1},$ and T_{n2}, are given by formulas (2.49) and (2.60).

2.3.7 Asymptotics of Optimal Empirical Centers

Recall that the following equality holds with probability tending to one:

$$b_n = \underset{b \in \{b_n^H, b_n^V\}}{\operatorname{argmin}} W_n(b).$$

Write approximation (1.5) for the empirical criterion function W_n near c^H and combine it with approximation (2.21) for the population criterion function W and approximation (2.39) for the centers b_n^H. Deduce that

$$W_n(b_n^H) - W_n(c_n^H) = O_p(n^{-1}).$$

Combine approximation (2.47) for the empirical criterion function W_n with approximations (2.59) and (2.61) for the centers b_n^V and deduce that

$$W_n(b_n^V) - W_n(c_n^V) = O_p(n^{-3/4}).$$

Conclude that if W^* denotes the minimum of the function W, the following approximation holds:

$$n^{1/2}\left(W_n(b_n^H) - W^*, W_n(b_n^V) - W^*\right) = \nu_n^z\left(\phi(z, c_n^H), \phi(z, c_n^V)\right) + O_p(n^{-1/4}).$$

Use this approximation together with approximations (2.39), (2.59) and (2.61) to derive the limiting distribution for the sequence of vectors

$$\left(n^{1/2}[W_n(b_n^H) - W^*], n^{1/2}[W_n(b_n^V) - W^*], n^{1/2}[b_n^H - c^H], n^{1/4}[b_n^V - c^V]\right).$$

With probability tending to one, the set of optimal sample centers b_n coincides with either b_n^H or b_n^H. The probability of $b_n = b_n^H$ converges to the probability of $A_1 < A_2$, where (A_1, A_2) is the weak limit of $\nu_n^z\left(\phi(z, c_n^H), \phi(z, c_n^V)\right)$. Observe that (A_1, A_2) has a centered gaussian distribution with covariance matrix

$$\begin{pmatrix} 20 & 12 \\ 12 & 8 \end{pmatrix},$$

and thus the probability of $b_n = b_n^H$ converges to $1/2$.

Combine approximations (2.21) and (2.38) for the population criterion function W near c^H and c^V with approximations (2.61) and (2.61) for b_n^H and b_n^V to deduce that vectors

$$\left(n[W(b_n^H) - W^*], n^{3/4}[W(b_n^V) - W^*]\right)$$

settle down to a non-degenerate two-dimensional limiting distribution. Hence, distortion redundancy $W(b_n) - W^*$ can be said to have two different rates at which it settles down.

Chapter 3

k-Means Asymptotics when Population Solution is Not Unique

In this chapter I establish some general results that handle non-uniqueness of the population solution. In addition to consistency I prove an $n^{-1/2}$ rate of convergence result and a central limit theorem.

3.1 Consistency

The following proposition is an extension of the main result in Pollard (1981). It is a natural extension and it has been recognized before (see, for example, Pollard 1982b.)

Proposition 1 *Suppose P is not concentrated on a set of points of cardinality smaller than k. The set b_n of optimal empirical centers converges almost surely to the set C of all population minima:*

$$\min_{a \in C} H(b_n, a) \to 0 \quad \text{a.s.}$$

Proof: The consistency proof in Pollard (1981) did not use uniqueness of the population minimum to establish the following facts for almost all sample points:

(i) All the optimal sample centers eventually lie in some compact region $\mathcal{K}$ of $\mathbb{R}^d$.

(ii) $W_n(b) - W(b)$ converges to zero uniformly over $\mathcal{E}_k$, a collection of all nonempty subsets of $\mathcal{K}$ that contain k or fewer points.

(iii) $W(b)$ is continuous on $\mathcal{E}_k$ with respect to the Hausdorff distance.

Note that continuity of function W implies compactness of C with respect to the Hausdorff metric. Fix a sample point for which the above statements are true and take any positive ϵ. Consider the following compact set:

$$F_\epsilon = \{b \in \mathcal{E}_k : \min_{a \in C} H(b, a) \geq \epsilon\}.$$

The minimum of the population criterion function W on the set F_ϵ, call it $m(F_\epsilon)$, is strictly greater than W^*, the global minimum of W. Suppose that for all b in $\mathcal{E}_k$, the difference $W_n(b) - W(b)$ is less than $m(F_\epsilon) - W^*$ whenever n is greater than some n_0. Then, for n greater than n_0, the set b_n of optimal sample centers lies outside of F_ϵ. □

3.2 Rate of Convergence

Recall the approximation to the empirical criterion function W_n given by Pollard (1982a), which holds for each element a of the set C:

$$W_n(a + h) - W_n(a) = W(a + h) - W(a) - n^{-1/2} h' Z_n(a) + o_p(n^{-1/2}|h|). \qquad (3.1)$$

A set $\{a_1, \ldots, a_k\}$ of centers partitions $\mathbb{R}^d$ into k polyhedral regions, the region A_i associated with a_i consists of those x closer to a_i than to any other center. This partition is called **Voronoi tesselation**, and the regions A_i are called **Voronoi polyhedra**. The assignment of boundary points of a Voronoi polyhedron to one of the closest centers can be handled

by a tie-breaking rule. Random variables $Z_n(a)$ in approximation (3.1) are defined as $\nu_n^x \Delta(x,a)$, where $\Delta(x,a)$ is an $\mathbb{R}^{kd}$ vector given by

$$\Delta(x,a) = 2\left(A_1(x-a_1), ..., A_k(x-a_k)\right). \tag{3.2}$$

The precise convention for allocating the boundary points of Voronoi polyhedra associated with a is unimportant, as will be seen from Lemma 8.

Pollard (1982a) derives approximation 3.1 by extracting a linear term in h from the empirical process $\nu_n^x \phi(x, a+h)$. Recall that $\phi(x,b)$ gives the squared distance from x to the closest point in b. Note that $\phi(x,a)$ as a function of x is non-differential on the boundaries of the Voronoi polyhedra associated with a. If the boundary of a Voronoi polyhedron contains positive probability, ϕ may no longer be differentiable at a in quadratic mean. Pollard (1982a) was interested in a central limit theorem for distributions that have continuous densities with respect to Lebesgue measure in $\mathbb{R}^d$. He derived approximation (3.1) under the assumption that the boundaries of the Voronoi polyhedra associated with a contain zero probability. The following lemma shows that this assumption can be dropped.

Lemma 8 *Suppose P has a finite second moment and does not concentrate on fewer than k points. Suppose that a set a minimizes the k-means criterion function W. Let k be greater than one. Then the boundary of each Voronoi polyhedron associated with a has zero P measure.*

Proof: Start with the simplest case of two means in $\mathbb{R}^d$. Let a_1 and a_2 be an optimal pair of centers. Denote the mean of P by μ. Define regions A_1 and A_2 by

$$\begin{aligned} A_1 &= \{x \in \mathbb{R}^d : |x-a_1|^2 < |x-a_2|^2\} \\ A_2 &= \{x \in \mathbb{R}^d : |x-a_2|^2 < |x-a_1|^2\}. \end{aligned}$$

Observe that regions A_1 and A_2 are the interiors of the Voronoi polyhedra associated with a. Denote the intersection of the closures of A_1 and A_2 by A_m. In other words,

$$A_m = \overline{A}_1 \cap \overline{A}_2 = \{x \in \mathbb{R}^d : |x - a_1|^2 = |x - a_2|^2\}.$$

Let p_1 and p_2 stand for PA_1 and PA_2 respectively. Denote the conditional means of A_1 and $\overline{A}_2$ by b_1 and b_2 and the conditional means of $\overline{A}_1$ and A_2 by c_1 and c_2.

Suppose that, contrary to the statement of the lemma, P assigns positive probability Δ to the hyperplane A_m. Observe that equalities $(a_1, a_2) = (b_1, b_2) = (c_1, c_2)$ can not hold simultaneously, because otherwise the equalities

$$\begin{aligned} (p_1 + \Delta)a_1 + p_2 a_2 &= \mu \\ p_1 a_1 + (p_2 + \Delta)a_2 &= \mu \end{aligned}$$

would lead to $a_1 = a_2$. Suppose without loss of generality that $(b_1, b_2) \neq (a_1, a_2)$. Note that PA_1 and $P\overline{A}_2$ are both positive because P is not concentrated on just one point. Observe that

$$\begin{aligned} W(a_1, a_2) &= P^x(x - a_1)^2 \wedge (x - a_2)^2 \\ &= P^x\{x \in A_1\}(x - a_1)^2 + P^x\{x \in \overline{A}_2\}(x - a_2)^2 \\ &= P^x\{x \in A_1\}(x - b_1)^2 + PA_1(b_1 - a_1)^2 + P^x\{x \in \overline{A}_2\}(x - b_2)^2 + P\overline{A}_2(b_2 - a_2)^2 \\ &> P^x\{x \in A_1\}(x - b_1)^2 + P^x\{x \in \overline{A}_2\}(x - b_2)^2 \\ &\geq P^x(x - b_1)^2 \wedge (x - b_2)^2 \\ &= W(b_1, b_2). \end{aligned}$$

The resulting strict inequality contradicts the assumption that the pair (a_1, a_2) minimizes

the function W. The proof caries over to the general case with minor adjustments. When k is greater than two, work with a particular pair of centers and substitute hyperplane A_m with the face of the corresponding Voronoi polyhedron. □

Approximation (3.1) is uniform over shrinking neighborhoods of a population minimum a. To handle the empirical criterion function in a neighborhood of the whole collection of minima, I will show that the remainder terms in (3.1) are small uniformly over the set C. The following lemma makes this statement precise.

Lemma 9 *Let $\mathcal{R}_n(a,h)$ be defined by*

$$W_n(a+h) - W_n(a) = W(a+h) - W(a) - n^{-1/2}h'Z_n(a) + n^{-1/2}|h|\mathcal{R}_n(a,h). \quad (3.3)$$

The following approximation holds for any sequence $\{r_n\}$ of positive numbers decreasing to zero:

$$\sup_{a\in C, |h|\le r_n} |\mathcal{R}_n(a,h)| = o_p(1).$$

Proof: Note that $\mathcal{R}(a,h) = \nu_n^x R(x,a,h)$, where $R(x,a,h)$ is defined by

$$\phi(x,a+h) = \phi(x,a) + h'\Delta(x,a) + |h|R(x,a,h).$$

Consider a class $\mathcal{F}$ of functions $R(\cdot,a,h)$ with a ranging over the set C and h in some bounded set. Follow Pollard (1982a) to conclude that each function in this class can be written as a sum of k^2 members of the class $\mathcal{G}$ of functions with the following properties:

(i) each g in $\mathcal{G}$ is of the form $g = LQ$ where L is a linear function and Q is a convex region expressible as an intersection of at most $2k$ open or closed half spaces.

(ii) $\mathcal{G}$ has a square integrable envelope.

Theorem 7 in the appendix implies that the empirical process $\{\nu_n f : f \in \mathcal{F}\}$ is stochastically equicontinuous at zero. The only thing left to show is that as n tends to infinity, functions $R(\cdot, a, h)$ go to zero uniformly with respect to the $L_2(P)$ norm, i.e.

$$\sup_{a \in C, |h| \leq r_n} \|R(\cdot, a, h)\|_2 \to 0 \qquad \text{as} \quad n \to \infty. \tag{3.4}$$

Consider an a in C. Label the points in the set a by $a_1, a_2, \ldots, a_k$. Define

$$E_n(a) = \{x \in \mathbb{R}^d, \quad \text{s.t.} \quad |x - a_i|^2 \vee |x - a_j|^2 \leq \phi(x, a) + 3r_n \quad \text{for some} \quad i \neq j\}. \tag{3.5}$$

Sets $E_n(a)$ consist of points near $l(a)$, the boundary of all the Voronoi polyhedra associated with a. Note that as n tends to infinity, $E_n(a)$ converges to $l(a)$ pointwise. For $|h| \leq r_n$ bound the contribution to $R(x, a, h)$ on the set $E_n(a)^c$:

$$R(x, a, h) E_n(a)^c = |h|^2 E_n(a)^c \leq |h|^2.$$

Use the above inequality to bound the $L_2(P)$ norm of $R(\cdot, a, h)$:

$$\|R(\cdot, a, h)\|_2 \leq \|R(\cdot, a, h) E_n(a)\|_2 + |h|^4.$$

The second term on the right hand side does not depend on a and can be easily handled:

$$\sup_{a \in C, |h| \leq r_n} \|R(\cdot, a, h)\|_2 \leq \sup_{a \in C, |h| \leq r_n} \|R(\cdot, a, h) E_n(a)\|_2 + r_n^4.$$

Since the class of functions $\{R(\cdot, a, h), a \in C, |h| \leq r_1\}$ has a square integrable envelope,

it is only left to show that

$$\sup_{a \in C} PE_n(a) \to 0, \qquad \text{as} \quad n \to \infty, \tag{3.6}$$

to establish convergence (3.4).

For each fixed population minimum a, indicator functions $E_n(a)$ converge pointwise to $l(a)$, the boundary of all the Voronoi polyhedra associated with a. By Lemma 8 there can be no mass on $l(a)$, hence

$$PE_n(a) \to 0, \quad \text{as} \quad n \to \infty. \tag{3.7}$$

The above convergence together with compactness of the set C imply (3.6) if for all ϵ and n the sets $\{a \in C : PE_n(a) < \epsilon\}$ are open. Thus, to complete the proof of the lemma it is only left to establish the upper semicontinuity with respect to the Hausdorff metric of $PE_n(a)$ as functions of a. An equivalent property is lower semicontinuity of $PE_n^c(a)$. By Fatou's Lemma,

$$\liminf_{a \to b} PE_n^c(a) \geq P \liminf_{a \to b} E_n^c(a), \tag{3.8}$$

where the convergence is understood with respect to the Hausdorff metric. Note that

$$\liminf_{a \to b} E_n^c(a) \geq E_n^c(b). \tag{3.9}$$

Indeed, suppose x belongs to the set $E_n^c(b)$. Label the points in b by $b_1, b_2, \ldots, b_k$ in such a way that $\phi(x, b)$ is exactly $|x - b_1|^2$. Then, by definition (3.5) of the sets $E_n(\cdot)$, there exists a positive δ, which depends on x, such that

$$|x - b_i|^2 > \phi(x, b) + 3r_n + \delta \qquad \text{for all} \quad i \geq 2.$$

Consider an a with $H(a,b) < \delta/5$. Suppose that δ is small enough to ensure that the labeling of the points in b induces a unique labeling of the points in a. Observe that

$$\begin{aligned} |x-a_1|^2 &< \phi(x,b) + \delta/5 \\ |x-a_i|^2 &> \phi(x,b) + 3r_n + 4\delta/5 \qquad \text{for all} \quad i \geq 2, \end{aligned}$$

and hence

$$\begin{aligned} |x-a_1|^2 &= \phi(x,a) \\ |x-a_i|^2 &> \phi(x,a) + 3r_n + 3\delta/5 \qquad \text{for all} \quad i \geq 2. \end{aligned}$$

Deduce that x belongs to the set $E_n(a)^c$ for all a in a small enough Hausdorff neighborhood of b. This completes the proof of inequality (3.9). Combine this inequality with inequality (3.8) to conclude

$$\liminf_{a \to b} PE_n^c(a) \geq PE_n^c(b).$$

Thus, functions $PE_n^c(a)$ are lower semicontinuous with respect to the Hausdorff metric, which completes the proof of the lemma. □

Definition 4 *For a compact set b in $\mathbb{R}^d$ and a compact subset E of the space of compact sets in $\mathbb{R}^d$ equipped with the Hausdorff metric define $d_H(b,E)$ by*

$$d_H(b,E) = \min_{e \in E} H(b,e).$$

Call $d_H(b,E)$ the 'Hausdorff distance from b to E'.

Theorem 1 *Consider an epsilon neighborhood of the set C of all population minima:*

$$\mathcal{N}_\epsilon(C) = \{b : d_H(b,C) < \epsilon\}.$$

Let W^* denote the minimum value of W, which is achieved on C. Suppose there exist a positive ϵ such that

$$\inf_{b \in \mathcal{N}_\epsilon(C)} \frac{W(b) - W^*}{d_H(b, C)^2} > 0. \tag{3.10}$$

Then the distance from the sample minimum $b_n$to the set C of all population minima decreases at the $O_p(n^{-1/2})$ rate:

$$d_H(b_n, C) = O_p(n^{-1/2}).$$

Proof: Denote the positive quantity on the left hand side of inequality (3.10) by α. The sample optimum b_n minimizes the empirical criterion function W_n. Let b_n^* be a set in C that satisfies

$$H(b_n, b_n^*) = d_H(b_n, C).$$

Consistency of b_n implies that there exist a sequence of positive numbers r_n decreasing to zero, such that $H(b_n, b_n^*) < r_n$ holds with probability tending to 1. Without loss of generality, suppose that r_n is less than ϵ. Also suppose that r_n is small enough to safely write $|b_n - b_n^*|$ rather than $H(b_n, b_n^*)$. Use approximation (3.3) to make a comparison between $W_n(b_n)$ and $W_n(b_n^*)$:

$$\begin{aligned} 0 &\geq W_n(b_n) - W_n(b_n^*) \\ &= W(b_n) - W(b_n^*) - n^{-1/2}(b_n - b_n^*)' Z_n(b_n^*) + n^{-1/2}|b_n - b_n^*| \mathcal{R}_n(b_n^*, b_n - b_n^*). \end{aligned}$$

Deduce that

$$W(b_n) - W(b_n^*) \leq n^{-1/2}(b_n - b_n^*)' Z_n(b_n^*) - n^{-1/2}|b_n - b_n^*| \mathcal{R}_n(b_n^*, b_n - b_n^*).$$

Note that $W(b_n^*) = W^*$. By inequality (3.10) the difference $W(b_n) - W(b_n^*)$ has a lower

bound of $\alpha|b_n - b_n^*|^2$. Conclude that

$$\alpha n^{1/2}|b_n - b_n^*|^2 \leq |b_n - b_n^*||Z_n(b_n^*)| + |b_n - b_n^*||\mathcal{R}_n(b_n^*, b_n - b_n^*|,$$

and hence

$$\alpha n^{1/2} d_H(b_n, C) \leq \sup_{a \in C} |Z_n(a)| + \sup_{a \in C, |h| \leq r_n} |\mathcal{R}_n(a, h)|.$$

Apply Lemma 9 to handle the last term. It is only left to show that

$$\sup_{a \in C} |Z_n(a)| = O_p(1). \tag{3.11}$$

Note that $\{Z_n(a), a \in C\}$ is the empirical process $\nu_n(\cdot)$ indexed by a collection of $\mathbb{R}^{kd}$-valued functions $\Delta(\cdot, a)$ defined by (3.2). Coordinate functions of $\Delta(\cdot, a)$ are members of the class $\mathcal{F}$ defined in the proof of Lemma 9. Apply statement (ii) of Theorem 7 in the appendix to derive (3.11). □

3.3 Central Limit Theorem

Theorem 2 *Suppose that for each a in C there exists a symmetric matrix $\Gamma(a)$, such that as h tends to zero,*

$$\sup_{a \in C} |W(a + h) - W(a) - h'\Gamma(a)h| = o(|h|^2). \tag{3.12}$$

Let λ_a be the smallest positive eigenvalue of $\Gamma(a)$. Suppose that

$$\inf_{a \in C} \lambda_a > 0. \tag{3.13}$$

Let $\ker^{\perp}\Gamma(a)$ *be the orthogonal complement to the linear subspace* $\{x \in \mathbb{R}^{kd} : \Gamma(a)x' = 0\}$. *Denote by* $K(a)$ *the collection of sets represented in* $\mathbb{R}^{kd}$ *by* $\ker^{\perp}\Gamma(a) + a$. *Suppose that there exists an* ϵ, *such that*

$$\mathcal{N}_{\epsilon}(C) \subseteq \bigcup_{a \in C} K(a), \tag{3.14}$$

and for each b *in* $\mathcal{N}_{\epsilon}$ *there exists a* b^* *that achieves the Hausdorff distance from* b *to* C, *such that*

$$b \in K(b^*). \tag{3.15}$$

Then for each n *there exists a random process* $\{h_n(a), a \in C\}$, *such that if* a_n *is defined as the minimizer of* $W_n(a + h_n(a))$ *over all* a *in* C, *then*

$$b_n = a_n + h_n(a_n).$$

As n *goes to infinity, the processes*

$$n^{1/2}\left\{ \left(W_n(a + h_n(a)) - W^* + P|x|^2 - P_n|x|^2,\ h_n(a) \right),\ a \in C \right\}$$

converge weakly to the centered gaussian process

$$\left\{ \left(Y(a), X(a) \right),\ a \in C \right\}$$

that takes values in $\mathbb{R} \times \mathbb{R}^{kd}$. *In addition, random variables* a_n *converge weakly to* $\operatorname{argmin}_C Y(a)$.

Proof: Verify that the conditions of Theorem 1 are satisfied. Without loss of generality, suppose that ϵ is small enough to safely use the $\mathbb{R}^{kd}$-representation in the neighborhood $\mathcal{N}_{\epsilon}$. For all a in C and all nonzero h in $\mathbb{R}^{kd}$ define $r(a, h)$ by

$$W(a + h) - W(a) - h'\Gamma(a)h = |h|^2 r(a, h).$$

Without loss of generality, assume that ϵ is small enough to guarantee

$$\sup_{a \in C, |h| < \epsilon} |r(a,h)| < \inf_{a \in C} \lambda_a. \tag{3.16}$$

Fix a set b in $\mathcal{N}_\epsilon(C)$. Let b^* be a set in C that achieves the Hausdorff distance from b to C and satisfies condition (3.15). Observe that

$$\frac{W(b) - W(b^*)}{H(b,b^*)^2} \geq \lambda_a + r(a, b-a).$$

Deduce that

$$\inf_{b \in \mathcal{N}_\epsilon(C)} \frac{W(b) - W^*}{d_H(b,C)^2} \geq \inf_{a \in C} \lambda_a - |r(a,b-a)| > 0.$$

The last inequality follows from the lower bound (3.16). Hence, by Theorem 1, the empirical minimum b_n approaches the set C of all population minima at the rate $O_p(n^{-1/2})$. Let γ_n be a $O_p(n^{-1/2})$ sequence of positive random variables, such that

$$b_n \in \mathcal{N}_{\gamma_n}(C).$$

For all $r > 0$ and all a in C define sets $K_r(a)$ by

$$K_r(a) = \{b : \ b \in K(a) \ \text{ and } \ H(b,a) < r\}.$$

It follows from assumptions (3.14) and (3.15) that for each b in $\mathcal{N}_{\gamma_n}(C)$ there exists a b^* in C, such that b lies in the set $K_{\gamma_n}(b^*)$. Hence,

$$\mathcal{N}_{\gamma_n}(C) \subseteq \bigcup_{a \in C} K_{\gamma_n}(a),$$

for all n. Split the minimization of W_n over $\mathcal{N}_{\gamma_n}(C)$ into two parts: minimization over

$K_{\gamma_n}(a)$ and minimization over all a in C.

Consider a population minimum a. Fix an orthonormal basis in $\ker^{\perp}\Gamma(a)$. For any vector h in R^{kd} let $\widetilde{h}$ be the coordinate form of h written with respect to this orthonormal basis. Let $\widetilde{\Gamma}(a)$ be the coordinate form of linear operator $\Gamma(a)$ with respect to the same orthonormal basis. For every a in C and every h in $\ker^{\perp}\Gamma(a)$, the quadratic approximation to the population criterion function has the form

$$W(a+h) - W(a) = \widetilde{h}'\widetilde{\Gamma}(a)\widetilde{h} + |h|^2 r(a,h).$$

Write the quadratic approximation to the empirical criterion function in $K(a)$:

$$W_n(a+h) - W_n(a) = \widetilde{h}'\widetilde{\Gamma}(a)\widetilde{h} - n^{-1/2}\widetilde{h}'\widetilde{Z}_n(a) + |h|^2 r(a,h) + n^{-1/2}|h|\mathcal{R}(a,h). \quad (3.17)$$

Lemma 9 guarantees that $\sup_{a\in C}|\mathcal{R}(a,h)| = o_p(1)$ holds uniformly over $|h| < \gamma_n$. Condition (3.12) implies that $\sup_{a\in C}|r(a,h)|$ is $o(1)$ as h goes to zero. Refine approximation (3.17) in the set $K_{\gamma_n}(a)$:

$$W_n(a+h) - W_n(a) = \widetilde{h}'\widetilde{\Gamma}(a)\widetilde{h} - n^{-1/2}\widetilde{h}'\widetilde{Z}_n(a) + o_p(n^{-1}). \quad (3.18)$$

The remainder term is $o_p(n^{-1})$ uniformly over all $a \in C$ and $|h| < \gamma_n$. Let $h_n(a)$ be the deviation from a of the empirical minimum in $K_{\gamma_n}(a)$:

$$h_n(a) = \operatorname*{argmin}_{K_{\gamma_n}(a)} W_n(\cdot) - a.$$

Use the standard comparison argument to deduce

$$\widetilde{h}_n(a) = \tfrac{1}{2}n^{-1/2}\widetilde{\Gamma}(a)^{-1}\widetilde{Z}_n(a) + o_p(n^{-1/2}),$$

uniformly over all a in C. In the original coordinate system,

$$h_n(a) = n^{-1/2}\xi_n(a) + o_p(n^{-1/2}), \tag{3.19}$$

where $\xi_n(a)$ is the $\mathbb{R}^{kd}$ coordinate form of $\frac{1}{2}\widetilde{\Gamma}(a)^{-1}\widetilde{Z}_n(a)$. Thus, $\xi_n(a)$ is of the form $L(a)Z_n(a)$, where $L(a)$ is a deterministic $\mathbb{R}^{kd} \times \mathbb{R}^{kd}$ matrix, which does not depend on n.

Approximation (3.18) implies that uniformly over a in C,

$$W_n(a + h_n(a)) = W_n(a) - \tfrac{1}{2}n^{-1}\widetilde{\Gamma}(a)^{-1}\widetilde{Z}_n(a) + o_p(n^{-1}).$$

Equation (3.11) together with inequality (3.13) give $\sup_{a\in C} |\widetilde{\Gamma}(a)^{-1}\widetilde{Z}_n(a)| = O_p(1)$. Conclude that

$$n^{1/2}\left(W_n(a + h_n(a)) - W^*\right) = \nu_n^x\phi(x, a) + O_p(n^{-1/2}),$$

uniformly over a in C. Recall that $Z_n(a)$ is defined as $\nu_n^x\Delta(x, a)$. Apply approximation (3.19) and deduce that the weak limit of the sequence of processes

$$n^{1/2}\left\{\left(W_n(a + h_n(a)) - W^* + P|x|^2 - P_n|x|^2,\ h_n(a)\right),\ a \in C\right\}$$

is the same as the weak limit of the sequence

$$\left\{\nu_n^x\left(\phi(x, a) - |x|^2, \Delta(x, a)\right),\ a \in C\right\}. \tag{3.20}$$

Functional classes $\{\phi(\cdot, a) - |\cdot|^2, a \in C\}$ and $\{\Delta(\cdot, a), a \in C\}$ satisfy the conditions of Theorem 7 in the appendix. By the functional central limit theorem (statement (iv) of Theorem 7) sequence (3.20) of stochastic processes converges in distribution to a centered multidimensional gaussian process. Apply continuous mapping theorem to derive the weak convergence of a_n. Note that a_n converges as a random element of the metric

space (C, ρ), where ρ is a pseudo-metric (see example 5) defined by

$$\rho(a_1, a_2) = P^x[\phi(x, a_1) - \phi(x, a_2)]^2.$$

To formulate the convergence of a_n in terms of the Hausdorff metric, use the almost sure representation (Dudley 1999, Theorem 3.5.1). Recall that metric d_H is defined by $d_H(a_0, A) = \min_{a \in A} H(a_0, a)$, where a is a set and A is a collection of sets. Dudley's almost sure representation provides a probability space on which copies of a_n converge almost surely with respect to d_H to the argmin a gaussian process with the correct distribution. □

The following example illustrates how the above limit theorem does not distinguish between population minima that are in the same equivalence class with respect to the pseudo-metric ρ.

Example 5 Consider a two means problem on the plane. Suppose the underlying distribution P is uniform over four points with the following Euclidean coordinates: $(-1, 1)$, $(1, 1)$, $(1, -1)$, $(-1, -1)$. These points are vertices of a square centered at the origin. The population within cluster sum of squares is 4 for partitions that put exactly three vertices in one of the clusters, and it is 2 for partitions that put two adjacent vertices in one cluster and the other two vertices in the other cluster. Hence there are two optimal population configurations of centers, call them c^V and c^H depending on wether the corresponding split line is vertical or horizontal:

$$c^V = \{(-1, 0), (1, 0)\} \qquad c^H = \{(0, -1), (0, 1)\}.$$

Figure 3.1 shows the four vertices of the square together with the two optimal pairs of centers. The triangles correspond to the pair c^H, and the empty circles correspond to c^V.

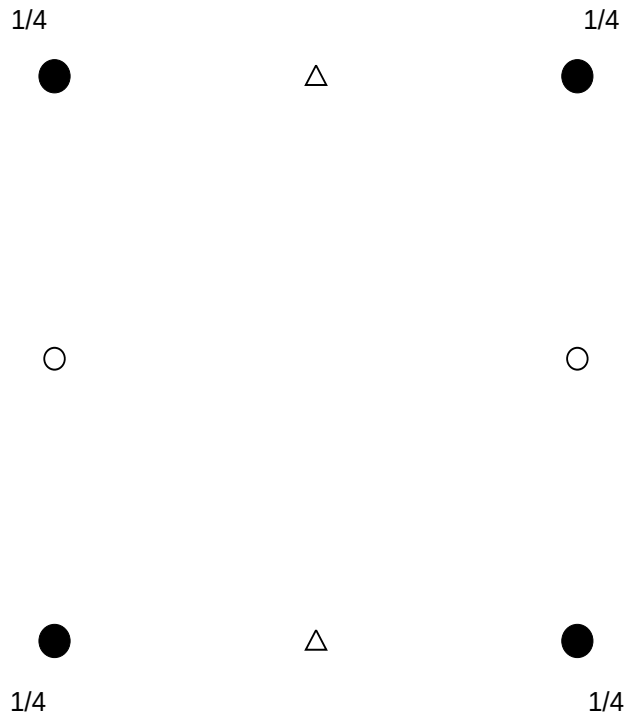

Figure 3.1: Illustration to example 5

By the consistency result, the optimal sample centers b_n converge to the set $\{c^V, c^H\}$ of the population optima. Note that $\rho(c^V, c^H)$ is zero, because for each vertex of the square, the distance to the closest point in c^V is the same as the distance to the closest point in c^H. Theorem 2 does not explain how the optimal sample centers decide which configuration of the optimal population centers they want to be closer to. The proof showed that this selection is governed by a minimization of the stochastic process $\{\nu_n^x \phi(x, a) : a \in C\}$, where C is the collection of population optima. For the example at hand, $\nu_n^x \phi(x, c^V)$ and $\nu_n^x \phi(x, c^H)$ are equal with probability one; the behavior of the optimal sample centers b_n is controlled by lower order terms.

Write approximation (3.18) for the empirical criterion function near c^H and c^V. Note

that no coordinate change is needed because the second derivative of W is nonsingular at both c^H and c^V. The underlying distribution concentrates on a finite number of points, hence the second derivative of W evaluated at its minima equals to the second derivative of the bias-squared function; it is exactly the identity matrix in both cases. Minimize the two quadratic approximations, and compare their minimum values. It follows that, up to smaller order terms, the optimal sample sample centers choose to be close to c^V whenever $(p_{2n} - p_{4n})(p_{1n} - p_{3n})$ is positive. Here I use p_{1n}, p_{2n}, p_{3n}, and p_{4n}, to denote the empirical probabilities assigned to the vertices of the square (the numeration starts from the top left corner and continues in the clock-wise direction).

Note that the same result could have been obtained by simply comparing the sample within cluster sum of squares of two partitions: the first splits the four vertices of the square the same way as do the centers c^H (horizontal split), the second splits the vertices the same way as do the centers c^V (vertical split).

Chapter 4

Nonlinear Least-Squares Estimation

4.1 Introduction

Consider the model where we observe y_i for $i = 1, \ldots, n$ with

$$y_i = f_i(\theta) + u_i \qquad \text{where } \theta \in \Theta. \tag{4.1}$$

The unobserved f_i can be random or deterministic functions. The unobserved errors u_i are random with zero means and finite variances. The index set Θ might be infinite dimensional. Later in the paper it will prove convenient to also consider triangular arrays of observations.

Think of $f(\theta) = (f_1(\theta), \ldots, f_n(\theta))'$ and $u = (u_1, \ldots, u_n)'$ as points in $\mathbb{R}^n$. The model specifies a surface $M_\Theta = \{f(\theta) : \theta \in \Theta\}$ in $\mathbb{R}^n$. The vector of observations $y = (y_1, \ldots, y_n)'$ is a random point in $\mathbb{R}^n$. The least squares estimator (LSE) $\widehat{\theta}_n$ is defined to minimize the distance of y to M_Θ,

$$\widehat{\theta}_n = \operatorname{argmin}_{\theta \in \Theta} |y - f(\theta)|^2,$$

where $|\cdot|$ denotes the usual Euclidean norm on $\mathbb{R}^n$. Many authors have considered the behavior of $\widehat{\theta}_n$ as $n \to \infty$ when the y_i are generated by the model for a fixed θ_0 in Θ.

When the f_i are deterministic, it is natural to express assertions about convergence of $\widehat{\theta}_n$ in terms of the n-dimensional Euclidean distance $\kappa_n(\theta_1, \theta_2) := |f(\theta_1) - f(\theta_2)|$. For example, Jennrich (1969) took Θ to be a compact subset of $\mathbb{R}^p$, the errors $\{u_i\}$ to be iid with zero mean and finite variance, and the f_i to be continuous functions in θ. He proved strong consistency of the least squares estimator under the assumption that $n^{-1}\kappa_n(\theta_1, \theta_2)^2$ converges uniformly to a continuous function that is zero if and only if $\theta_1 = \theta_2$. He also gave conditions for asymptotic normality.

Under similar assumptions Wu (1981, Theorem 1) proved that existence of a consistent estimator for θ_0 implies that

$$\kappa_n(\theta) := \kappa_n(\theta, \theta_0) \to \infty \qquad \text{at each } \theta \neq \theta_0. \tag{4.2}$$

If Θ is finite, the divergence (4.2) is also a sufficient condition for the existence of a consistent estimator (Wu 1981, Theorem 2). His main consistency result (his Theorem 3) may be reexpressed as a general convergence assertion.

Theorem 3 *Suppose the $\{f_i\}$ are deterministic functions indexed by a subset Θ of $\mathbb{R}^p$. Suppose also that $\sup_i \text{var}(u_i) < \infty$ and $\kappa_n(\theta) \to \infty$ at each $\theta \neq \theta_0$. Let S be a bounded subset of $\Theta \backslash \{\theta_0\}$ and let $R_n := \inf_{\theta \in S} \kappa_n(\theta)$. Suppose there exist constants $\{L_i\}$ such that*

(i) $\sup_{\theta \in S} |f_i(\theta) - f_i(\theta_0)| \leq L_i$ *for each i;*

(ii) $|f_i(\theta_1) - f_i(\theta_2)| \leq L_i|\theta_1 - \theta_2|$ *for all $\theta_1, \theta_2 \in S$;*

(iii) $\sum_{i \leq n} L_i^2 = O(R_n^\alpha)$ *for some $\alpha < 4$.*

Then $\mathbb{P}\{\widehat{\theta}_n \notin S \text{ eventually }\} = 1$.

Remark. Assumption (i) implies $\sum_{i\le n} L_i^2 \ge \kappa_n(\theta)^2 \to \infty$ for each θ in S, which forces $R_n \to \infty$.

If Θ is compact and if for each $\theta \neq \theta_0$ there is a neighborhood $S = S_\theta$ satisfying the conditions of the Lemma then $\widehat{\theta}_n \to \theta_0$ almost surely.

Wu's paper was the starting point for several authors. For example, both Lai (1994) and Skouras (2000) generalized Wu's consistency results by taking the functions $f_i(\theta) = f_i(\theta, \omega)$ as random processes indexed by θ. They took the $\{u_i\}$ as a martingale difference sequence, with $\{f_i\}$ a predictable sequence of functions with respect to a filtration $\{\mathcal{F}_i\}$.

Another line of development is typified by the work of Van de Geer (1990) and Van de Geer and Wegkamp (1996). They took $f_i(\theta) = f(x_i, \theta)$, where $\mathcal{F} = \{f_\theta : \Theta\}$ is a set of deterministic functions and the (x_i, u_i) are iid pairs. (In fact they identified Θ with the index set $\mathcal{F}$.) Under a stronger assumption about the errors, they established rates of convergence of $\kappa_n(\widehat{\theta}_n)$ in terms of $\mathcal{L}^2$ entropy conditions on $\mathcal{F}$, using empirical process methods that were developed after Wu's work.

The stronger assumption was that the errors are uniformly subgaussian. In general, we say that a random variable W has a subgaussian distribution if there exists some finite τ such that

$$\mathbb{P}\exp(tW) \le \exp\left(\tfrac{1}{2}\tau^2 t^2\right) \qquad \text{for all } t \in \mathbb{R}.$$

We write $\tau(W)$ for the smallest such τ. Van de Geer assumed that $\sup_i \tau(u_i) < \infty$.

Remark. Notice that we must have $\mathbb{P}W = 0$ when W is subgaussian because the linear term in the expansion of $\mathbb{P}\exp(tW)$ must vanish. When $\mathbb{P}W = 0$, subgaussianity is equivalent to existence of a finite constant β for which $\mathbb{P}\{|W| \ge x\} \le 2\exp(-x^2/\beta^2)$ for all $x \ge 0$.

In our paper we try to bring together the two lines of development. Our main motivation for working on nonlinear least squares was an example presented by Wu (1981,

page 507). He noted that his consistency theorem has difficulties with a simple model,

$$f_i(\theta) = \lambda i^{-\mu} \qquad \text{for } \theta = (\lambda, \mu) \in \Theta, \text{ a compact subset of } \mathbb{R} \times \mathbb{R}^+. \tag{4.3}$$

For example, condition (4.2) does not hold for $\theta_0 = (0, 0)$ at any θ with $\mu > 1/2$. When $\theta_0 = (\lambda_0, 1/2)$, Wu's method fails in a more subtle way, but Van de Geer (1990)'s method would work if the errors satisfied the subgaussian assumption. In Section 4.4, under only second moment assumptions on the errors, we establish weak consistency and a central limit theorem.

The main idea behind all the proofs—ours, as well as those of Wu and Van de Geer—is quite simple. The LSE also minimizes the random function

$$\begin{aligned} G_n(\theta) := |y - f(\theta)|^2 - |u|^2 = \kappa_n(\theta)^2 - 2Z_n(\theta) \\ \text{where } Z_n(\theta) := u' f(\theta) - u' f(\theta_0). \end{aligned} \tag{4.4}$$

In particular, $G_n(\widehat{\theta}_n) \le G_n(\theta_0) = 0$, that is, $\frac{1}{2}\kappa_n(\widehat{\theta}_n)^2 \le Z_n(\widehat{\theta}_n)$. For every subset S of Θ,

$$\mathbb{P}\{\widehat{\theta}_n \in S\} \le \mathbb{P}\{\exists \theta \in S : Z_n(\theta) \ge \tfrac{1}{2}\kappa_n(\theta)^2\} \le 4\mathbb{P} \sup_{\theta \in S} |Z_n(\theta)|^2 / \inf_{\theta \in S} \kappa_n(\theta)^4. \tag{4.5}$$

The final bound calls for a maximal inequality for Z_n.

Our methods for controlling Z_n are similar in spirit to those of Van de Geer. Under her subgaussian assumption, for every class of real functions $\{g_\theta : \theta \in \Theta\}$, the process

$$X(\theta) = \sum\nolimits_{i \le n} u_i g_i(\theta) \tag{4.6}$$

has subgaussian increments. Indeed, if $\tau(u_i) \le \tau$ for all i then

$$\tau \left(X(\theta_1) - X(\theta_2) \right)^2 \le \sum\nolimits_{i \le n} \tau(u_i)^2 \left(g_i(\theta_1) - g_i(\theta_2) \right)^2 \le \tau^2 |g(\theta_1) - g(\theta_2)|^2.$$

That is, the tails of $X(\theta_1) - X(\theta_2)$ are controlled by the n-dimensional Euclidean distance between the vectors $g(\theta_1)$ and $g(\theta_2)$. This property allowed her to invoke a chaining bound (similar to our Theorem 4) for the tail probabilities of $\sup_{\theta \in S} |Z_n(\theta)|$ for various annuli $S = \{\theta : R \le \kappa_n(\theta) < 2R\}$.

Under the weaker second moment assumption on the errors, we apply symmetrization arguments to transform to a problem involving a new process $Z_n^\circ(\theta)$ with conditionally subgaussian increments. We avoid Van de Geer's subgaussianity assumption at the cost of extra Lipschitz conditions on the $f_i(\theta)$, analogous to Assumption (ii) of Theorem 3, which lets us invoke chaining bounds for conditional second moments of $\sup_{\theta \in S} |Z_n^\circ(\theta)|$ for various S.

In Section 4.3 we prove a new consistency theorem (Theorem 5) and a new central limit theorem (Theorem 6, generalizing Wu's Theorem 5) for nonlinear LSEs. More precisely, our consistency theorem corresponds to an explicit bound for $\mathbb{P}\{\kappa_n(\widehat{\theta}_n) \ge R\}$, but we state the result in a form that makes comparison with Theorem 3 easier. Our Theorem does not imply almost sure convergence, but our techniques could easily be adapted to that task. We regard the consistency as a preliminary to the next level of asymptotics and not as an end in itself. We describe the local asymptotic behavior with another approximation result, Theorem 6, which can easily be transformed into a central limit theorem under a variety of mild assumptions on the $\{u_i\}$ errors. For example, in Section 4.4 we apply the Theorem to the model (4.3), to sharpen the consistency result at $\theta_0 = (1, 1/2)$ into the approximation

$$\left(\ell_n^{1/2}(\widehat{\lambda}_n - 1), \ell_n^{3/2}(1 - 2\widehat{\mu}_n) \right) = \sum\nolimits_{i \le n} u_i \zeta'_{i,n} + o_p(1) \tag{4.7}$$

where $\ell_n := \log n$ and

$$\zeta_{i,n} = i^{-1/2}\ell_n^{-1/2} \begin{pmatrix} 2 & -6 \\ -6 & 24 \end{pmatrix} \begin{pmatrix} 2 \\ \ell_i/\ell_n \end{pmatrix}.$$

The sum on the right-hand side of (4.7) is of order $O_p(1)$ when $\sup_i \text{var}(u_i) < \infty$. If the $\{u_i\}$ are also identically distributed, the sum has a limiting multivariate normal distribution.

4.2 Maximal Inequalities

Assumption (ii) of Theorem 3 ensures that the increments $Z_n(\theta_1) - Z_n(\theta_2)$ are controlled by the ordinary Euclidean distance in Θ; we allow for control by more general metrics. Wu invoked a maximal inequality for sums of random continuous processes, a result derived from a bound on the covering numbers for M_θ as a subset of $\mathbb{R}^n$ under the usual Euclidean distance; we work with covering numbers for other metrics.

Definition 6 *Let (T, d) be a metric space. The covering number $N(\delta, S, d)$ is defined as the size of the smallest δ-net for S, that is, the smallest N for which there are points $t_1, \ldots, t_N$ in T with $\min_i d(s, t_i) \le \delta$ for every s in S.*

> **Remark.** The definition is the same for a pseudometric space, that is, a space where $d(\theta_1, \theta_2) = 0$ need not imply $\theta_1 = \theta_2$. In fact, all results in our paper that refer to metric spaces also apply to pseudometric spaces. The slight increase in generality is sometimes convenient when dealing with metrics defined by $\mathcal{L}^p$ norms on functions.

Standard chaining arguments (see, for example, Pollard 1989), give maximal inequalities for processes with subgaussian increments controlled by a metric on the index set.

Theorem 4 *Let $\{W_t : t \in T\}$ be a stochastic process, indexed by a metric space (T, d), with subgaussian increments. Let T_δ be a δ-net for T. Suppose:*

(i) there is a constant K such that $\tau(W_s - W_t) \le K d(s,t)$ for all $s, t \in T$;

(ii) $J_\delta := \int_0^\delta \rho(N(y, S, d))\, dy < \infty$, where $\rho(N) := \sqrt{1 + \log N}$.

Then there is a universal constant c_1 such that

$$\frac{1}{c_1}\sqrt{\mathbb{P}\sup_t |W_t|^2} \le K J_\delta + \rho(N(\delta, T, d)) \max_{s \in T_\delta} \tau(W_s).$$

Remark. We should perhaps work with outer expectations because, in general, there is no guarantee that a supremum of uncountably many random variables is measurable. For concrete examples, such as the one discussed in Section 4.4, measurability can usually be established by routine separability arguments. Accordingly, we will ignore the issue in this paper.

Under the assumption that $\text{var}(u_i) \le \sigma^2$, the X process from (4.6) need not have subgaussian increments. However, it can be bounded in a stochastic sense by a symmetrized process $X^\circ(\theta) := \sum_{i \le n} \epsilon_i u_i g_i(\theta)$, where the $2n$ random variables $\epsilon_1, \ldots, \epsilon_n, u_1, \ldots, u_n$ are mutually independent with $\mathbb{P}\{\epsilon_i = +1\} = 1/2 = \mathbb{P}\{\epsilon_i = -1\}$. In fact, for each subset S of the index set Θ,

$$\mathbb{P}\sup_{\theta \in S} |X(\theta)|^2 \le 4\mathbb{P}\sup_{\theta \in S} |X^\circ(\theta)|^2. \tag{4.8}$$

For a proof see, for example, van der Vaart and Wellner (1996, Lemma 2.3.1). The process X° has conditionally subgaussian increments with

$$\tau_u \left(X^\circ_{\theta_1} - X^\circ_{\theta_2} \right)^2 \le \sum\nolimits_{i \le n} u_i^2 \left(g_i(\theta_1) - g_i(\theta_2) \right)^2. \tag{4.9}$$

The subscript u indicates the conditioning on u.

Corollary 1 *Let S_δ be a δ-net for S and let X be as in (4.6). Suppose*

(i) $\mathbb{P}u_i = 0$ *and* $\text{var}(u_i) \le \sigma^2$ *for* $i = 1, \dots, n$

(ii) there is a metric d for which $J_\delta := \int_0^\delta \rho(N(y, S, d))\, dy < \infty$

(iii) there are constants $L_1, \dots, L_n$ for which

$$|g_i(\theta_1) - g_i(\theta_2)| \le L_i d(\theta_1, \theta_2) \qquad \text{for all } i \text{ and all } \theta_1, \theta_2 \in S$$

(iv) there are constants $b_1, \dots, b_n$ for which $|g_i(\theta)| \le b_i$ for all i and all θ in S.

Then there is a universal constant c_2 such that

$$\mathbb{P} \sup_{\theta \in S} |X_\theta|^2 \le c_2^2 \sigma^2 \left(L J_\delta + B\rho(N(\delta, S, d))\right)^2$$

where $L := \sqrt{\sum_i L_i^2}$ and $B := \sqrt{\sum_i b_i^2}$.

Proof: From (4.9),

$$\tau_u(X^\circ_{\theta_1} - X^\circ_{\theta_2}) \le L_u d(\theta_1, \theta_2) \qquad \text{where } L_u := \sqrt{\textstyle\sum_{i \le n} L_i^2 u_i^2}$$

and

$$\tau_u(X^\circ_\theta) \le B_u := \sqrt{\textstyle\sum_{i \le n} b_i^2 u_i^2}$$

Apply Theorem 4 conditionally to the process X° to bound $\mathbb{P}_u \sup_{\theta \in S} |X^\circ_\theta|^2$. Then invoke inequality (4.8), using the fact that $\mathbb{P}L_u^2 \le \sigma^2 L^2$ and $\mathbb{P}B_u^2 \le \sigma^2 B^2$. □

4.3 Limit Theorems

Inequality (4.5) and Corollary 1, with $g_i(\theta) = f_i(\theta) - f_i(\theta_0)$, give us some probabilistic control over $\widehat{\theta}_n$.

Theorem 5 *Let S be a subset of Θ equipped with a pseudometric d. Let $\{L_i : i = 1, \ldots, n\}$, $\{b_i : i = 1, \ldots, n\}$, and δ be positive constants such that*

(i) $|f_i(\theta_1) - f_i(\theta_2)| \le L_i d(\theta_1, \theta_2)$ *for all* $\theta_1, \theta_2 \in S$

(ii) $|f_i(\theta) - f_i(\theta_0)| \le b_i$ *for all* $\theta \in S$

(iii) $J_\delta := \int_0^\delta \rho\left(N(y, S, d)\right)\, dy < \infty$

Then

$$\mathbb{P}\{\widehat{\theta}_n \in S\} \le 4c_2^2\sigma^2\left(B\rho\left(N(\delta, S, d)\right) + LJ_\delta\right)^2/R^4,$$

where $R := \inf\{\kappa(\theta) : \theta \in S\}$, and $L^2 = \sum_i L_i^2$, and $B^2 := \sum_i b_i^2$.

The Theorem becomes more versatile in its application if we partition S into a countable union of subsets S_k, each equipped with its own pseudometric and Lipschitz constants. We then have $\mathbb{P}\{\widehat{\theta}_n \in \cup_k S_k\}$ smaller than a sum over k of bounds analogous to those in the Theorem. As shown in Section 4.4, this method works well for the Wu example if we take $S_k = \{\theta : R_k \le \kappa_n(\theta) < R_{k+1}\}$, for an $\{R_k\}$ sequence increasing geometrically.

A similar appeal to Corollary 1, with the $g_i(\theta)$ as partial derivatives of $f_i(\theta)$ functions, gives us enough local control over Z_n to go beyond consistency. To accommodate the application in Section 4.4, we change notation slightly by working with a triangular array: for each n,

$$y_{in} = f_{in}(\theta_0) + u_{in}, \qquad \text{for } i = 1, 2, \ldots, n,$$

where the $\{u_{in} : i = 1, \ldots, n\}$ are unobserved independent random variables with mean zero and variance bounded by σ^2.

Theorem 6 *Suppose $\widehat{\theta}_n \to \theta_0$ in probability, with θ_0 an interior point of Θ, a subset of $\mathbb{R}^p$. Suppose also:*

(i) Each f_{in} is continuously differentiable in a neighborhood $\mathcal{N}$ of θ_0 with derivatives $D_{in}(\theta) = \partial f_{in}(\theta)/\partial\theta$.

(ii) $\gamma_n^2 := \sum_{i \le n} |D_{in}(\theta_0)|^2 \to \infty$ as $n \to \infty$.

(iii) There are constants $\{M_{in}\}$ with $\sum_{i \le n} M_{in}^2 = O(\gamma_n^2)$ and a metric d on $\mathcal{N}$ for which $|D_{in}(\theta_1) - D_{in}(\theta_2)| \le M_{in} d(\theta_1, \theta_2)$ for $\theta_1,\ \theta_2 \in \mathcal{N}$.

(iv) The smallest eigenvalue of the matrix $V_n = \gamma_n^{-2} \sum_{i \le n} D_{in}(\theta_0) D_{in}(\theta_0)'$ is bounded away from zero for n large enough.

(v) $\int_0^1 \rho\left(N(y, \mathcal{N}, d)\right)\, dy < \infty$

(vi) $d(\theta, \theta_0) \to 0$ as $\theta \to \theta_0$.

Then $\kappa_n(\widehat{\theta}_n) = O_p(1)$ and

$$\gamma_n(\widehat{\theta}_n - \theta_0) = \sum\nolimits_{i \le n} \xi_{i,n} u_{in} + o_p(1) = O_p(1).$$

where $\xi_{i,n} = \gamma_n^{-1} V_n^{-1} D_{in}(\theta_0)$.

Proof: Let D be the $p \times n$ matrix with ith column $D_{in}(\theta_0)$, so that $\gamma_n^2 = \mathrm{trace}(DD')$ and $V_n = \gamma_n^{-2} DD'$. The main idea of the proof is to replace $f(\theta)$ by $f(\theta_0) + D'(\theta - \theta_0)$, thereby approximating $\widehat{\theta}_n$ by the least-squares solution

$$\overline{\theta}_n := \theta_0 + (DD')^{-1} Du = \operatorname*{argmin}_{\theta \in \mathbb{R}^p} |y - f(\theta_0) - D'(\theta - \theta_0)|.$$

To simplify notation, assume with no loss of generality, that $f(\theta_0) = 0$ and $\theta_0 = 0$. Also, drop extra n subscripts when the meaning is clear. The assertion of the Theorem is that $\widehat{\theta}_n = \overline{\theta}_n + o_p(\gamma_n^{-1})$.

Without loss of generality, suppose the smallest eigenvalue of V_n is larger than a fixed constant $c_0^2 > 0$. Then

$$\gamma_n^2 = \text{trace}(DD') \geq \sup_{|t|\leq 1} |D't|^2 \geq \inf_{|t|\leq 1} |D't|^2 = c_0^2 \gamma_n^2,$$

from which it follows that

$$c_0|t| \leq |D't|/\gamma_n \leq |t| \qquad \text{for all } t \in \mathbb{R}^p. \tag{4.10}$$

Similarly, $\mathbb{P}|Du|^2 = \text{trace}\left(D\mathbb{P}(uu')D'\right) \leq \sigma^2\gamma_n^2$, implying that $|Du| = O_p(\gamma_n)$ and

$$\overline{\theta}_n = \gamma_n^{-2} V_n^{-1} Du = O_p(\gamma_n^{-1}).$$

In particular, $\mathbb{P}\{\overline{\theta}_n \in \mathcal{N}\} \to 1$, because θ_0 is an interior point of Θ. Note also that

$$\mathbb{P}|\textstyle\sum_{i\leq n} \xi_i u_i|^2 \leq \sigma^2 \text{trace}(\sum_{i\leq n} \xi_i \xi_i') = \sigma^2 \text{trace}(V_n^{-1}) < \infty.$$

Consequently $\sum_{i\leq n} \xi_i u_i = O_p(1)$.

From the assumed consistency, we know that there is a sequence of balls $\mathcal{N}_n \subseteq \mathcal{N}$ that shrink to $\{0\}$ for which $\mathbb{P}\{\widehat{\theta}_n \in \mathcal{N}_n\} \to 1$. From (vi) and (v), it follows that both $r_n := \sup\{d(\theta, 0) : \theta \in \mathcal{N}_n\}$ and $J_{r_n} = \int_0^{r_n} \rho\left(N(y, \mathcal{N}, d)\right) dy$ converge to zero as $n \to \infty$.

The $n \times 1$ remainder vector $R(\theta) := f(\theta) - D'\theta$ has ith component

$$R_i(\theta) = f_i(\theta) - D_i(0)'\theta = \theta' \int_0^1 D_i(t\theta) - D_i(0)\, dt. \tag{4.11}$$

Uniformly in the neighborhood $\mathcal{N}_n$ we have

$$|R(\theta)| \leq |\theta| \left(\sum\nolimits_{i \leq n} M_{in}^2 \right)^{1/2} r_n = o(|\theta|\gamma_n),$$

which, together with the upper bound from inequality (4.10), implies

$$|f(\theta)|^2 = |D'\theta|^2 + o(\gamma_n^2|\theta|^2) = O(\gamma_n^2|\theta|^2) \qquad \text{as } |\theta| \to 0. \tag{4.12}$$

In the neighborhood $\mathcal{N}_n$, via (4.11) we also have,

$$|u'R(\theta)| \leq |\theta| \sup_{s \in \mathcal{N}_n} | \sum\nolimits_i u_i \left(D_i(s) - D_i(0) \right) |.$$

From Corollary 1 with $g_i(\theta) = D_i(\theta) - D_i(0)$ deduce that

$$\mathbb{P} \sup_{s \in \mathcal{N}_n} | \sum\nolimits_i u_i \left(D_i(s) - D_i(0) \right) |^2 \leq c_2^2 \sigma^2 J_{r_n}^2 \sum\nolimits_i M_{in}^2 = o(\gamma_n^2),$$

which implies

$$|u'R(\theta)| = o_p(\gamma_n|\theta|) \qquad \text{uniformly for } \theta \in \mathcal{N}_n. \tag{4.13}$$

Approximations (4.12) and (4.13) give us uniform approximations for the criterion functions in the shrinking neighborhoods $\mathcal{N}_n$:

$$\begin{aligned} G_n(\theta) &= |u - f(\theta)|^2 - |u|^2 \\ &= -2u' f(\theta) + |f(\theta)|^2 \\ &= -2u' D'\theta + |D'\theta|^2 + o_p(\gamma_n|\theta|) + o_p(\gamma_n^2|\theta|^2) \\ &= |u - D'\overline{\theta}_n|^2 - |u|^2 + |D'(\theta - \overline{\theta}_n)|^2 + o_p(\gamma_n|\theta|) + o_p(\gamma_n^2|\theta|^2). \end{aligned} \tag{4.14}$$

The uniform smallness of the remainder terms lets us approximate G_n at random points that are known to lie in $\mathcal{N}_n$.

The rest of the argument is similar to that of Chernoff (1954). When $\widehat{\theta}_n \in \mathcal{N}_n$ we have $G_n(\widehat{\theta}_n) \le G_n(0)$, implying

$$|D'(\widehat{\theta}_n - \overline{\theta}_n)|^2 + o_p(\gamma_n|\widehat{\theta}_n|) + o_p(\gamma_n^2|\widehat{\theta}_n|^2) \le |D'\overline{\theta}_n|^2.$$

Invoke (4.10) again, simplifying the last approximation to

$$c_0^2|\gamma_n\widehat{\theta}_n - \gamma_n\overline{\theta}_n|^2 \le O_p(1) + o_p\left(|\gamma_n\widehat{\theta}_n| + |\gamma_n\widehat{\theta}_n|^2\right).$$

It follows that $|\widehat{\theta}_n| = O_p(\gamma_n^{-1})$ and, via (4.12),

$$\kappa_n(\widehat{\theta}_n) = |f(\widehat{\theta}_n)| = O_p(1).$$

We may also assume that $\mathcal{N}_n$ shrinks slowly enough to ensure that $\mathbb{P}\{\overline{\theta}_n \in \mathcal{N}_n\} \to 1$. When both $\widehat{\theta}_n$ and $\overline{\theta}_n$ lie in $\mathcal{N}_n$ the inequality $G_n(\widehat{\theta}_n) \le G_n(\overline{\theta}_n)$ and approximation (4.14) give

$$|D'(\widehat{\theta}_n - \overline{\theta}_n)|^2 + o_p(1) \le o_p(1).$$

It follows that $\widehat{\theta}_n = \overline{\theta}_n + o_p(\gamma_n^{-1})$. $\square$

Remark. If the errors are iid and $\max|\xi_{i,n}| = o(1)$ then the distribution of $\sum_{i\le n}\xi_{i,n}u_{in}$ is asymptotically $N\big(0, \sigma^2 V_n^{-1}\big)$.

4.4 Analysis of Model (4.3): Wu's Example

The results in this section illustrate the work of our limit theorems in a particular case where Wu's method fails. We prove both consistency and a central limit theorem for the model (4.3) with $\theta_0 = (\lambda_0, 1/2)$. In fact, without loss of generality, $\lambda_0 = 1$.

As before, let $\ell_n = \log n$. Remember $\theta = (\lambda, \mu)$ with $a \in \mathbb{R}$ and $0 \le \mu \le C_\mu$ for a finite constant C_μ greater than $1/2$, which ensures that $\theta_0 = (1, 1/2)$ is an interior point of the parameter space. Taking $C_\mu = 1/2$ would complicate the central limit theorem only slightly. The behavior of $\widehat{\theta}_n$ is determined by the behavior of the function

$$G_n(\gamma) := \textstyle\sum_{i\le n} i^{-1+\gamma} \qquad \text{for } \gamma \le 1,$$

or its standardized version

$$g_n(\beta) := G_n(\beta/\ell_n)/G_n(0) = \textstyle\sum_{i\le n}\left(i^{-1}/G_n(0)\right)\exp\left(\beta\ell_i/\ell_n\right),$$

which is the moment generating function of the probability distribution that puts mass $i^{-1}/G_n(0)$ at ℓ_i/ℓ_n, for $i = 1, \ldots, n$. For large n, the function g_n is well approximated by the increasing, nonnegative function

$$g(\beta) = \begin{cases} (e^\beta - 1)/\beta & \text{for } \beta \ne 0 \\ 1 & \text{for } \beta = 0 \end{cases},$$

the moment generating function of the uniform distribution on $(0,1)$. More precisely, comparison of the sum with the integral $\int_1^n x^{-1+\gamma}\,dx$ gives

$$G_n(\gamma) \;= \ell_n g(\gamma\ell_n) + r_n(\gamma) \qquad \text{with } 0 \le r_n(\gamma) \le 1 \text{ for } \gamma \le 1. \tag{4.15}$$

The distributions corresponding to both g_n and g are concentrated on $[0,1]$. Both functions have the properties described in the following lemma.

Lemma 10 *Suppose $h(\gamma) = P\exp(\gamma x)$, the moment generating function of a probability distribution concentrated on $[0,1]$. Then*

(i) $\log h$ *is convex*

(ii) $h(\gamma)^2/h(2\gamma)$ *is unimodal: increasing for $\gamma < 0$, decreasing for $\gamma > 0$, achieving its maximum value of 1 at $\gamma = 0$*

(iii) $h'(\gamma) \le h(\gamma)$

Proof: Assertion (i) is just the well known fact that the logarithm of a moment generating function is convex. Thus h'/h, the derivative of $\log h$, is an increasing function, which implies (ii) because

$$\frac{d}{d\gamma}\log\left(\frac{h(\gamma)^2}{h(2\gamma)}\right) = 2\frac{h'(\gamma)}{h(\gamma)} - 2\frac{h'(2\gamma)}{h(2\gamma)}.$$

Property (iii) comes from the representation $h'(\gamma) = P\left(xe^{\gamma x}\right)$. □

Remark. Direct calculation shows that $g(\gamma)^2/g(2\gamma)$ is a symmetric function.

Reparametrize by putting $\beta = (1-2\mu)\ell_n$, with $(1-2C_\mu)\ell_n \le \beta \le \ell_n$, and $\alpha = \lambda\sqrt{G_n(\beta/\ell_n)}$. Notice that $|f(\theta)| = |\alpha|$ and that θ_0 corresponds to $\alpha_0 = \sqrt{G_n(0)} \approx \sqrt{\ell_n}$ and $\beta_0 = 0$. Also

$$f_i(\theta) = \alpha\nu_i(\beta/\ell_n) \qquad \text{where} \quad \nu_i(\gamma) := i^{-1/2}\exp(\gamma\ell_i/2)/\sqrt{G_n(\gamma)},$$

and

$$\kappa_n(\theta)^2 = G_n(0)\left(\lambda^2 g_n(\beta) - 2\lambda g_n(\beta/2) + 1\right). \tag{4.16}$$

We define $\nu_i := \sup_{\gamma \le 1} \nu_i(\gamma)$.

Lemma 11 *For all* (α, β) *corresponding to* $\theta = (\lambda, \mu) \in \mathbb{R} \times [0, C_\mu]$*:*

(i) $\kappa_n(\theta) - \sqrt{G_n(0)} \le |\alpha| \le \kappa_n(\theta) + \sqrt{G_n(0)}$

(ii) $\sum_{i \le n} \nu_i^2 = O\left(\log\log n\right)$

(iii) $|d\nu_i(\beta/\ell_n)/d\beta| \le \frac{1}{2}\nu_i(\beta/\ell_n)$

(iv) $|f_i(\alpha_1, \beta_1) - f_i(\alpha_2, \beta_2)| \le \left(|\alpha_1 - \alpha_2| + \frac{1}{2}|\alpha_2||\beta_1 - \beta_2|\right)\nu_i$

(v) $|f_i(\theta) - f_i(\theta_0)| \le i^{-1/2} + |\alpha|\nu_i$

Proof: Inequalities (i) and (v) follow from the triangle inequality.

For inequality (ii), first note that $\nu_1^2 \le 1$. For $i \ge 2$, separate out contributions from three ranges:

$$\nu_i^2 = \max\left(\sup_{1 \ge \gamma \ge 1/\ell_n} \nu_i(\gamma)^2, \sup_{|\gamma| < 1/\ell_n} \nu_i(\gamma)^2, \sup_{\gamma \le -1/\ell_n} \nu_i(\gamma)^2\right).$$

For $\gamma \ge 1/\ell_n$, invoke (4.15) to get a tractable upper bound:

$$\nu_i(\gamma)^2 \le i^{-1}\frac{\exp(\gamma\ell_i)}{\ell_n g(\gamma\ell_n)} \le i^{-1}\gamma\frac{\exp(\gamma\ell_i)}{\exp(\gamma\ell_n) - 1} \le i^{-1}\frac{\exp\left(\log\gamma + \gamma\log(i/n)\right)}{1 - e^{-1}}.$$

The last expression achieves its maximum over $[1/\ell_n, 1]$ at

$$\gamma_0 := \begin{cases} 1/\log(n/i) & \text{if } 1 \le i \le n/e \\ 1 & \text{if } n/e \le i \le n \end{cases},$$

which gives

$$\sup_{1\geq\gamma\geq 1/\ell_n} \nu_i(\gamma)^2 \leq \frac{(e-1)^{-1}}{n} H\left(\frac{i\wedge(n/e)}{n}\right) \qquad \text{where } H(x) := 1/\left(x\log(1/x)\right). \tag{4.17}$$

Similarly, if $-1 < \gamma\ell_n < 1$,

$$\nu_i(\gamma)^2 \leq \frac{\exp(\gamma\ell_i)}{i\ell_n g(\gamma\ell_n)} \leq \frac{\exp(\ell_i/\ell_n)}{i\ell_n g(-1)} \leq \frac{e/g(-1)}{i\ell_n}.$$

The last term is smaller than a constant multiple of the bound from (4.17). Finally, if $-\gamma = \delta \geq 1/\ell_n$ and $i \geq 2$ then

$$\nu_i(\gamma)^2 \leq i^{-1}\delta\frac{\exp(-\delta\ell_i)}{1-\exp(-\delta\ell_n)} \leq i^{-1}\frac{\exp\left(\log\delta - \delta\ell_i\right)}{1-e^{-1}} \leq \frac{e^{-1}/(1-e^{-1})}{i\ell_i}.$$

In summary, for some universal constant C,

$$\nu_i^2 \leq C\max\left(n^{-1}H\left(\frac{i\wedge(n/e)}{n}\right), \frac{1}{i\log i}\right) \qquad \text{if } 2\leq i\leq n.$$

Bounding sums by integrals we thus have

$$C^{-1}\sum_{i=2}^{n}\nu_i^2 \leq \int_{1/n}^{1/e} H(x)\,dx + H(1/e)/n + \int_2^n \left(x\log x\right)^{-1} dx = O\left(\log\log n\right).$$

For (iii) note that

$$2\frac{d}{d\beta}\nu_i(\beta/\ell_n) = 2\frac{d}{d\beta}\exp\left(\tfrac{1}{2}\beta\ell_i/\ell_n\right)\left(G_n(0)g_n(\beta)\right)^{-1/2} = \left(\frac{\ell_i}{\ell_n} - \frac{g_n'(\beta)}{g_n(\beta)}\right)\nu_i(\beta),$$

which is bounded in absolute value by $\nu_i(\beta)$ because $0 \leq g_n'(\beta) \leq g_n(\beta)$.

For (iv)

$$\begin{aligned} |f_i(\alpha_1,\beta_1) - f_i(\alpha_2,\beta_2)| &\le |(\alpha_1-\alpha_2)\nu_i(\beta_1/\ell_n)| + |\alpha_2||\nu_i(\beta_1/\ell_n) - \nu_i(\beta_2/\ell_n)| \\ &\le |(\alpha_1-\alpha_2)|\nu_i + |\alpha_2||(\beta_1-\beta_2)|\tfrac{1}{2}\nu_i, \end{aligned}$$

the bound for the second term coming from the mean-value theorem and (iii). □

Lemma 12 *For $\epsilon > 0$, let $\mathcal{N}_\epsilon = \{\theta : \max\left(|\lambda - 1|, |\beta|\right) \ge \epsilon$. If ϵ is small enough, there exists a constant $C_\epsilon > 0$ such that $\inf\{\kappa_n(\theta) : \theta \notin \mathcal{N}_\epsilon\} \ge C_\epsilon\sqrt{\ell_n}$ when n is large enough.*

Proof: Suppose $|\beta| \ge \epsilon$. Remember that $G_n(0) \ge \ell_n$. Minimize over λ the lower bound (4.16) for $\kappa_n(\theta)^2$ by choosing $\lambda = g_n(\beta/2)/g_n(\beta)$, then invoke Lemma 10(ii).

$$\frac{\kappa_n(\theta)^2}{\ell_n} \ge 1 - \frac{g_n(\beta/2)^2}{g_n(\beta)} \ge 1 - \max\left(\frac{g_n(\epsilon/2)^2}{g_n(\epsilon)}, \frac{g_n(-\epsilon/2)^2}{g_n(-\epsilon)}\right) \to 1 - \frac{g(\epsilon/2)^2}{g(\epsilon)} > 0.$$

If $|\beta| \le \epsilon$ and ϵ is small enough to make $(1-\epsilon)e^{\epsilon/2} < 1 < (1+\epsilon)e^{-\epsilon/2}$, use

$$\kappa_n(\theta)^2 = \textstyle\sum_{i\le n} i^{-1}\left(\lambda\exp(\beta\ell_i/2\ell_n) - 1\right)^2.$$

If $\lambda \ge 1+\epsilon$ bound each summand from below by $i^{-1}((1+\epsilon)e^{-\epsilon/2} - 1)^2$. If $\lambda \le 1-\epsilon$ bound each summand from below by $i^{-1}(1-(1-\epsilon)e^{\epsilon/2})^2$. □

4.4.1 Consistency

On the annulus $S_R := \{R \le \kappa_n(\theta) < 2R\}$ we have

$$\begin{aligned} |a| &\le K_R := 2R + \sqrt{G_n(0)} \\ |f_i(\theta_1) - f_i(\theta_2)| &\le K_R\nu_i d_R(\theta_1,\theta_2) \\ &\qquad \text{where } d_R(\theta_1,\theta_2) := |\alpha_1-\alpha_2|/K_R + \tfrac{1}{2}|\beta_1-\beta_2| \\ |f_i(\theta) - f_i(\theta_0)| &\le b_i := i^{-1/2} + K_R\nu_i. \end{aligned}$$

Note that

$$\sum_{i\le n}\left(i^{-1/2}+K_R\nu_i\right)^2 = O(\ell_n + K_R^2\log\ell_n) = O(K_R^2\mathfrak{L}_n) \qquad \text{where } \mathfrak{L}_n := \log\log n.$$

The rectangle $\{|\alpha| \le K_R, |\beta| \le c\ell_n\}$ can be partitioned into $O(y^{-1}\ell_n/y)$ subrectangles of d_R-diameter at most y. Thus $N(y, S_R, d_R) \le C_0\ell_n/y^2$ for a constant C_0 that depends only on C_μ, which gives

$$\int_0^1 \rho\left(N(y,S_R,d_R)\right)\,dy = O\left(\sqrt{\mathfrak{L}_n}\right).$$

Apply Theorem 5 with $\delta = 1$ to conclude that

$$\mathbb{P}\{\widehat{\theta}_n \in S_R\} \le C_1K_R^2\mathfrak{L}_n^2/R^4 \le C_2(R^2+\ell_n)\mathfrak{L}_n^2/R^4.$$

Put $R = C_3 2^k(\ell_n\mathfrak{L}_n^2)^{1/4}$ then sum over k to deduce that

$$\mathbb{P}\{\kappa_n(\widehat{\theta}_n) \ge C_3(\ell_n\mathfrak{L}_n^2)^{1/4}\} \le \epsilon \qquad \text{eventually}$$

if the constant C_3 is large enough. That is $\kappa_n(\widehat{\theta}_n) = O_p\left((\ell_n\mathfrak{L}_n^2)^{1/4}\right)$ and, via Lemma 12,

$$|\widehat{\lambda}_n - 1| = o_p(1) \qquad \text{and} \qquad 2\ell_n|\widehat{\mu}_n - \mu_0| = |\widehat{\beta}| = o_p(1).$$

4.4.2 Central Limit Theorem

This time work with the (λ,β) reparametrization, with

$$\begin{aligned} f_i(\lambda,\beta) &= \lambda i^{-1/2+\beta/2\ell_n} \\ D_i(\lambda,\beta)' = \left(\tfrac{\partial f_i(\lambda,\beta)}{\partial\lambda}, \tfrac{\partial f_i(\lambda,\beta)}{\partial\beta}\right) &= \left(1/\lambda, \ell_i/2\ell_n\right) f_i(\lambda,\beta) \end{aligned}$$

and $\theta_0 = (\lambda_0, \beta_0) = (1, 0)$. Take d as the usual two-dimensional Euclidean distance in the (λ, β) space. For simplicity of notation, we omit some n subscripts, even though the relationship between θ and (λ, β) changes with n.

We have just shown that the LSE $(\widehat{\lambda}_n, \widehat{\beta}_n)$ is consistent.

Comparison of sums with analogous integrals gives the approximations

$$\sum_{i \le n} i^{-1} \ell_i^{p-1} = \ell_n^p / p + r_p \qquad \text{with } |r_p| \le 1 \text{ for } p = 0, 1, 2, \ldots. \tag{4.18}$$

In consequence,

$$\gamma_n^2 = \sum\nolimits_{i \le n} |D_i(\lambda_0, \beta_0)|^2 = \sum\nolimits_{i \le n} i^{-1} \left(1 + \ell_i^2 / 4\ell_n^2\right) = \tfrac{13}{12} \ell_n + O(1)$$

and

$$V_n = \gamma_n^{-2} \textstyle\sum_{i \le n} i^{-1} \begin{pmatrix} 1 & \ell_i / 2\ell_n \\ \ell_i / 2\ell_n & \ell_i^2 / 4\ell_n^2 \end{pmatrix} = V + O(1/\ell_n) \qquad \text{where } V = \tfrac{1}{13} \begin{pmatrix} 12 & 3 \\ 3 & 1 \end{pmatrix}.$$

The smaller eigenvalue of V_n converges to the smaller eigenvalue of the positive definite matrix V, which is strictly positive.

Within the neighborhood $\mathcal{N}_\epsilon := \{\max\left(|\lambda - 1|, |\beta|\right) \le \epsilon\}$, for a fixed $\epsilon \le 1/2$, both $|f_i(\lambda, \beta)|$ and $|D_i(\lambda, \beta)|$ are bounded by a multiple of $i^{-1/2}$. Thus

$$|D_i(\theta_1) - D_i(\theta_1)| \le \left|\lambda_1^{-1} - \lambda_2^{-1}\right| \, |f_i(\theta_1)| + 3|f_i(\theta_1) - f_i(\theta_2)| \le C_\epsilon i^{-1/2} d(\theta_1, \theta_2).$$

That is, we may take M_i as a multiple of $i^{-1/2}$, which gives $\sum_{i \le n} M_i^2 = O(\ell_n)$.

All the conditions of Theorem 6 are satisfied. We have

$$\sqrt{\ell_n}(\widehat{\lambda_n} - 1, \widehat{\beta_n}) = \tfrac{12}{13} \textstyle\sum_{i \le n} u_i i^{-1/2} \ell_n^{-1/2} (1, \ell_i / 2\ell_n) V^{-1} + o_p(1).$$

Appendix A

Some Empirical Process Tools

Denote the $L_2(P)$ norm by $\|\cdot\|_2$. The following theorem adapts several general results in empirical process theory to handle a particular class of functions that appears throughout this thesis.

Theorem 7 *Let $\mathcal{F}$ be a class of functions that are defined on $\mathbb{R}^d$ and take values in $\mathbb{R}^{d'}$. Let N be an integer. Suppose that each function in $\mathcal{F}$ can be written as a sum of at most N members of the class $\mathcal{G}$ with the following properties:*

- *There exists a function G, such that $PG^2 < \infty$ and $|g| < G$ for every g in $\mathcal{G}$.*
- *Every function in $\mathcal{G}$ is of the form LQ, where L is a linear function and Q is a convex region expressible as an intersection of at most N open or closed half spaces.*

Then

(i) Class $\mathcal{F}$ is Glivenko-Cantelli, i.e.

$$\sup_{f\in\mathcal{F}} |P_n f - Pf| \to 0$$

almost surely.

(ii) There exists a constant C, such that

$$P \sup_{f \in \mathcal{F}} |\nu_n f|^2 < C$$

for all n.

*(iii) Empirical process $\{\nu_n f : f \in \mathcal{F}\}$ is **stochastically equicontinuous** at zero, i.e. for every $\epsilon > 0$ and $\eta > 0$ there exists a $\delta > 0$ for which*

$$\limsup_n \mathbb{P}\{\sup_{[\delta]} |\nu_n(f-g)| > \eta\} < \epsilon,$$

where $[\delta] = \{(f,g) : f, g \in \mathcal{F} \quad \text{and} \quad \|f-g\|_2 \leq \delta\}$.

(iv) Class $\mathcal{F}$ satisfies the functional central limit theorem, i.e.

$$\{\nu_n f : f \in \mathcal{F}\} \rightsquigarrow \{Y(f) : f \in \mathcal{F}\}$$

as random elements of the space of all bounded $\mathbb{R}^{d'}$-valued functions on $\mathcal{F}$ equipped with the uniform norm. The limit process Y has joint normal finite-dimensional distributions with zero means and covariance matrix

$$PY(f)Y(g)' = P(fg') - (Pf)(Pg)'.$$

Each sample path of Y is bounded and uniformly continuous with respect to the L_2 seminorm on $\mathcal{F}$.

See, for example, Pollard (1984) or van der Vaart and Wellner (1996) for proofs of statements (i), (iii), and (iv). A proof of statement (ii) is given in Pollard (1989).

Bibliography

Bartlett, P. L., T. Linder, and G. Lugosi (1998). The minimax distortion redundancy in empirical quantizer design. *IEEE Transactions on Information Theory 44*, 1802–1813.

Chernoff, H. (1954). On the distribution of the likelihood ratio. *Annals of Mathematical Statistics 25*, 573–578.

Devroye, L., L. Györfi, and G. Lugosi (1996). *A Probabilistic Theory of Pattern Recognition*. New York: Springer-Verlag.

Dudley, R. M. (1999). *Uniform Central Limit Theorems*. Cambridge University Press.

Graf, S. and H. Luschgy (2000). *Foundations of Quantization for Probability Distributions*, Volume 1730 of *Springer Lecture Notes in Mathematics*. Berlin: Springer-Verlag.

Hartigan, J. (1975). *Clustering Algorithms*. New York: Wiley.

Hartigan, J. (1978). Asymptotic distributions for clustering criteria. *Annals of Statistics 6*, 117–131.

Jennrich, R. I. (1969). Asymptotic properties of non-linear least squares estimators. *Annals of Mathematical Statistics 40*, 633–643.

Komlós, J., P. Major, and G. Tusnády (1975). An approximation of partial sums of independent rv-s, and the sample df. I. *Zeitschrift für Wahrscheinlichkeitstheorie*

und Verwandte Gebiete 32, 111–131.

Lai, T. L. (1994). Asymptotic properties of nonlinear least-squares estimates in stochastic regression models. *Annals of Statistics 22*, 1917–1930.

Linder, T., G. Lugosi, and K. Zeger (1994). Rates of convergence in the source coding theorem, in empirical quantizer design, and in universal lossy source coding. *IEEE Transactions on Information Theory 40*, 1728–1740.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In L. Le Cam and J. Neyman (Eds.), *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, Volume 1, Berkeley, pp. 281–297. University of California Press.

Pollard, D. (1981). Strong consistency of k-means clustering. *Annals of Statistics 9*, 135–140.

Pollard, D. (1982a). A central limit theorem for k-means clustering. *Annals of Probability 10*, 919–926.

Pollard, D. (1982b). Quantization and the method of k-means. *IEEE Transactions on Information Theory 28*, 199–205.

Pollard, D. (1984). *Convergence of Stochastic Processes*. New York: Springer.

Pollard, D. (1989). Asymptotics via empirical processes (with discussion). *Statistical Science 4*, 341–366.

Pollard, D. and P. Radchenko (2003). Nonlinear least-squares estimation. Available at http://pantheon.yale.edu/˜pvr4/.

Serinko, R. J. and G. J. Babu (1992). Weak limit theorems for univariate k-mean clustering under a nonregular condition. *Journal of Multivariate Analysis 41*, 273–296.

Shorack, G. and J. Wellner (1986). *Empirical Processes with Applications to Statistics*. New York: Wiley.

Skouras, K. (2000). Strong consistency in nonlinear stochastic regression models. *Annals of Statistics 28,* 871–879.

Van de Geer, S. (1990). Estimating a regression function. *Annals of Statistics 18,* 907–924.

Van de Geer, S. and M. Wegkamp (1996). Consistency for the least squares estimator in nonparametric regression. *Annals of Statistics 24,* 2513–2523.

van der Vaart, A. W. and J. A. Wellner (1996). *Weak Convergence and Empirical Process: With Applications to Statistics*. Springer-Verlag.

Wu, C.-F. (1981). Asymptotic theory of nonlinear least squares estimation. *Annals of Statistics 9,* 501–513.