ABSTRACT

Title of Dissertation: A VARIATIONAL APPROACH TO CLUSTERING WITH LIPSCHITZ DECISION FUNCTIONS

Xiaoyu Zhou
Doctor of Philosophy, 2023

Dissertation Directed by: Professor Eric Slud
Department of Mathematics

This dissertation proposes an objective function based clustering approach using Lipschitz functions to represent the clustering function. We establish some mathematical properties including two optimality conditions and a uniqueness result; some statistical properties including two consistency results; and some computational development. This work is a step forward building upon existing work about Lipschitz classifiers to proceed from classification to clustering, also covering more theoretical and computational aspects. The mathematical contents strongly suggest further future analysis of the method. The general objective function might be of independent interest.

A VARIATIONAL APPROACH TO CLUSTERING WITH LIPSCHITZ DECISION FUNCTIONS

by

Xiaoyu Zhou

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2023

Advisory Committee:
Professor Eric Slud, Chair/Advisor
Professor Wojciech Czaja
Professor Vince Lyzinski
Professor Antonio De Rosa
Professor Tianzhou Ma, Dean's Representative

Acknowledgments

First, I would like to thank my advisor, Professor Eric Slud, for his constant support and encouragement in the past four years. I was always filled with motivation and fruitful thoughts after each weekly meeting. When everything went online during the pandemic, he was always very supportive and made sure that I did not feel isolated. This dissertation would be impossible without his patient guidance, as well as careful reading and commenting on my dozens of writings. I also want to thank my other committee members, Professor Wojciech Czaja, Vince Lyzinski, Antonio De Rosa, Tianzhou Ma for helpful feedback on the dissertation from their own expertise, and providing the flexibility for a summer defense. Thanks to Cristina Garcia, Jemma Natanson, and Bijoya Chakraborty for their assistance with many administrative obstacles in my last year of PhD which has seen several changes of plans. Lastly, I thank my parents for their unconditional support.

Table of Contents

List of Notations

Generic notations

g	generic notation for clustering function, or a feasible point of (1.5)
$I(g)$	main objective function/energy functional
g^*	optimal point of (1.5) or one optimal point if not unique
$L(g)$	(global) Lipschitz constant of g
Ω	support of P_X
U, U_1, U_2	level set and lower, upper level set at $1/2$ of g, or g^*
$P[f]$	$E_P[f] = \int f dP$

Notations under model C1

S_k	kth cluster
π_k	proportion of kth cluster
α_0	$\max_k \pi_k$, maximum proportion
L_0	$\frac{1}{d(S_1,S_2)}$
$\tilde{g}$	a Lipschitz extension of function g such that $g\|_{S_1} = 0, g\|_{S_2} = 1$

Chapter 1: Introduction and preliminary result

1.1 Preliminary literature review

Clustering is one of the most widely used techniques for exploratory data analysis in many disciplines. From a statistical point of view, data points are drawn from an underlying probability distribution, and the aim is to classify them into homogeneous groups, typically when no external information is given. For historical references, we refer to [24, 28].

Classical methods as exploratory data analysis tools Classical clustering methods include K-means clustering [24] and hierarchical clustering [28]. While theoretical analysis is often rudimentary in the early references, there has been good progress in the last few decades: for K-means, we refer to the elegant work of David Pollard [43]; for hierarchical clustering, we refer to [54] where the analysis is centered around the tree structure.

Modern algorithmic advances The computer age has led to the popularity of many efficient and scalable clustering methods, e.g., spectral clustering [62] and convex clustering [12]. These methods are computationally efficient and well-suited to many modern applications where the sample size is quite large. Some algorithms do have some theoretical guarantees (discussed in the next section). At the same time, there have also been algorithmic advances on classical methods such as K-means [14, 22, 31].

Clustering in scientific applications As data science approaches become more and more

popular, there is an increasing literature of clustering in scientific applications, including but not limited to neuroscience [46], genomics [61], and astronomy [57]. These applications demand more theoretical understanding of the clustering problem, where it is not sufficient to merely provide a clustering result, but one should also be able to answer further validation questions, such as how significant are the clustering findings.

Current status of clustering theory

Machine learning literature Much of the modern clustering theory was pioneered by the machine learning community. In [64], it was explained that the generalization bound from statistical learning theory is not suitable in a general clustering framework, and proposed that convergence proofs and stability considerations should play key roles. In [63], it was clarified that in a statistical setting of clustering two questions need to be answered:

Question 1 Given a probability distribution over the data space, how is the data space separated into clusters as a function of the probability distribution? (conceptual question)

Question 2 Given a finite sample of data from the probability distribution, how can the clusters be approximated? (algorithmic and statistical question)

A well-studied line of this type of work is consistency theory in the context of spectral clustering [58, 63]. The nature of these works in some sense follows Pollard's earlier work about consistency of K-means clustering, but they require more advanced proof techniques (related to operator theory and calculus of variations).

Two methods for establishing clustering consistency can be distinguished. One starts from an algorithmic procedure and then studies its limiting behavior, e.g., [58] studies the ideal limit of

spectral clustering as a PDE problem. The other method starts from an ideal population problem, but implements an analytical algorithm on data samples from the population, e.g., in model-based clustering when the population is specified as a mixture of Gaussian distributions. The latter is the point of view taken in classical statistics. We refer to a culminating discussion in section 2.2.1 regarding the issue of Pollard-type clustering consistency.

The topic of cluster stability is discussed in the tuning parameter selection section 3.8 in Chapter 3.

Statistical literature In model-based clustering ([6], with Adrian Raftery being the main early contributor), clusters are modeled as components in a statistical mixture model, e.g., Gaussian mixtures. Recent literature has shifted from multivariate to high dimensional and nonparametric settings [3, 8]. The importance of studying clustering from the nonparametric statistics point of view should be emphasized here: when people use K-means or many modern machine learning methods for clustering, they are working nonparametrically. As in classification, people often do not want to impose parametric assumptions on the data ("Gaussian" or mixtures of Gaussian or any other specific distribution). The use of nonparametric statistics has been successful and mature in classification literature, leading to the so-called distribution free theory of classification [15, 21]. For the unsupervised clustering problem, however, there is difficulty: a nonparametric mixture model is not easy to make identifiable or to implement. For example, mixture of sub-Gaussians is not an identifiable model [3]. High dimensional mixture models will be discussed only briefly in Chapter 5 as they are not the focus of the thesis.

Geometric perspectives on clustering Geometric considerations arise from density-based clustering [11, 48], as structures of the level sets of a density-estimate. There are also interesting geometric models studied recently by applied mathematicians, for example, the Low-

Dimensional Large Noise (LDLN) model in [33]. We believe statisticians could play a bigger role here. The geometric perspective will become relevant in Chapter 2 of the thesis.

Statistical network clustering There are lines of research about clustering in the statistical networks literature, sometimes under the name community detection [1]. The limit theorems therein characterize the ranges for the network parameters (in stochastic block model, or more generally, random dot product graphs) that lead to consistent clustering in the limit, see [32, 36]. These results are mathematically rigorous, but require some different concepts and definitions.

1.2 Problem formulation

Given a distribution P_X on some data space $\mathcal{X}$ (we will mostly restrict attention to Euclidean space), let Y be a random variable indicating class membership, $Y \in \{1, \cdots, K\}$. To simplify the notation, we write $E[g]$ for $E[g(X)]$ if there is no confusion that the expected value is taken with respect to P_X.

As pointed out in Question 1, we need first a reasonable definition of a clustering as a function of the underlying probability distribution P_X. To this end, suppose the number of clusters is known a priori to be K. We define a clustering of P_X as the minimizer of the following general criterion

$$\min_{Y|X \in \mathcal{G}} \left\{ \min_{f: \mathcal{X} \to \{1, \cdots, K\}} E_{(X,Y)}[l(Y, f(X))] \right\}, \tag{1.1}$$

where the two things to be minimized over are the conditional probability functions $Y|X$ (i.e., $P(Y = k|X = x), k = 1, \cdots, K$) and the mapping f. Both the set of conditional probability functions ($\mathcal{G}$) and set of classification rules (f) can be further restricted in the minimization.

In this thesis we develop relevant theory when l is 0-1 loss (i.e., $l(Y, f(X)) = I\{Y \neq$

$f(X)\}$), and $\mathcal{G}$ belongs to the class of Lipschitz functions with bounded Lipschitz constant. Specifically, denote $g_k(x) = P(Y = k|X = x)$ (so $\sum_{k=1}^{K} g_k(x) = 1$), $p_k = P(Y = k) = E[g_k(X)], k = 1, \cdots, K$, we study

$$\mathcal{G}_{C,\alpha} = \{Y|X : |g_k(x) - g_k(y)| \leq C \cdot d(x, y) \text{ for all } x, y \in \mathcal{X}, k = 1, \cdots, K; \min_k p_k \geq \alpha\},$$

for some finite positive constants C, α, where C is a universal Lipschitz constant, and α is a lower bound on the smallest class probability. The class $\mathcal{G}_{C,\alpha}$ also depends on P_X through p_k. We make no further assumption on the set of classification rules here, in which case the inner minimum in (1.1) can be derived to be $E_X[\min_k(1 - g_k(X))]$, achieved by

$$f(x) = \arg\max_k g_k(x). \tag{1.2}$$

In particular for $K = 2$, we arrive at

$$\text{minimize } E_X[g \wedge (1 - g)] \text{ subject to } L(g) \leq C, E[g] \in [\alpha, 1 - \alpha] \tag{1.3}$$

for specified constants $C > 0, \alpha \in (0, 1)$. The presence of α excludes two trivial solutions: the constant 0 and constant 1 function.

We will work with the penalized form of this optimization problem instead:

$$\text{minimize } E_X[g \wedge (1 - g)] + \lambda_2 L(g) + \lambda_3 \max\{E[g], 1 - E[g]\}. \tag{1.4}$$

At first glance, the problem might appear a bit unusual from the point of view of classical op-

timization theory [7]: it is nonconvex and therefore hard to use the Lagrangian to characterize the solution (in fact we will prove that it in general does not satisfy strong duality). However, through some preliminary analysis and experiments, this formulation turns out to have some good mathematical properties and special features. Therefore, we commit to this formulation in the thesis.

(1.4) is a variational problem whose solution may be viewed as a Lipschitz-regularized Bayes classifier. The term "Bayes classifier" refers to the conventional name for the classification rule (1.2) in the machine learning literature ([39]), and is not formulated directly in terms of Bayesian decision theory ([18, 67]) in statistics.

Remark (motivation for (1.1)). In (1.1), the inner minimization is a classification problem (see next remark). The outer minimization charaterizes clustering as the "easiest" classification problem.

Remark (definition of a classification problem). The goal of classification is, under some joint distribution $P_{X,Y}$, to find a mapping $f : \mathcal{X} \to \{1, \cdots, K\}$ that minimizes the expected classification error, possibly within some subclass $\mathcal{F}$:

$$\min_{f \in \mathcal{F}} E_{(X,Y)}[l(Y, f(X))].$$

1.3 Outline of thesis

Chapter 1 will focus on preliminary results, including a preliminary optimality result (Theorem 1.1), a Pollard-type consistency result (Theorem 1.3), and Theorem 1.5 which sets up the basis for computation. Chapter 2 studies aspects of the variational problem in more depth. Im-

portant aspects include: a further optimality result with geometric interpretation (Theorem 2.1), a model-based consistency result (Theorem 2.3), and uniqueness (Theorem 2.6). Chapter 3 deals with computational issues, including a main algorithm (Algorithm 1). Chapter 4 discusses possible extensions. Chapter 5 summarizes contribution and future work.

1.4 Necessary condition for optimal g, part 1

This section studies necessary condition for a minimizer of the variational problem for $K = 2$. From now on denote

$$
\begin{aligned}
I_1 &= E[g \wedge (1-g)] \\
I_2 &= \lambda_2 L(g) \\
I_3 &= \lambda_3 \max\{E[g], 1 - E[g]\} \\
I &= I_1 + I_2 + I_3,
\end{aligned}
$$

where the expectation is with respect to the underlying distribution P_X. The problem is

$$
\underset{g\colon \mathbb{R}^d \to [0,1]}{\text{minimize}} \quad I(g). \tag{1.5}
$$

For results in this section we work with $\mathbb{R}^d$ instead of general metric space because of technical reasons: version of Rademacher's theorem [A.6] and the gradient formula [A.4] used in the proof hold in $\mathbb{R}^d$.

Throughout this section, assume $\lambda_3 < 1$, support$(P_X) = \Omega \subset \mathbb{R}^d$, where Ω can be $\mathbb{R}^d$ or

some compact subset of $\mathbb{R}^d$. Let $g^* \in \arg\min_g I(g)$ be an optimal solution, $L = L(g^*)$ denotes its Lipschitz constant if there is no confusion.

Remark. For any measure μ on $\mathbb{R}^n$, support$(\mu) = \{x \in \mathbb{R}^n : \mu(B_\rho(x)) > 0, \forall \rho > 0\}$.

First, we establish below a key lemma that will be used in the proof of several later results. It comes from the particular form of the objective function (or energy functional) $I(g)$.

Lemma 1.1 (comparison lemma)**.** *Let $g \in [0,1]$ be a Lipschitz function, and let $U = \{g = 1/2\}, U_1 = \{g < 1/2\}, U_2 = \{g > 1/2\}$. Suppose another Lipschitz function $g' \in [0,1]$ satisfies*

$$g' \leq g \ \text{ on } U_1, \tag{1.6}$$

$$g' \geq g \ \text{ on } U_2, \tag{1.7}$$

$$L(g') \leq L(g), \tag{1.8}$$

and if $\lambda_3 < 1$, then

$$I(g') \leq I(g).$$

The inequality becomes strict if any of the following holds: either (1.6) or (1.7) is strict for some $x \in U_1 \cap \Omega$ or $x \in U_2 \cap \Omega$, or $L(g') < L(g)$, or $\{g' = 1/2\} \cap \Omega \subsetneq U \cap \Omega$.

The proof is in section 1.8.1.

Remark. (1.6)-(1.8) and the conditions where strict inequality holds indicate several ways to find a local (or global) variation that reduces $I(g)$.

Remark (several uses of Lemma 1.1). We list below several applications of Lemma 1.1 in the thesis. For clarity, let us denote $U = \{g^* = 1/2\}, U_1 = \{g^* < 1/2\}, U_2 = \{g^* > 1/2\}$ where

g^* is an optimal solution of (1.5). That is, when applying Lemma 1.1, we are thinking of g as a candidate for the optimal and g' as a variation of g.

1. In Step 1 of proof of Theorem 1.1, g' is a local variation of g where $g' < g$ at some point $x \in U_1 \cap \Omega$ and $g' = g$ outside some neighborhood of x.

2. In Step 2 of proof of Theorem 1.1, Lemma 1.1 is used to show that $U \cap \Omega$ does not contain any ball. A local variation g' is constructed to satisfy $\{g' = 1/2\} \cap \Omega \subsetneq U \cap \Omega$ if otherwise a ball is contained in $U \cap \Omega$.

3. In Theorem 2.1, the form of g^* in the theorem satisfies $g^* \leq g < 1/2$ on U_1, $g^* \geq g > 1/2$ on U_2 and $L(g^*) = L(g)$, for any g that shares the same U, U_1, U_2, which roughly explains why it gives a necessary condition for optimality.

4. In Corollary 2.1, Lemma 1.1 is used to show that U is "thin", because a better function g' can be found by reducing U to a smaller set if it is "fat".

5. In Step 4 of proof of uniqueness Theorem 2.6, it is shown that if g_0, g_1 are two solutions, and if $L_0 > L_1$, then $g_0 > g_1$ on $S_1 = U_1 \cap \Omega$, $g_0 < g_1$ on $S_2 = U_2 \cap \Omega$. By Lemma 1.1, we would have $I(g_0) > I(g_1)$, which is a contradiction.

Now we present a first result regarding the form of g^*:

Theorem 1.1 (necessary condition 1, a.e. version)**.**

$$\textit{For any } x,\ g^*(x) = 0 \ \textit{ or } \ g^*(x) = 1 \textit{ or } ||\nabla g^*(x)|| = L, \textit{a.e. in } \Omega$$

The proof is in section 1.8.2.

Remark. $||\nabla g^*(x)||$ is the local Lipschitz constant of g^* at x. Note that a.e. is in Lebesgue measure, not in P_X.

Remark (Proof outline). Let $U = \{x : g^*(x) = 1/2\}$. We divide the proof into three steps:

Step 1. For any $x \in \Omega$ such that $g^*(x) < 1/2$ and differentiable, $||\nabla g^*(x)|| = L$ or $g^*(x) = 0$; for any $x \in \Omega$ that $g^*(x) > 1/2$ and differentiable, $||\nabla g^*(x)|| = L$ or $g^*(x) = 1$.

Step 2. For any ball $B_r(x), x \in \Omega$, $B_r(x) \backslash U$ has positive measure. By Rademacher's theorem [A.6], this implies there exists a differentiable point x_r within any ball $B_r(x), r > 0$.

Step 3. For any $x \in U \cap \Omega$ where $g^*(x)$ is differentiable, show $||\nabla g^*(x)|| = L$. Thus the statement of the theorem holds for all differentiable points in Ω, which, again by Rademacher's theorem, are almost everywhere in Ω.

Remark. In Theorem 1.1, $||\nabla g(x)||$ can be understood as the local Lipschitz constant at a differentiable point x. We make a more complete result below by extending the almost constant local Lipschitz constant property to every point in the space, including nondifferentiable points. In other words, the local Lipschitz constant at a nondifferentiable point is determined by those of surrounding differentiable points, which, by Theorem 1.1, are always equal to the global Lipschitz constant L. For this purpose we need the concept of generalized gradient (denoted by ∂_C) for Lipschitz functions, a generalization of subgradient (usually denoted by ∂) for convex functions, and gradient (denoted by ∇) for differentiable functions. We refer to the definition [A.1.2] and a list of properties that follows, including sum rule, mean value theorem and gradient formula in the appendix section. A more comprehensive coverage of generalized gradient can be found in Chapter 10 of [13].

The gradient formula in nonsmooth analysis says, in $\mathbb{R}^n$, the generalized gradient of a Lipschitz function can be generated by gradients at nearby points where derivative exist [A.4]. The everywhere version below is a direct consequence of the a.e. version (Theorem 1.1) and the gradient formula.

Theorem 1.2 (necessary condition 1, everywhere version)**.**

$$\textit{For any } x \in \Omega,\ g^*(x) = 0 \ \textit{ or } \ g^*(x) = 1 \textit{ or } \sup ||\partial_C g^*(x)|| = L,$$

where $\sup ||\partial_C g(x)||$ *can be understood as the local Lipschitz constant of a function* g *at point* x*.*

The proof is in section 1.8.3.

Remark. Since g^* is Lipschitz, $\partial_C g^*$ is well-defined. Note that $\partial_C g^*(x)$ is a compact convex set [A.1.2], so $||\partial_C g^*(x)||$ is a compact interval in $\mathbb{R}$. If g^* is locally convex, this reduces to $||\partial g(x)||$, where $\partial g(x)$ is the subgradient of g at x. When g^* is differentiable at x, this further reduces to $||\nabla g(x)||$ as in the differentiable case, and there is no supremum to take. See [A.6-A.9] for relationship between these quantities and the Lipschitz constant.

We end this section by some remarks on the distributional assumptions of P_X. We haven't made much assumption on P_X in Theorem 1.1 and 1.2, only that P_X has support Ω which is either $\mathbb{R}^d$ or some compact set. When making these results, we are mostly interested in P_X that are absolutely continuous on $\mathbb{R}^d$. The case $\Omega = \mathbb{R}^d$ is typical for many statistical models for clustering, such as the Gaussian mixture model. Another case that will be of interest in Chapter

2 is when $\Omega = S_1 \cup S_2$ for some compact, connected disjoint sets S_1, S_2. In this case care needs to be taken with what can be said about the behavior of g^* outside the support.

A slightly different but interesting case is when P_X is atomic, which is related to the data problem we will start to consider later in Chapter 1. In this case the a.e. version (Theorem 1.1) might be vacuous simply because the support is finite or countable, so is of measure 0 in $\mathbb{R}^d$. But the everywhere version still gives some restriction on the form of g^*, that the local Lipschitz constant at each point of the support is always equal to the global, whenever g^* is not 0 or 1. In particular, the local variational argument in Step 1 within the proof of Theorem 1.1 works also at these atoms, so that "rigid" solutions are still encouraged.

1.5 Pollard's consistency proof

In this section, we summarize Pollard's contribution in his classical paper on the consistency of K-means clustering [43]. This serves two purposes: first, the proof strategy for our problem will be very similar to [43] (see next section); second, the proof itself is elegant and the techniques therein can be generalized to study consistency of many other clustering or unsupervised learning methods.

K-means clustering

For a set of points $X_1, ..., X_n$ in $\mathbb{R}^d$, K-means clustering is based on the criterion of minimizing the within cluster sum of squares:

$$\text{minimize} \quad \sum_{k=1}^{K} \sum_{X_i \in C_k} ||X_i - c_k||^2, \tag{1.9}$$

where $C_1, ..., C_K$ are K distinct subsets of $X_1, \cdots, X_n$, with cluster centers $c_1, ..., c_K$, and the minimization is taken with respect to both $c_1, \cdots, c_K$ and cluster memberships. This criterion is the same as

$$\min_{c_1,\cdots,c_K} \frac{1}{n} \sum_{i=1}^{n} \min_{k=1,...,K} ||X_i - c_k||^2.$$

The paper [43] rewrites the above using the notation $C = \{c_1, ..., c_K\}$ and the empirical measure $P_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{X_i}$:

$$W(C, P_n) = \int \min_{c \in C} ||x - c||^2 dP_n, \tag{1.10}$$

$$\underset{|C| \leq K}{\text{minimize}} \quad W(C, P_n),$$

where $|C| \leq K$ is used instead of $|C| = K$, to allow the possibility that two cluster centers coincide. A natural population version for (1.10) is

$$W(C, P) := \int \min_{c \in C} ||x - c||^2 dP. \tag{1.11}$$

Consistency of K-means clustering

The paper [43] shows under some general conditions that the set of optimal cluster centers for the sample (the minimizer of (1.10)), converges to the set of centers that minimizes (1.11). The measure of closeness used is the Hausdorff metric $H(\cdot, \cdot)$, which is defined for compact subsets A, B of $\mathbb{R}^d$ by: $H(A, B) < \delta$ iff every point of A is within (Euclidean) distance δ of at least

one point of B, and vice versa. The convergence is almost surely, that $\arg\min_C W(C, P_n) \xrightarrow{a.s.} \arg\min_C W(C, P)$.

The proof is achieved by first showing the optimal sample cluster centers eventually lie in some compact region $\mathcal{E}_K$ of $\mathbb{R}^d$, then the convergence result is derived following a uniform law of large numbers statement. By strong law of large numbers, $W(A, P_n) - W(A, P)$ converges almost surely to zero for any set A such that $W(A, P) < \infty$; it turns out this difference also converges to zero uniformly over all size-K subsets of $\mathcal{E}_K$. This is shown via techniques from empirical process theory [A.2.1]. Similar technique is applied within the proof of [63].

Proof sketch

By strong law of large numbers, if $W(C, P) < \infty$, then

$$W(C, P_n) \xrightarrow{a.s.} W(C, P).$$

Let

$$C_n := \arg\min_C W(C, P_n),$$

$$C_0 := \arg\min_C W(C, P). \tag{1.12}$$

Suppose C_0 is unique, this is interpreted as true cluster centers. The major question is,

$$\text{does } C_n \xrightarrow{a.s.} C_0 \text{ hold?}$$

Step 1: When n large enough, one can show C_n eventually lies in some compact region $\mathcal{E}$ of $\mathbb{R}^d$.

Step 2: By a uniform law of large numbers

$$\sup_{C\subset\mathcal{E},\ |C|\leq K} |W(C,P_n) - W(C,P)| \xrightarrow{a.s.} 0,$$

we have

$$W(C_n,P_n) - W(C_n,P) \xrightarrow{a.s.} 0, \tag{1.13}$$

$$W(C_0,P_n) - W(C_0,P) \xrightarrow{a.s.} 0. \tag{1.14}$$

By definition of C_n, C_0,

$$W(C_n,P_n) \leq W(C_0,P_n),$$

$$W(C_0,P) \leq W(C_n,P),$$

so we have

$$\begin{aligned}
0 &\leq W(C_n,P) - W(C_0,P) \\
&= W(C_n,P) - W(C_n,P_n) + W(C_n,P_n) - W(C_0,P_n) + W(C_0,P_n) - W(C_0,P) \\
&\leq (W(C_n,P) - W(C_n,P_n)) + (W(C_0,P_n) - W(C_0,P)).
\end{aligned}$$

Therefore by (1.13)&(1.14),

$$W(C_n, P) \xrightarrow{a.s.} W(C_0, P).$$

Then by continuity (in Hausdorff metric) of the map $C \longrightarrow W(C, P)$ on $\mathcal{E}$ and uniqueness of C_0,

$$C_n \xrightarrow{a.s.} C_0.$$

Remark (Algorithmic consideration). Because the memberships are also variables, (1.9) is not a convex optimization problem, but a combinatorial one, thus finding its global minimum is computationally difficult. In practice a heuristic algorithm is often carried out to find a locally optimal partition (the most widely-used one is Lloyd's algorithm [34]). There are also efficient approximation algorithms now, see [14, 22, 31]. Pollard's paper did not include these algorithmic considerations. The result means if we can indeed find the global minimum of (1.9), then the cluster centers attained by finite sample will converge to the true cluster centers (1.12) defined only through the underlying distribution P that generates the data.

Remark (Drawback of Pollard-type consistency). The consistency result applies to any distribution P such that $\int ||x||^2 dP < \infty$. It presumes the minimizer of the population criterion (1.11) to be the truth, then the empirical minimizer will converge to it. We call this type of clustering consistency "Pollard-type consistency". The drawback of such consistency is that even though the result looks distribution free, the minimizer of (1.11) may not be interpretable, e.g. when P has no clustering structure, or when the clusters are not spherical or convex. Consistency of

a clustering method in some statistical model is harder to attain. We will consider this in later chapters.

1.6 Consistency, part 1: general case

Consider a data-based version of (1.5). A positive feature of the formulation is that consistency result follows in a similar way as in Pollard's paper [43]. In [43], the optimal sample cluster centers C_n converge to the optimal centers C_0 derived from P, provided that the criterion function (1.11) has unique minimum. Here we can similarly show that the estimated clustering function g_n converges to the optimal g derived from P, provided that the variational problem (1.5) has unique solution.

1.6.1 Consistency for $K = 2$

Following the notation from the last section, let

$$W(g, P) = E[g \wedge (1 - g)] + \lambda_2 L(g) + \lambda_3 \max\{E[g], 1 - E[g]\},$$

$$W(g, P_n) = \frac{1}{n} \sum_{i=1}^{n} \min\{g(X_i), 1 - g(X_i)\} + \lambda_2 L(g) + \lambda_3 \max\{\overline{g(X)}, 1 - \overline{g(X)}\},$$

where $\overline{g(X)} = \frac{1}{n} \sum_{i=1}^{n} g(X_i)$.

Theorem 1.3. *Suppose $\mathcal{X}$ is a totally bounded metric space. Let $g_n \in \underset{g:\mathcal{X}\to[0,1]}{\arg\min}\, W(g, P_n)$. Suppose $W(g, P)$ has unique minimum g_0. Then $||g_n - g_0||_{L_1(P)} \xrightarrow{a.s.} 0$.*

Remark. This result holds regardless of the dimension of $\mathcal{X}$. The essential part is a finite covering number for the Lipschitz function ball on the metric space $\mathcal{X}$, so that uniform law of large

numbers applies on the function class $\{f \in Lip(\mathcal{X}) : L(f) \leq C\}$ for any positive constant C.

Remark (Proof outline). First, it can be shown that the Lipschitz constant of g_n is bounded. We have

$$\lambda_2 L(g_n) \leq W(g_n, P_n) \leq W(0, P_n) = \lambda_3,$$

here 0 denotes the constant 0 function. Therefore

$$L(g_n) \leq \frac{\lambda_3}{\lambda_2}, \quad \text{for any } n.$$

Let $C = \frac{\lambda_3}{\lambda_2}$. We can break down the proof into four parts (the norm $L_1(P)$ in the theorem is chosen so that (3) is convenient to show):

1. $\sup_{g \in \mathcal{G}} |P_n g - Pg| \xrightarrow{a.s.} 0$, where $\mathcal{G} = \{g : \mathcal{X} \to [0,1] \mid L(g) \leq C\}$.
2. $\sup_{g \in \mathcal{G}} |W(g, P_n) - W(g, P)| \xrightarrow{a.s.} 0$.
3. $W(g, P)$ is continuous in g with $|| \cdot ||_{L_1(P)}$.
4. $W(g_n, P) \xrightarrow{a.s.} W(g_0, P)$, then by 3. and uniqueness of g_0, $||g_n - g_0||_{L_1(P)} \xrightarrow{a.s.} 0$.

Step 4 is along the same lines as Step 2 of the last section (i.e., along the same lines as [43]): continuity of the map $g \to W(g, P)$ on $\mathcal{G}$ and uniqueness of g_0 implies $||g_n - g_0||_{L_1(P)} \xrightarrow{a.s.} 0$. Proofs of Step 1, 2, 3 are in section 1.8.4.

We immediately obtain the following two corollaries (from the proof of Theorem 1.3) about consistency of the objective function value, and pointwise convergence of g_n to g_0 in $supp(P)$.

Corollary 1.1 (Consistency of objective value)**.** $I(g_n) \xrightarrow{a.s.} I(g_0)$.

This is within the statement of step 4 in the proof of Theorem 1.3.

Corollary 1.2 (From L_1 consistency to pointwise)**.** *Suppose g_0 is a unique solution for (1.5), then for any $x \in supp(P)$, $g_n(x) \xrightarrow{p} g_0(x)$.*

Theorem 1.3 is about g_n being $L_1(P)$ consistent to g_0, if the latter is unique. This, along with continuity of g_n and g_0, implies pointwise convergence. The statement may not hold outside the support, where g_0 is in general not unique.

1.6.2 Consistency for $K > 2$

For general K, we need to measure the distance between the set of K estimated functions which minimizes the data-based formulation, and the set of K functions that minimizes the population formulation. It is shown that their Hausdorff distance will converge to zero.

Theorem 1.4. *Denote $\underline{g} = \{g_1, ..., g_K\}$ satisfying $\sum_{k=1}^{K} g_k = 1$ and let*

$$W(\underline{g}, P) = E[(1 - g_1) \wedge (1 - g_2) \cdots \wedge (1 - g_K)] + \lambda_2 \max_k L(g_k) + \lambda_3 \max_k (1 - E[g_k]),$$

$$W(\underline{g}, P_n) = \frac{1}{n} \sum_{i=1}^{n} (1 - g_1(X_i)) \wedge (1 - g_2(X_i)) \cdots \wedge (1 - g_k(X_i)) + \lambda_2 \max_k L(g_k)$$
$$+ \lambda_3 \max_k \{1 - \overline{g_k(X)}\}$$

where $\overline{g_k(X)} = \frac{1}{n} \sum_{i=1}^{n} g_k(X_i)$.

Let $\underline{g_n} \in \arg\min_{\underline{g}\in\mathcal{G}} W(\underline{g}, P_n)$, $\mathcal{G} = \{\{g_1, \cdots, g_k\} | g_1, \cdots, g_k : \mathcal{X} \to [0,1], \sum_{k=1}^{K} g_k = 1\}$. *Suppose* $\min_{\underline{g}\in\mathcal{G}} W(g, P)$ *has unique minimum* $\underline{g_0}$. *Then*

$$d_H(\underline{g_n}, \underline{g_0}) \xrightarrow{a.s.} 0,$$

where $d_H(\cdot,\cdot)$ *is the Hausdorff distance between two sets of functions with norm* $||\cdot||_{L_1(P)}$. *E.g., if* $\underline{g^{(1)}} = \{g_1^{(1)}, \cdots, g_{k_1}^{(1)}\}, \underline{g^{(2)}} = \{g_1^{(2)}, \cdots, g_{k_2}^{(2)}\}$, *then* $d_H(\underline{g^{(1)}}, \underline{g^{(2)}}) = \max_i \min_j ||g_i^{(1)} - g_j^{(2)}||_{L_1(P)}$.

The proof is in section 1.8.6.

1.7 Computing with data

A representer theorem

Even though (1.5) is in general a variational problem, we show a data-based version of (1.5) can be represented by finitely many parameters, which are the values of g on the data points. Therefore with data we can solve a finite dimensional optimization problem, which makes the method practical, providing the basis for an algorithm.

Theorem 1.5. *The following two problems are equivalent:*

$$\min_{g:\mathcal{X}\to[0,1]} \left\{ \frac{1}{n}\sum_{i=1}^{n} \min\{g(x_i), 1-g(x_i)\} + \lambda_2 L(g) + \lambda_3 \max\{\overline{g(x)}, 1-\overline{g(x)}\} \right\} \tag{1.15}$$

and

$$\min_{a_1,\cdots,a_n\in[0,1]} \frac{1}{n}\sum_{i=1}^{n} \min\{a_i, 1-a_i\} + \lambda_2 \max_{d(x_i,x_j)\neq 0} \frac{a_i - a_j}{d(x_i,x_j)} + \lambda_3 \max\{\bar{a}, 1-\bar{a}\}. \tag{1.16}$$

The proof is in section 1.8.7.

R function implementation

Computation based on Theorem 1.5 is coded in R software. The actual algorithm, based on linear programming and an alternating minimization strategy to deal with non-convexity (or, to avoid combinatorial optimization schemes), will be introduced in Chapter 3.

1.8 Proofs of chapter 1

1.8.1 Proof of Lemma 1.1

Notice that (1.6) and (1.7) implies $g' < 1/2$ on U_1 and $g' > 1/2$ on U_2, from which we deduce

$$
\begin{aligned}
I_1(g') - I_1(g) &= E[g' \wedge (1-g')] - E[g \wedge (1-g)] \\
&= E[g' I_{U_1}] + E[(1-g') I_{U_2}] - (E[g I_{U_1}] + E[(1-g) I_{U_2}]) \\
&\quad + E[(g' \wedge (1-g') - \frac{1}{2}) I_U] \\
&= E[(g'-g) I_{U_1}] + E[(g-g') I_{U_2}] + E[(g' \wedge (1-g') - \frac{1}{2}) I_U] \\
&\leq 0 \text{ (by (1.6), (1.7))}, \\
I_3(g') - I_3(g) &= \lambda_3 (\max\{E[g'], 1-E[g']\} - \max\{E[g], 1-E[g]\}) \\
&\leq \lambda_3 |E[g'-g]| \text{ (by [A.14])} \\
&\leq \lambda_3 (E|(g'-g) I_{U_1}| + E|(g'-g) I_{U_2}| + E|(g' - \frac{1}{2}) I_U|) \\
&= \lambda_3 (E|(g'-g) I_{U_1}| + E|(g'-g) I_{U_2}| + E|(g' \wedge (1-g') - \frac{1}{2}) I_U|) \\
&\quad \text{(for any number } a, |a - \frac{1}{2}| = |a \wedge (1-a) - \frac{1}{2}|).
\end{aligned}
$$

For any function $f \leq 0, Ef + \lambda_3 E|f| = (1-\lambda_3)Ef$. Apply this separately to the three functions $f = (g'-g)I_{U_1}, (g-g')I_{U_2}, [g' \wedge (1-g') - \frac{1}{2}]I_U$ respectively to obtain

$$\begin{aligned} I_1(g') - I_1(g) + I_3(g') - I_3(g) &\leq (1-\lambda_3)(E[(g'-g)I_{U_1}] + E[(g-g')I_{U_2}] \\ &\quad + E[(g' \wedge (1-g') - \frac{1}{2})I_U]) \\ &= (1-\lambda_3)(I_1(g') - I_1(g)) \leq 0. \end{aligned}$$

Finally, since $I_2(g') - I_2(g) = \lambda_2(L(g') - L(g)) \leq 0$,

$$I(g') - I(g) \leq I_1(g') - I_1(g) + I_2(g') - I_2(g) + I_3(g') - I_3(g) \leq 0.$$

Now we look at when the inequality holds strictly. If $g' < g$ for some point in $U_1 \cap \Omega$, then $E[(g'-g)I_{U_1}] < 0$; if $g' > g$ for some point in $U_2 \cap \Omega$, then $E[(g-g')I_{U_2}] < 0$; if $\{g' = 1/2\} \cap \Omega \subsetneq U \cap \Omega$, then $E[(g' \wedge (1-g') - \frac{1}{2})I_U] < 0$; if $L(g') < L(g)$, then $I_2(g') < I_2(g)$. When either of these happens, we have $I(g') < I(g)$.

1.8.2 Proof of Theorem 1.1

Step 1.

Let g be any Lipschitz function with values in $[0,1]$ such that there exists $x \in \Omega, 0 < g(x) < 1/2$ and $||\nabla g(x)|| < L$.

Consider local variation $g_\epsilon(x) = g(x) + \epsilon\eta$, where $\eta, ||\nabla\eta||$ are both bounded in a neighborhood of x such that $0 < g < 1/2$ and $||\nabla g|| < L$, and $\eta = 0$ outside the neighborhood. When

ϵ small enough, we have $0 < g_\epsilon < 1/2$ and $||\nabla g_\epsilon|| < L$ in the neighborhood, so

$$I_1(g_\epsilon) - I_1(g) = E[g_\epsilon] - E[g] = \epsilon E[\eta],$$

$$I_2(g_\epsilon) = I_2(g),$$

$$I_3(g_\epsilon) - I_3(g) \leq \lambda_3 |E[g_\epsilon] - E[g]| = \epsilon \lambda_3 |E[\eta]|.$$

Thus $I(g_\epsilon) - I(g) = (I_1(g_\epsilon) - I_1(g)) + (I_3(g_\epsilon) - I_3(g)) \leq \epsilon(E[\eta] + \lambda_3 |E[\eta]|)$. As long as η is chosen such that $\eta < 0$ on the neighborhood of x, since $\lambda_3 < 1$, we have $I(g_\epsilon) < I(g)$. Therefore g cannot be the minimizer.

In conclusion, g cannot be a minimizer unless for every x with $||\nabla g(x)|| < L$ and $g(x) < 1/2$, $g(x) = 0$. Similarly, g cannot be a minimizer unless for every x with $||\nabla g(x)|| < L$ and $g(x) > 1/2$, $g(x) = 1$.

Remark. A particular construction for the local variation can be taken as

$$g_\epsilon(y) = \begin{cases} g(y), & y \notin B_\epsilon(x) \\ \inf\limits_{z \in S_\epsilon} \{g(z) - L \cdot d(y,z)\}, & y \in B_\epsilon(x) \end{cases},$$

for some ϵ small enough such that $||\nabla g(y)|| < L$ for any $y \in B_\epsilon(x)$, and $g_\epsilon|_{B_\epsilon(x)} > 0$. This can be seen as a Lipschitz extension of $g\Big|_{S_\epsilon(x)}$ into $B_\epsilon(x)$, where S_ϵ is the boundary of $B_\epsilon(x)$. By Lemma 1.1, $I(g_\epsilon) < I(g)$. Such argument works also when P_X has point mass at x and has measure 0 in $B_\epsilon(x) \backslash \{x\}$: since $||\nabla g(y)|| < L$ for any $y \in B_\epsilon(x)$, we deduce $g_\epsilon(x) < g(x)$.

Step 2.

Let us first show $B_r(x) \nsubseteq U$ for any point $x \in \Omega$. Suppose $B_r(x) \subset U$ for some point $x \in \Omega$ and for some $r > 0$, define a local modification g' of g^* within $B_r(x)$ (Figure 1.1) as

$$g'(y) = \begin{cases} g^*(y), & y \notin B_r(x) \\ 1/2 - L \cdot d(y, S_r(x)), & y \in B_r(x) \end{cases},$$

where $S_r(x)$ is the boundary of $B_r(x)$. From the construction, $I_1(g') < I_1(g^*), L(g') = L(g^*)$.

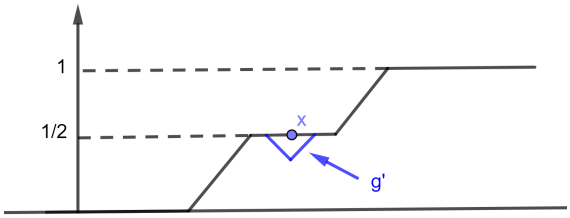


Figure 1.1: 1-d illustration of the local variation g'

By Lemma 1.1, $I(g') < I(g^*)$, a contradiction. This shows $U = \{g^* = 1/2\}$ does not contain any ball.

Let m denote Lebesgue measure in $\mathbb{R}^d$. Next we show $m(B_r(x) \backslash U) > 0$ for any $x \in \Omega, r > 0$. It suffices to show by contradiction that U cannot have full measure in $B_r(x)$ for any $r > 0$. Suppose $m(U \cap B_r(x)) = m(B_r(x))$ for some $r > 0$ and $x \in \Omega$, i.e., by definition of U, $g^* = 1/2$ a.e. on $B_r(x)$. By continuity of g^*, we deduce $g^* \equiv 1/2$ on $B_r(x)$. By the construction of g' above, such g^* cannot be optimal, a contradiction. Therefore, $m(U \cap B_r(x)) < m(B_r(x))$. In other words, $m(B_r(x) \backslash U) > 0$, i.e., $B_r(x) \backslash U$ has positive measure. By Rademacher's theorem [A.6], this allows us to pick a differentiable point $x_r \subset B_r(x)$ such that $g^*(x_r) \neq 1/2$. Further, when r small enough, g^* is bounded away from 0 and 1, i.e., there exists $r_0 > 0$ such that for any $r < r_0, 0 < g^*(x_r) < 1$. By Step 1, $||\nabla g^*(x_r)|| = L$ for every $r < r_0$.

Step 3.

From Step 2, for any differentiable point $x \in U$, there is a sequence of differentiable points $x_r \to x$ as $r \to 0$. Moreover, since the gradients $\nabla g(x_r)$ are bounded, there is a convergent subsequence $\nabla g(x_{r_n}), n = 1, 2, \cdots$. By gradient formula [A.4], $||\nabla g(x)|| = || \lim_{n\to\infty} \nabla g(x_{r_n})|| = L$.

This concludes the proof.

1.8.3 Proof of Theorem 1.2

By gradient formula [A.4], $\partial_C g(x) = \text{co}\{\lim_n \nabla g(x_n) : \lim_n \nabla g(x_n) \text{ exists}, x_n \to x\}$, the convex envelope of all limiting gradients arising from neighborhood of x. All differentiable points are dealt with in the almost everywhere version. For non-differentiable points that are not 0 or 1, by [A.8],

$$\sup ||\partial_C g(x)|| = \sup_{\lim_n \nabla g(x_n) \text{ exists}, x_n \to x} || \lim_n \nabla g(x_n)|| = L.$$

1.8.4 Proof of Theorem 1.3

(1) Step 1 follows from the following lemma:

Lemma 1.2 (uniform law of large numbers for Lipschitz function class)**.** *Suppose $(\mathcal{X}, d)$ is a totally bounded metric space. Let $\mathcal{G}_C := \{g : \mathcal{X} \to [0, 1] \mid L(g) \leq C\}$ for some constant $C > 0$,*

then for any probability measure P on $\mathcal{X}$,

$$\sup_{g \in \mathcal{G}_C} |P_n g - Pg| \xrightarrow{a.s.} 0.$$

The proof is in section 1.8.5.

(2) The analysis can be broken down into two parts since

$$\sup_{g \in \mathcal{G}_C} |W(g, P_n) - W(g, P)| \leq \sup_{g \in \mathcal{G}_C} |E[g \wedge (1-g)] - \frac{1}{n} \sum_{i=1}^{n} \min\{g(X_i), 1 - g(X_i)\}|$$

$$+\lambda_3 \sup_{g \in \mathcal{G}_C} |\max\{E[g], 1 - E[g]\} - \max\{\overline{g(X)}, 1 - \overline{g(X)}\}|.$$

For the first part we can consider

$$\mathcal{H} = \{h : h = g \wedge (1-g), g \in \mathcal{G}_C\}.$$

By [A.2], for any $h \in \mathcal{H}, L(h) \leq C$, so $\mathcal{H}$ is a subset of $\mathcal{G}_C$. It follows that

$$\sup_{h \in \mathcal{H}} |P_n h - Ph| \xrightarrow{a.s.} 0.$$

For the second part, by [A.14],

$$|\max\{E[g], 1 - E[g]\} - \max\{\overline{g(X)}, 1 - \overline{g(X)}\}| \leq |E[g] - \overline{g(X)}| = |Pg - P_n g|,$$

so it follows that

$$\sup_{g\in\mathcal{G}_C} |\max\{E[g], 1-E[g]\} - \max\{\overline{g(X)}, 1-\overline{g(X)}\}| \le \sup_{g\in\mathcal{G}_C} |Pg - P_n g| \xrightarrow{a.s.} 0.$$

(3) We have

$$\begin{aligned}
|W(g_1, P) - W(g_2, P)| &\le \int |g_1 \wedge (1-g_1) - g_2 \wedge (1-g_2)| dP \\
&\quad + \lambda_3(\max\{E[g_1], 1-E[g_1]\} - \max\{E[g_2], 1-E[g_2]\}) \\
&\le \int |g_1 - g_2| dP + \lambda_3 \int |g_1 - g_2| dP \\
&= (1+\lambda_3) E|g_1 - g_2|.
\end{aligned}$$

Therefore $W(g, P)$ is continuous in g with $||\cdot||_{L_1(P)}$. The proof is complete.

1.8.5 Proof of Lemma 1.2

Result of this type can be found in the empirical process theory literature, a brief exposition of the background is given in the appendix section [A.2.1]. Here we refer to Example 19.11 of [60] for the 1-d case; in the general metric space setting, the result basically follows from a finite covering number for the Lipschitz function ball. The following only serves to address the technical difference between these existing results and the version we need in Lemma 1.2.

Proof on $\mathbb{R}^1$: Suppose P has bounded support $[a, b]$, then $\mathcal{G}_C$ has bounded variation $C(b-a)$. Therefore $\mathcal{G}_C$ is a subset of the bounded variation class. It suffices to show that the bounded variation class is P-Glivenko-Cantelli. By [A.10], since $||\cdot||_{L_1(P)} \le ||\cdot||_{L_2(P)}$, the bounded variation class has finite bracketing number in $||\cdot||_{L_1(P)}$ for any $\epsilon > 0$. By [A.9], it is P-

Glivenko-Cantelli.

Proof on general metric space: Suppose $(\mathcal{X}, d)$ is a totally bounded metric space. By [A.12], $N(\epsilon, \mathcal{G}_C, ||\cdot||_\infty) < \infty$. Let F be the constant 1 function on $\mathcal{X}$, then F is an envelope function for $\mathcal{G}_C$ (i.e., $|g(x)| \leq F(x)$ for any $g \in \mathcal{G}_C$ and $x \in \mathcal{X}$) with $||F||_{L_1(Q)} = 1$ for any probability measure Q on $\mathcal{X}$. Note that $N(\epsilon, \mathcal{G}_C, ||\cdot||_{L_1(Q)}) \leq N(\epsilon, \mathcal{G}_C, ||\cdot||_\infty)$, since for any two functions $f, g \in \mathcal{G}_C$, $||f-g||_\infty < \epsilon$ implies $||f-g||_{L_1(Q)} \leq ||f-g||_\infty < \epsilon$ for any probability measure Q. Therefore,

$$\sup_Q N(\epsilon||F||_{L_1(Q)}, \mathcal{G}_C, L_1(Q)) \leq N(\epsilon, \mathcal{G}_C, ||\cdot||_\infty) < \infty,$$

by [A.11], $\mathcal{G}_C$ is P-Glivenko-Cantelli.

1.8.6 Proof of Theorem 1.4

For any k, we have

$$L(g_k) \leq \max_k L(g_k) \leq W(\underline{g}_n, P_n) \leq W(0, P_n) = \lambda_3,$$

Therefore

$$L(g_k) \leq \frac{\lambda_3}{\lambda_2} := C, k = 1, \cdots, K.$$

Similar to Theorem 1.3, the proof contains four parts.

1. $\sup_{L(g_k) \leq C} |P_n g_k - P g_k| \xrightarrow{a.s.} 0, k = 1, \cdots, K$, by Lemma 1.2, as in Theorem 1.3. Let $\mathcal{G}_C := \{\underline{g} \in \mathcal{G} : L(g_k) \leq C, k = 1, \cdots, K\}$.

2. $\sup_{\underline{g}\in\mathcal{G}_C} |W(\underline{g}, P_n) - W(\underline{g}, P)| \leq \sup_{h\in\mathcal{H}} |P_n h - Ph| + \lambda_3 \sup_{\underline{g}\in\mathcal{G}_C} |\max_k(1 - E[g_k]) - \max_k\{1 - \overline{g_k(X)}\}|$, where $\mathcal{H} = \{h : h = (1 - g_1) \wedge (1 - g_2) \cdots \wedge (1 - g_K), \{g_1, \cdots, g_K\} \in \mathcal{G}_C\}$.

   By [A.2], for any $h \in \mathcal{H}, L(h) \leq C$. It follows that $\sup_{h\in\mathcal{H}} |P_n h - Ph| \xrightarrow{a.s.} 0$.

   Also, by [A.14], $|\max_k(1-E[g_k])-\max_k\{1-\overline{g_k(X)}\}| \leq \max_k |E[g_k]-\overline{g_k(X)}| = \max_k |Pg_k - P_n g_k|$, so

$$\sup_{\underline{g}\in\mathcal{G}_C} |\max_k(1 - E[g_k]) - \max_k\{1 - \overline{g_k(X)}\}| \leq \max_k \sup_{L(g_k)\leq C} |Pg_k - P_n g_k| \xrightarrow{a.s.} 0.$$

   Therefore, $\sup_{\underline{g}\in\mathcal{G}_C} |W(\underline{g}, P_n) - W(\underline{g}, P)| \xrightarrow{a.s.} 0$.

3. When $d_H(\underline{g^{(1)}}, \underline{g^{(2)}}) < \delta$,

$$\begin{aligned} |W(\underline{g^{(1)}}, P) - W(\underline{g^{(2)}}, P)| \leq \int & |(1 - g_1^{(1)}) \wedge (1 - g_2^{(1)}) \cdots \wedge (1 - g_K^{(1)}) \\ & - (1 - g_1^{(2)}) \wedge (1 - g_2^{(2)}) \cdots \wedge (1 - g_K^{(2)})| dP \\ & + \lambda_3(\max_k(1 - E[g_k^{(1)}]) - \max_k(1 - E[g_k^{(2)}])). \end{aligned}$$

We have

$$
\begin{aligned}
&\int |(1-g_1^{(1)}) \wedge (1-g_2^{(1)}) \cdots \wedge (1-g_K^{(1)}) - (1-g_1^{(2)}) \wedge (1-g_2^{(2)}) \cdots \wedge (1-g_K^{(2)})| dP \\
&\leq \int \max_i \min_j |g_i^{(1)}(x) - g_j^{(2)}(x)| dx \\
&\leq \sum_i \int \min_j |g_i^{(1)}(x) - g_j^{(2)}(x)| dx \\
&\leq \sum_i \min_j \int |g_i^{(1)}(x) - g_j^{(2)}(x)| dx \\
&\leq K \max_i \min_j \int |g_i^{(1)}(x) - g_j^{(2)}(x)| dx \\
&= K d_H(\underline{g^{(1)}}, \underline{g^{(2)}}),
\end{aligned}
$$

and

$$
\max_k (1 - E[g_k^{(1)}]) - \max_k (1 - E[g_k^{(2)}]) \leq \max_i \min_j |g_i^{(1)} - g_j^{(2)}| = d_H(\underline{g^{(1)}}, \underline{g^{(2)}}).
$$

Therefore $W(\underline{g}, P)$ is continuous in $\underline{g}$ with $d_H(\cdot, \cdot)$.

4. By 2 &3, and uniqueness of g_0, $d_H(\underline{g_n}, \underline{g_0}) \xrightarrow{a.s.} 0$.

1.8.7 Proof of Theorem 1.5

Suppose $g(x_i) = a_i, i = 1, \cdots, n$. By [A.5] and [A.1], there exists a Lipschitz extension $\tilde{g}$ of g such that

$$
\tilde{g}(x_i) = g(x_i) = a_i, i = 1, \cdots, n; \min_{i=1,\cdots,n} a_i \leq \tilde{g}(x) \leq \max_{i=1,\cdots,n} a_i, \text{ for any } x, \tag{1.17}
$$

given by

$$\tilde{g}(x) = \frac{1}{2} \min_{i=1,\cdots,n} \{a_i + Ld(x, x_i)\} + \frac{1}{2} \max_{i=1,\cdots,n} \{a_i - Ld(x, x_i)\}, \tag{1.18}$$

where $L = \max_{d(x_i,x_j)\neq 0} \frac{a_i - a_j}{d(x_i,x_j)}$ is the Lipschitz constant of g on $\{x_1, \cdots, x_n\}$.

Let

$$I^{(1)}(g) = \frac{1}{n} \sum_{i=1}^{n} \min\{g(x_i), 1 - g(x_i)\} + \lambda_2 L(g) + \lambda_3 \max\{\overline{g(x)}, 1 - \overline{g(x)}\},$$
$$I^{(2)}(a) = \frac{1}{n} \sum_{i=1}^{n} \min\{a_i, 1 - a_i\} + \lambda_2 \max_{d(x_i,x_j)\neq 0} \frac{a_i - a_j}{d(x_i, x_j)} + \lambda_3 \max\{\bar{a}, 1 - \bar{a}\}, a = \{a_1, \cdots, a_n\},$$

and denote their minimizers by g_1^*, a_2^*, respectively. Let g_2^* be the interpolation function of a_2^* by the construction in (1.17), and $a_1^* = \{g_1(x_1), \cdots, g_1(x_n)\}$. Note that $I^{(2)}(a_1^*) \leq I^{(1)}(g_1^*)$, because $\max_{d(x_i,x_j)\neq 0} \frac{g(x_i) - g(x_j)}{d(x_i,x_j)} \leq L(g)$. Then,

$$I^{(2)}(a_2^*) \leq I^{(2)}(a_1^*) \leq I^{(1)}(g_1^*) \leq I^{(1)}(g_2^*).$$

From (1.17), $I^{(2)}(a_2^*) = I^{(1)}(g_2^*)$. It follows that $I^{(2)}(a_2^*) = I^{(2)}(a_1^*) = I^{(1)}(g_1^*) = I^{(1)}(g_2^*)$, i.e. the two minimizations are equivalent.

1.9 Appendix: no strong duality

We are interested in whether the constrained version (1.3) and the penalized version (1.4) of the problem can be solved by each other. We prove a negative result: strong duality in general does not hold for this problem. Therefore solving either version of the problem may lead to a solution that does not come from solving the other version for any parameter. Note that strong

duality and the KKT condition can sometimes hold for nonconvex problems ([7]).

We study a toy example below, which will be used soon to establish the difference between the constrained problem and the penalized problem.

Let P_X has density $f_X(x) = p(I_{[a_1,b_1]} + I_{[a_2,b_2]})$, where $a_1 < b_1 < a_2 < b_2, b_1 - a_1 = b_2 - a_2, p = \frac{1}{2|b_1-a_1|}$. Let $g(x) = L(x - b_1)I_{[b_1,a_2)}(x) + I_{[a_2,\infty)}(x)$, where $L = \frac{1}{a_2-b_1}$.

Remark. P_X is a uniform distribution on two disjoint intervals with equal probability mass and "margin" $1/L$. Such example with well-separated compact clusters will be an important generative model to study in Chapter 2, see C1. In general, suppose we have two well-separated compact clusters S_1, S_2, we can define their margin to be $\frac{1}{d(S_1,S_2)}$.

Below we show that g is not the optimal solution of (1.5) for any λ. In fact, we can relax the Lipschitz constant of g to get a better solution.

Let $g'(x) = \frac{1}{1/L+2\epsilon}(x - (b_1 - \epsilon))I_{[b_1-\epsilon,a_2+\epsilon)}(x) + I_{[a_2+\epsilon,\infty]}(x)$. When $\epsilon/(1/L + 2\epsilon) < 1/2$,

$$
\begin{aligned}
I_1(g') - I_1(g) = I_1(g') &= 2\int_{b_1-\epsilon}^{b_1} g'(x)\, p\, dx = \frac{pL}{1+2\epsilon L}\epsilon^2, \\
I_2(g') - I_2(g) &= \lambda_2(\frac{1}{1/L+2\epsilon} - L) = \frac{-2\lambda_2\epsilon L^2}{1+2\epsilon L}, \\
I_3(g') &= I_3(g),
\end{aligned}
$$

so $I(g') - I(g) = \frac{\epsilon L(p\epsilon - 2\lambda_2 L)}{1+2\epsilon L} < 0$ when $\lambda_2 > 0$ and ϵ is small enough. Therefore g is not optimal.

Strong duality does not hold

Suppose strong duality holds for (1.5), then the optimal solution can be characterized by KKT condition (see e.g., section 5.5.3 in [7]). We show that such necessary condition for opti-

mality does not apply here. Consider the constrained minimization problem:

$$\min_g I_1(g) \text{ s.t. } L(g) \leq C, \max\{E[g], 1 - E[g]\} \leq \alpha. \tag{1.19}$$

Let g^*, p^* be the optimal function and optimal value for (1.19). The Lagrangian associated with (1.19) is

$$L(g, \lambda_2, \lambda_3) = I_1(g) + \lambda_2(L(g) - C) + \lambda_3(\max\{E[g], 1 - E[g]\} - \alpha).$$

Define the dual function $h(\lambda_2, \lambda_3) = \inf_g L(g, \lambda_2, \lambda_3)$. The dual problem associated with (1.19) is

$$\max_{\lambda_2, \lambda_3} h(\lambda_2, \lambda_3) \text{ s.t. } \lambda_2 \geq 0, \lambda_3 \geq 0 \tag{1.20}$$

Let $\lambda_2^*, \lambda_3^*, d^*$ be the dual optimal variables and optimal value for (1.20). We always have $d^* \leq p^*$ (weak duality). To see whether strong duality holds in general, i.e., whether $d^* = p^*$, consider the uniform distribution example studied previously and let $C \geq \frac{1}{a_2 - b_1} = L$. In this case $g^* = L(x - b_1)I_{[b_1, a_2)} + I_{[a_2, \infty)}$ is an optimal function for (1.19) (though it may be not unique) because $I_1(g^*) = 0$, and so $p^* = 0$. Suppose $d^* = p^*$, and $g^*, \lambda_2^*, \lambda_3^*$ are the primal and dual optimal variables, then by "complementary slackness" (see remark below), any optimal function g^* should also minimize $L(g, \lambda_2^*, \lambda_3^*)$ (minimizing $L(g, \lambda_2^*, \lambda_3^*)$ is equivalent to minimizing $I_1(g) + \lambda_2^* L(g) + \lambda_3^* \max\{E[g], 1 - E[g]\}$, after throwing out constants), and that $L(g^*, \lambda_2^*, \lambda_3^*) = 0$. However, we have shown in the example that for any λ_2, λ_3, this g^* cannot be the minimizer of the Lagrangian under the uniform distribution setting. Therefore strong duality does not hold in general, so the constrained problem and the penalized problem may have

different properties.

Remark (complementary slackness). Suppose $d^* = p^*$, then

$$\begin{aligned} I_1(g^*) = h(\lambda_2^*, \lambda_3^*) &= \inf_g L(g, \lambda_2^*, \lambda_3^*) \\ &\leq L(g^*, \lambda_2^*, \lambda_3^*) \\ &= I_1(g^*) + \lambda_2^*(L(g^*) - C) + \lambda_3^*(\max\{E[g^*], 1 - E[g^*]\} - \alpha) \\ &\leq I_1(g^*). \end{aligned}$$

Therefore the two inequalities become equalities: the first one implies that g^* is the minimizer of $L(g, \lambda_2^*, \lambda_3^*)$; the second one implies that $L(g^*, \lambda_2^*, \lambda_3^*) = 0$.

Chapter 2: Optimal g under ideal population model, further results

This chapter focuses on the ideal problem

$$\text{minimize } E[g \wedge (1-g)] + \lambda_2 L(g) + \lambda_3 \max\{E[g], 1 - E[g]\}$$

and attempts to characterize the optimal solution g^* as clearly as we can. There are two main directions. One is to give necessary conditions under general P_X. The other is to assume that P_X is a probability measure supported on K sharp clusters (we focus on $K = 2$). In either case, the main idea is that finding the optimal g reduces to first finding an optimal U - the level set of g^* at $1/2$, then g^* is determined by U almost uniquely by a Lipschitz extension.

The motivation of this variational problem is described in Chapter 1, and Chapter 2 is written in such a way that it can be read independently. Recall our variational problem (1.5) for

$K = 2$:

$$I_1 = E[g \wedge (1-g)]$$

$$I_2 = \lambda_2 L(g)$$

$$I_3 = \lambda_3 \max\{E[g], 1 - E[g]\},$$

$$I = I_1 + I_2 + I_3$$

$$\underset{g:\mathcal{X}\to[0,1]}{\text{minimize}} \quad I(g).$$

Let $g^* \in \arg\min_g I(g)$ be an optimal solution. When the dependence on λ_2, λ_3 is stressed, it is denoted by $g^*(\cdot, \lambda_2, \lambda_3)$.

Organization. This chapter is organized as follows. In section 2.1, we give a better qualitative description of the optimal solution than in Theorem 1.1, and this will set up the foundation for later results in the chapter. In section 2.2, we study consistency of our variational procedure in recovering true clusters in a model with various cluster shapes. The difference between this and the consistency result established in Chapter 1 (Theorem 1.3) is discussed in section 2.2.1. This model will be used throughout later sections. A bipartite result (Theorem 2.4, which says $\frac{1}{2} - g^*$ has different signs on the clusters) is developed out of the consistency result, but allows for a wider range of tuning parameters. In section 2.5, we study uniqueness of the variational problem (1.5) under the model and under bipartite condition (that $g^* < 1/2$ on one cluster and $g^* > 1/2$ on the other). The remaining part in this line of results is essentially a geometric variational problem – some toy examples with underlying symmetry are studied in section 2.4.

2.1 Necessary condition for optimal g, part 2

The difficulty of the variational problem (1.5) is in its nonconvexity (in I_1) and nonsmoothness (in I_2): classical method in calculus of variations such as the Euler Lagrange equation cannot be directly applied, and general nonsmooth extensions of Euler-Lagrange [13] do not lead to useful first order necessary conditions. In this section we take another route: to exploit the close relation between Lipschitz functions and distance functions, and give a semi-constructive necessary condition. Part of the motivation for results in this section comes from explicit constructions of Lipschitz extension, such as Mcshane [A.1], and the fact that distance functions are in general neither smooth nor convex.

We characterize g^* in the following main theorem of the chapter, relating g^* to its level set at $\frac{1}{2}$:

Theorem 2.1. *Suppose $\lambda_3 < 1$, support$(P_X) = \Omega \subset \mathbb{R}^d$. Let $U = \{x : g^*(x) = 1/2\}, U_1 = \{x : g^*(x) < 1/2\}, U_2 = \{x : g^*(x) > 1/2\}$, $L = L(g^*)$, then g^* must have the form:*

$$g^*(x) = \begin{cases} \max\{\frac{1}{2} - Ld(x, U), 0\}, & x \in U_1 \cap \Omega; \\ \frac{1}{2}, & x \in U; \\ \min\{\frac{1}{2} + Ld(x, U), 1\}, & x \in U_2 \cap \Omega. \end{cases}$$

Moreover, when $\Omega \subsetneq \mathbb{R}^d$, there is always an optimal g^* that has the form:

$$g^*(x) = \begin{cases} \max\{\frac{1}{2} - Ld(x, U), 0\}, & x \in U_1; \\ \frac{1}{2}, & x \in U; \\ \min\{\frac{1}{2} + Ld(x, U), 1\}, & x \in U_2. \end{cases}$$

and all other optimals can be modified (in the way given by (2.1) and (2.2)) to have this canonical form.

The complete proof of Theorem 2.1 is in 2.7.1. See Remark (5) for motivation of the overall plan of the proof.

Remark (1). Once L and U are determined, g^* is uniquely determined on Ω. This turns the main problem from minimizing over g to minimizing over L, U, a geometric variational problem. Although in higher dimensions, solving the geometric variational problem of minimizing over U can still be hard, Theorem 2.1 gives us a better mental picture of the optimal solution, see further exposition in the end of this section.

Remark (2). Since g is continuous, and U is the preimage of $\{1/2\}$ under g, then U must be a closed set. Therefore for any x, its distance to U, $d(x, U) = \inf_{u \in U} ||x - u||$ is always achieved at some point $u \in U$, and $d(x, U) = 0$ iff $x \in U$.

Remark (3). Theorem 1.1 suggested that U cannot have positive measure. However, the measure 0 of U does not come out directly from the arguments here.

Remark (4). The form of g in the theorem satisfies previous necessary conditions (Theorem 1.1 and Theorem 1.2) on U^C:

$$\text{for a.e. } x \in U^C, g(x) = 0 \text{ or } g(x) = 1 \text{ or } ||\nabla g(x)|| = L;$$

$$\text{for any } x \in U^C, g(x) = 0 \text{ or } g(x) = 1 \text{ or } \sup ||\partial_C g(x)|| = L,$$

which follows from properties of distance functions [A.10]:

$$||\nabla d(x,U)|| = 1, a.e. \text{ in } U^C;$$

$$\sup ||\partial_C d(x,U)|| = 1, \forall x \in U^C.$$

Moreover, suppose we are willing to assume that U has measure 0, then for any $x \in U$,

$$\partial_C g(x) = \text{co}\{\lim_n \nabla g(x_n), x_n \to x\} = \text{co}\{\lim_n \nabla g(x_n), x_n \notin U, x_n \to x\},$$

by the gradient formula [A.4] and the fact that generalized gradient $\partial_C g(x)$ won't change if any set of measure 0 is excluded when building the sequence $\{x_n\}$ from neighborhood of x. It follows that the property $\sup ||\partial_C g(x)|| = L$ also extends (from U^C) to any $x \in U$ here. Therefore, Theorem 2.1 improves the necessary condition in Theorem 1.1, giving a more detailed description of optimal solution.

Remark (5). Below is motivation of the overall plan of the proof:

Suppose g is optimal but violates the above form on either U_1 or U_2 ($U = \{g = 1/2\}, U_1 = \{g < 1/2\}, U_2 = \{g > 1/2\}, L = L(g)$), then let g^* be a modification of g such that

$$g^* = g, \forall x \in U; \tag{2.1}$$

$$g^*(x) = \max\{\frac{1}{2} - Ld(x,U), 0\}, \forall x \in U_1; g^*(x) = \min\{\frac{1}{2} + Ld(x,U), 1\}, \forall x \in U_2. \tag{2.2}$$

The main arguments have two parts. First, it can be shown that $L(g^*) = L(g)$, that is, doing this modification does not change the Lipschitz constant.

Then to see why $I(g^*) < I(g)$, note that the first term in the objective is small when g is close to either 0 or 1. The form of g^* achieves this goal most "efficiently" while preserving the Lipschitz constant, among all functions that share the same level set U.

Specifically, we can show that

$$g^*(x) \leq g(x) < \frac{1}{2}, \forall x \in U_1;\ g^*(x) \geq g(x) > \frac{1}{2}, \forall x \in U_2. \tag{2.3}$$

By Lemma 1.1, this implies $I(g^*) \leq I(g)$.

The following result makes the "U" in Theorem 2.1 less mysterious.

Corollary 2.1 (U is "thin")**.** *Let g^* be an optimal solution with the form specified by Theorem 2.1, and $U = \{g^* = 1/2\}$. Suppose $U \supset M$ where M is a closed, connected $(d-1)$ dimensional manifold that separates $\mathbb{R}^d$ into two connected components M_1 and M_2, define*

$$g^{**}(x) := \begin{cases} \max\{\frac{1}{2} - Ld(x, M), 0\}, & x \in M_1; \\ \frac{1}{2}, & x \in M; \\ \min\{\frac{1}{2} + Ld(x, M), 1\}, & x \in M_2. \end{cases}$$

Then $L(g^{**}) = L(g^*)$, and $I(g^{**}) \leq I(g^*)$.

The proof is in section 2.7.2.

Remark. Corollary 2.1 shows, if one could first establish a result that U contains some simple, separating manifold M (e.g., a hyperplane, a sphere), then M is "essential". However, this does not prove such separating manifold exists.

Theorem 2.1 gives us a better mental picture of the optimal solution. We give here a glimpse of what U and g typically look like in a 2-cluster model, leaving the details to section 2.2-2.4. Under suitable conditions on $\lambda_2, \lambda_3, g^*$ is "bipartite", that is, w.l.o.g, $g^*|_{S_1} < 1/2, g^*|_{S_2} > 1/2$. By continuity of g^*, this implies that U must separate S_1, S_2, so the solution can then be understood as follows: starting from a separating surface U (a $(d-1)$ dimensional manifold) where $g^* = 1/2$, on one side of U where S_1 lies, the value of $1/2 - g^*$ is proportional to the distance from U when g^* is positive, and g^* stays at 0 on the far side; on the other side where S_2 lies, the value of $g^* - 1/2$ is proportional to the distance from U when g^* is less than 1, and g^* stays at 1 on the far side. U contains those most ambiguous points for clustering (since g^* indicates membership probability), the further away from U the clearer membership becomes.

2.1.1 Result for arbitrary level set

We give a result that suggests the U in Theorem 2.1 may be replaced by other level sets. This can be regarded as a refinement of Theorem 2.1.

Corollary 2.2. *Let g^* be an optimal solution that has the form in Theorem 2.1, where $U = \{g^* = 1/2\}, U_1 = \{g^* < 1/2\}, U_2 = \{g^* > 1/2\}$ as before. Then g^* has the following property: for any level set $U_\alpha = \{g^* = \alpha\}, \alpha < 1/2$, and corresponding lower level set $U_{1,\alpha} = \{g^* < \alpha\}$,*

$$g^*(x) = \max\{\alpha - Ld(x, U_\alpha), 0\}, \forall x \in U_{1,\alpha}.$$

Further, we have

$$d(x, U) - d(x, U_\alpha) = \frac{1/2 - \alpha}{L}, \forall x \in U_{1,\alpha}.$$

The proof is in section 2.7.3.

Remark. The upper α level set of g^* ($\alpha < 1/2$) may not be recovered in the same way. Suppose

$$g_\alpha(x) = \begin{cases} \max\{\alpha - Ld(x, U_\alpha), 0\}, & x \in \{g^* < \alpha\} \\ \alpha, & x \in U_\alpha \\ \min\{\alpha + Ld(x, U_\alpha), 1\}, & x \in \{g^* > \alpha\} \end{cases},$$

then g_α may not agree with g^* on $\{g^* > \alpha\}$. In particular, $\{g_\alpha = 1/2\}$ may not be equal to U, see Figure 2.1.

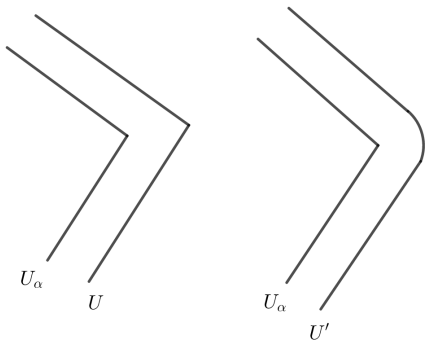


Figure 2.1: Counterexample: $U' = \{g_\alpha = 1/2\} \neq U$ – the lower level sets can be recovered from U by Corollary 2.2, but the reverse may not be true.

2.2 Consistency, part 2: under sharp cluster model

In this section we show that our objective function-based procedure is not only Pollard-type consistent (Theorem 1.3), but also converges to meaningful solutions under some ideal models. The two notions of consistency is distinguished in the next subsection. The next two subsections define the model, and what we mean by clustering risk. The last subsection gives the main result.

2.2.1 Pollard-type consistency v.s. consistency in statistical models

The Pollard-type consistency explained in section 1.5 and 1.6 tells only that the data problem is asymptotic to the ideal one, not that the ideal-problem solution is possible or useful. The same thing is true in Luxburg's theory surrounding spectral clustering [63]. These consistency results do not imply consistency to the clusters in a particular statistical model with a meaningful notion of clusters, while subsequent statistical inference is only possible for the latter notion of consistency. In our case, the difference between the two notions of consistency can also be seen in the proof: proving Theorem 2.3 takes more effort than Theorem 1.3.

Figure 2.2 illustrates issue of Pollard-type consistency: K-means with $K = 2$ cannot consistently estimate the two clusters (disk and annulus) even with infinite amount of data (the figure is generated by $n = 100000$).

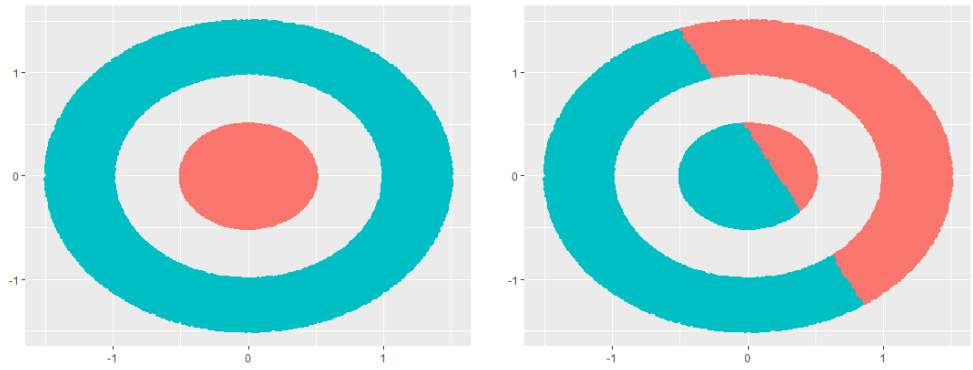


Figure 2.2: Left: 2 clusters–disk and annulus. Right: clustering result by K-means.
This is not surprising, since by the form of K-means criterion, it can only give out linearly separable clusters.

2.2.2 Sharp cluster model

Throughout section 2.2-2.5, consider the generating distribution P to be two sharp clusters $S_1, S_2 \subset \mathbb{R}^d$ with $P(S_1) = \pi_1, P(S_2) = \pi_2 = 1 - \pi_1$, where "sharp" means:

(1) density exists for P and is lower bounded away from 0 on $S_1 \cup S_2$, constant 0 on $(S_1 \cup S_2)^C$

(C1)

(2) S_1, S_2 are compact, connected and disjoint.

Denote $\alpha_0 = \max\{\pi_1, \pi_2\}, L_0 = \frac{1}{d(S_1,S_2)}$. Let $\tilde{g}$ be any Lipschitz function such that

$$\tilde{g}|_{S_1} = 0, \tilde{g}|_{S_2} = 1, \text{ and } \tilde{g}|_{(S_1 \cup S_2)^C} \text{ is a Lipschitz extension [A.5] of } \tilde{g}|_{S_1 \cup S_2}, \text{ so that } L(\tilde{g}) = L_0. \tag{2.4}$$

Remark. "Sharp" density on the clusters may be a strong technical assumption. For some results later in this chapter it will be enough to assume that support$(P_X) = S_1 \cup S_2$. We make this strong assumption here to avoid any potential technical issue we might run into – since our major interest is having this geometric model with general cluster shapes, it will be a distraction to constantly discuss what is the weakest smoothness assumption to make for every theorem and corollary.

Sharp clusters have also been considered in the density clustering literature, we refer to [48] for some related background.

2.2.3 Clustering risk of a clustering function

Recall the classical 0-1 loss function in classification. Let Y be a $\{0, 1\}$-valued random variable indicating class membership. Under $P_{X,Y}$ and 0-1 loss, the classification risk of a clas-

sifier f that assigns point x to class 1 with probability $g(x)$, class 0 with probability $1 - g(x)$ ($g : \mathbb{R}^d \to [0, 1]$, $f : \Omega \times \mathbb{R}^d \to \{0, 1\}$), is

$$R(g) = E[f(X) \neq Y] = E[g(X)I_{[Y=0]} + (1 - g(X))I_{[Y=1]}].$$

In clustering, the risk function should be invariant under permutation of class labels, so for a clustering function g, the clustering risk "induced" by the classification risk is

$R(g) = \min_{\pi \in \mathcal{P}_2} E[f(X) \neq \pi(Y)]$, where $\mathcal{P}_2 = \{\pi : \{0, 1\} \to \{0, 1\}\}$ is a set of permutation functions,

or equivalently,

$$R(g) = \min\{E[f(X) \neq Y], E[f(X) \neq (1 - Y)]\}.$$

In the sharp cluster model, we have $P(Y = 0|X = x, x \in S_1) = P(Y = 1|X = x, x \in S_2) = 1$, so

$$E[f(X) \neq Y] = E[g(X)I_{S_1}(X)] + E[(1 - g(X))I_{S_2}(X)] = E[|g - \tilde{g}|] = ||g - \tilde{g}||_{L_1(P)},$$

it follows that

$$R(g) = \min\{||g - \tilde{g}||_{L_1(P)}, ||g - (1 - \tilde{g})||_{L_1(P)}\}. \tag{2.5}$$

Empirical risk

We define empirical risk under the same setting as above. Let $\{(X_i, Y_i)\}_{i=1}^n$ be a sample from $P_{X,Y}$, the corresponding empirical classification risk under 0-1 loss is

$$R_n(g) = \frac{1}{n}\sum_{i=1}^{n} I\{f(X_i) \neq Y_i\}.$$

The induced empirical clustering risk is

$$R_n(g) = \min_{\pi \in \mathcal{P}_2} \frac{1}{n}\sum_{i=1}^{n} I\{f(X_i) \neq \pi(Y_i)\}. \tag{2.6}$$

The clustering risk and its empirical version are ideal quantities (because they demand true labels), and is not to be confused with I_1 and its empirical version $I_{n,1}$ in the variational approach we take, where

$$I_{n,1}(g) := P_n[g \wedge (1-g)], \tag{2.7}$$

although we will sometimes call this the "classification error term" in the objective function. The latter may be understood as an "unsupervised estimate" to the true (whether population or empirical) risk.

2.2.4 Consistency of clustering risk

For model C1, we show the clustering risk of g_n converges to 0, under some conditions on tuning parameters.

Under model C1, and let C be any constant, the following two theorems holds.

Theorem 2.2 (population version)**.** *Denote*

$$I(g, \lambda_2, \lambda_3) = E[g \wedge (1-g)] + \lambda_2 L(g) + \lambda_3 \max\{E[g], 1 - E[g]\},$$

and let

$$g^*(\cdot, \lambda_2, \lambda_3) \in \underset{g:\mathbb{R}^d \to [0,1]}{\arg\min}\, I(g, \lambda_2, \lambda_3),$$

then

$$\lim_{\substack{\frac{2L_0}{1-\alpha_0}\lambda_2 < \lambda_3 \le C\lambda_2, \\ \lambda_3 \to 0^+}} R(g^*(\cdot, \lambda_2, \lambda_3)) = 0,$$

Remark. This is understood as: for any sequence of solutions g_n^* that are individually optimal for parameters $\lambda_{2,n},\ \lambda_{3,n}$ satisfying the bounds underneath this limit, the corresponding $R(g_n^*)$ values must tend to 0.

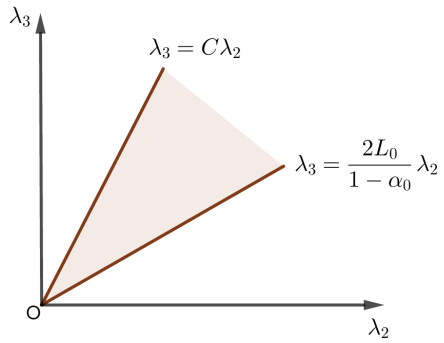


Figure 2.3: consistency cone for (λ_2, λ_3)

Theorem 2.3 (sample version)**.** *Let*

$$I_n(g, \lambda_{2,n}, \lambda_{3,n}) = P_n[g \wedge (1-g)] + \lambda_{2,n} L(g) + \lambda_{3,n} \max\{P_n[g], 1 - P_n[g]\},$$

$$g_n(\cdot, \lambda_{2,n}, \lambda_{3,n}) \in \underset{g:\mathbb{R}^d \to [0,1]}{\arg\min}\ I_n(g, \lambda_{2,n}, \lambda_{3,n}).$$

Suppose $\lambda_{2,n}$ *and* $\lambda_{3,n}$ *are chosen such that*

$$\frac{2L_0}{1-\alpha_0}\lambda_{2,n} < \lambda_{3,n} \leq C\lambda_{2,n}, \lambda_{3,n} \overset{n}{\to} 0, \tag{C2}$$

then

$$\lim_n R(g_n(\cdot, \lambda_{2,n}, \lambda_{3,n})) = 0 \textit{ a.s.}.$$

Proofs of the two theorems is in section 2.7.4 and 2.7.5. Further, we have control on the Lipschitz constant:

Under the same condition on $(\lambda_{2,n}, \lambda_{3,n})$,

$$\limsup_n L(g_n) \leq L_0,$$

and under some further smoothness assumption on P_X,

$$\lim_n L(g_n) = L_0.$$

We refer forward to section 2.7.6 for a formal corollary of this with a formal definition of the smoothness assumption needed to establish the lower bound, which ensures convergence of Lipschitz constant.

2.3 Perfect separation

As λ_2, λ_3 goes to 0 in the way specified by Theorem 2.2, g^* gets arbitrarily close to 0 on S_1 and 1 on S_2. This implies that g^* will first appear as being "bipartite" ($g^*|_{S_1} < 1/2, g^*|_{S_2} > 1/2$) before it gets to 0 and 1 in the limit (either the large sample limit or its population counterpart). This requires weaker condition than the last section.

Theorem 2.4 (Sufficient condition for bipartite g)**.** *Under model C1, suppose λ_2, λ_3 satisfies*

$$\frac{L_0}{1-c-\alpha_0} < \frac{\lambda_3}{\lambda_2} \leq C, \quad 0 < \lambda_2 L_0 + \lambda_3 \alpha_0 \leq c, \tag{C3}$$

where C is any constant, and c is some sufficiently small constant depending on P_X. Then $\frac{1}{2} - g^*$ has different signs on S_1, S_2.

The proof is in section 2.7.7.

Remark. The proof arguments are mostly borrowed from the proof of Theorem 2.2. Theorem 2.4 is thus a weaker result: here λ_2, λ_3 do not have to go to zero as in Theorem 2.2, it is sufficient that they are reasonably small. One question is whether g^* is unique in this case, this will be addressed in section 2.5.

We do not pursue finding the optimal constants in the sufficient condition.

Discussion

Dependence of these conditions (C2, C3) on unknown constants is theoretical. In practice, by resampling, stability plots of the solution against λ_2, λ_3 can be generated to select these tuning

parameters, similar to the cross validation procedure in supervised problems. See section 3.8 in Chapter 3.

2.4 Some examples

In this section, we show how to theoretically derive the optimal U and L assuming bi-partition (λ_2, λ_3 satisfy C3) in simple examples. A general method to derive the optimal U and L is not available at this time, some ideas will be discussed in the appendix section 2.9.

Throughout this section, assume that (λ_2, λ_3) satisfies bipartite condition (C3).

A particular nice property here is convexity. Observe that when $g|_{S_1} < 1/2, g|_{S_2} > 1/2$,

$$I_1(g) = E[gI_{S_1}] + E[(1-g)I_{S_2}],$$

which becomes linear in g. By [A.3] and [A.13], the Lipschitz constant functional is convex, and the third term is convex as well, so

$$I(g) = I_1(g) + \lambda_2 L(g) + \lambda_3 \max\{E[g], 1 - E[g]\}$$

is convex in g. Therefore any locally optimal g is globally optimal.

Remark. This may also be understood as "restricted convexity": a non-convex function can be convex when restricted to a certain region. In particular, if the minimizer can be first shown to lie in a region where the function is convex, we can then do convex analysis as usual.

By convexity, taking an average of two functions or a family of functions will produce a function that gives a smaller $I(g)$ value, if not equal. This is useful to determine U when P_X has

some underlying symmetry. The following gives a good example.

2.4.1 Example 2.1 (hyperplane separation)

Suppose P_X is some distribution supported on S_1, S_2 symmetric about a hyperplane H. We would like to show a "U" of the optimal g is indeed H.

Let g be a candidate function that

$$g(x) = \max\{\frac{1}{2} - Ld(x, U), 0\}, \forall x \in U_1 \supset S_1; g(x) = \min\{\frac{1}{2} + Ld(x, U), 1\}, \forall x \in U_2 \supset S_2,$$

the goal is to construct another function extending from H that is better than g, unless $U = H$ already. We achieve this goal in two steps. Let σ_H denote rigid reflection about H, the reflection of U about H is $\sigma_H(U)$. Define a function g' as

$$g'(x) = 1 - g(\sigma_H(x)),$$

which is built by first reflecting g about H, then "flipping" it (one may find it helpful to first picture this in 1-d when H is a single point). The "U" corresponding to g' is $\sigma_H(U)$, and g' satisfies

$$g'(x) = \max\{\frac{1}{2} - Ld(x, \sigma_H(U)), 0\}, \forall x \in S_1; g'(x) = \min\{\frac{1}{2} + Ld(x, \sigma_H(U)), 1\}, \forall x \in S_2.$$

By definition, U_1, U, U_2 denotes $\{g < 1/2\}, \{g = 1/2\}, \{g > 1/2\}$, respectively. The next observation is that $\sigma_H(U_2), \sigma_H(U), \sigma_H(U_1)$ are $\{g' < 1/2\}, \{g' = 1/2\}, \{g' > 1/2\}$, respectively.

To see this, note that

$$g'(x) > 1/2 \iff g(\sigma_H(x)) < 1/2 \iff \sigma_H(x) \in U_1 \iff x \in \sigma_H(U_1),$$

similarly,

$$g'(x) = 1/2 \iff x \in \sigma_H(U), g'(x) < 1/2 \iff x \in \sigma_H(U_2).$$

By symmetry, g' has the same I_1, I_2, I_3 values as g. If $U \neq H$, then $\frac{g+g'}{2}$ is a better (if not equally good) candidate, by convexity of $I(g)$. The function $\frac{g+g'}{2}$ is the function to which we will apply Theorem 2.1. Before that, we show the "U" of $\frac{g+g'}{2}$ contains H.

Claim: $H \subset \{\frac{g+g'}{2} = 1/2\}$.

Proof. Suppose $x \in H$ and $x \in U_1$, then $\sigma_H(x) = x \in \sigma_H(U_1)$. This implies $g(x) < 1/2, g'(x) > 1/2$, we have

$$g(x) = \max\{\frac{1}{2} - Ld(x, U), 0\}, g'(x) = \min\{\frac{1}{2} + Ld(x, \sigma_H(U)), 1\}.$$

Note that $d(x, U) = d(\sigma_H(x), U) = d(x, \sigma_H(U))$, so

$$\frac{g+g'}{2}(x) = \frac{1}{2} \implies x \in \{\frac{g+g'}{2} = 1/2\}.$$

Similarly, we can prove the claim for any $x \in H \cap U_2$. The case for $x \in H \cap U$ is trivial. □

By Corollary 2.1 (where H plays the role of M in the corollary), we can reconstruct another function that extends from H, shares the same "L" with $\frac{g+g'}{2}$, retains the same form as g (but with different L and U), and is better. Therefore, the "U" of the optimal g is indeed H.

2.4.2 Example 2.2 (1-d)

We characterize g^* in 1-d. Let S_1, S_2 be two disjoint intervals, a, b be the boundary of S_1, S_2 where $d(a,b) = d(S_1, S_2)$, P_X has density $f(x)$. Under suitable conditions on λ_2, λ_3, g^* is bipartite, so we can assume w.l.o.g that $g^*|_{S_1} < 1/2, g^*|_{S_2} > 1/2$. In particular, $g^*(a) < 1/2, g^*(b) > 1/2$. For this case, U reduces to a single point x_0, and the form of g^* can be re-expressed as

$$g^*(x) = \begin{cases} 0, & x \in (-\infty, x_0 - \frac{1}{2L}]; \\ L(x - x_0) + \frac{1}{2}, & x \in (x_0 - \frac{1}{2L}, x_0 + \frac{1}{2L}); \\ 1, & x \in [x_0 + \frac{1}{2L}, \infty), \end{cases}$$

where $x_0 \in (a, b)$. Therefore we are able to write $I(g^*) = I(x_0, L)$, and equivalently optimize for (x_0, L) within the compact region $x_0 - \frac{1}{2L} \le a \le x_0 \le b \le x_0 + \frac{1}{2L}$, which may also be written as $x_0 \in [a, b], L \in [0, \frac{1}{2(x_0 - a)} \wedge \frac{1}{2(b - x_0)}]$, so $I(x_0, L)$ achieves its infimum. For L fixed, we can then take the partial derivative of I with respect to x_0 to find the optimal x_0.

Corollary 2.3. *For fixed L, let $H(x_0) = E[g^*]$, the expectation of g^* with respect to P_X expressed by x_0. Denote the ratio $\frac{P([x_0 - \frac{1}{2L}, a])}{P([b, x_0 + \frac{1}{2L}])} := r(x_0)$, then the optimality condition for x_0 is*

$$r(x_0) = \begin{cases} \frac{1 - \lambda_3}{1 + \lambda_3}, & H(x_0) > \frac{1}{2} \\ \frac{1 + \lambda_3}{1 - \lambda_3}, & H(x_0) < \frac{1}{2} \\ [\frac{1 - \lambda_3}{1 + \lambda_3}, \frac{1 + \lambda_3}{1 - \lambda_3}], & H(x_0) = \frac{1}{2} \end{cases}$$

and there is a unique x_0^* that satisfies the above condition.

The detail is in section 2.7.8.

Remark. 1. The three cases in the optimality condition come from taking the derivative of $\max\{H(x_0), 1 - H(x_0)\}$, i.e., the I_3 term.

2. One can continue to take partial derivative with respect to L to find the optimal L. We will go through this computation in the next example in 2-d.

2.4.3 Example 2.3 (disk and annulus)

Let S_1, S_2 be a disk inside and an annulus outside (as in Figure 2.2, Left), and P_X be rotationally symmetric around origin. Suppose g is one optimal solution of the problem, then by convexity, taking average over all rotations of g will produce another (if not the same) rotational symmetric optimal solution. Therefore, an optimal U is a circle S_r^1 that lies in between the annulus and disk. We have

$$\nabla d(y, U) = \begin{cases} -(y_1, y_2), & y \text{ in annulus} \\ (y_1, y_2), & y \text{ in disk} \end{cases},$$

$$d(y, U) = \begin{cases} ||y|| - r, & y \text{ in annulus} \\ r - ||y||, & y \text{ in disk} \end{cases}.$$

Denote the radius of U by r_U. Then the optimal g and the objective $I(g)$ has the form

$$g(x) = \begin{cases} 0, & r \le r_U - \frac{1}{2L} \\ \frac{1}{2} - L(r_U - r), & r \in [r_U - \frac{1}{2L}, r_U + \frac{1}{2L}] \\ 1, & r \ge r_U + \frac{1}{2L} \end{cases},$$

where $r = ||x||$.

Finding optimal U

Fixing L, with some abuse of notation, we now write the objective function I as a function of r_U:

$$I(r_U) = I_1(r_U) + \lambda_2 L + I_3(r_U),$$

$$\begin{aligned}
I_1(r_U) &= \int_{y\in S_1, ||y||\geq r_U - \frac{1}{2L}} (\frac{1}{2} - L(r_U - ||y||))dP + \int_{y\in S_2, ||y||\leq r_U + \frac{1}{2L}} (\frac{1}{2} - L(||y|| - r_U))dP, \\
E[g] &= \int_{y\in S_1, ||y||\geq r_U - \frac{1}{2L}} (\frac{1}{2} - L(r_U - ||y||))dP + \int_{y\in S_2, ||y||\leq r_U + \frac{1}{2L}} (\frac{1}{2} - L(r_U - ||y||))dP \\
&\quad + \int_{y\in S_2, ||y|| > r_U + \frac{1}{2L}} 1dP \\
&:= h(r_U), \\
I_3(r_U) &= \max\{E[g], 1 - E[g]\} = \max\{h(r_U), 1 - h(r_U)\}.
\end{aligned}$$

Next we will re-express $I(r_U)$ in polar coordinates. Denote the rotational symmetric density $f(r\cos\theta, r\sin\theta) := f(r)$, let r_d, r_a be the radius of disk and radius of the inner circle of annulus, respectively. For example,

$$\begin{aligned}
\int_{y\in S_1, ||y||\geq r_U - \frac{1}{2L}} (\frac{1}{2} - L(r_U - ||y||))dP &= \int_0^{2\pi} \int_{r_U - \frac{1}{2L}}^{r_d} (\frac{1}{2} - L(r_U - r))f(r)rdrd\theta \\
&= 2\pi \int_{r_U - \frac{1}{2L}}^{r_d} (\frac{1}{2} - L(r_U - r))f(r)rdr,
\end{aligned}$$

taking derivative with respect to r_U,

$$\frac{d}{dr_U} \int_{r_U - \frac{1}{2L}}^{r_d} (\frac{1}{2} - L(r_U - r))f(r)rdr = -\int_{r_U - \frac{1}{2L}}^{r_d} Lf(r)rdr.$$

Let $A_1(r_U) = -2\pi L \int_{r_U - \frac{1}{2L}}^{r_d} f(r) r dr$, $A_2(L) = 2\pi L \int_{r_a}^{r_U + \frac{1}{2L}} f(r) r dr$, then

$$I_1'(r_U) = A_1(r_U) + A_2(r_U),$$

$$h'(r_U) = A_1(r_U) - A_2(r_U) := A,$$

$$\partial \max\{h(r_U), 1 - h(r_U)\} = \begin{cases} A, & h(r_U) > \frac{1}{2} \\ -A, & h(r_U) < \frac{1}{2} \\ [-|A|, |A|], & h(r_U) = \frac{1}{2} \end{cases},$$

so we have

$$\partial(I_1 + I_3)(r_U) = \begin{cases} (1+\lambda_3)A_1(r_U) + (1-\lambda_3)A_2(r_U), & h(r_U) = E[g] > \frac{1}{2} \\ (1-\lambda_3)A_1(r_U) + (1+\lambda_3)A_2(r_U), & h(r_U) = E[g] < \frac{1}{2} \\ [\min, \max], & h(r_U) = E[g] = \frac{1}{2} \end{cases},$$

$[\min, \max]$ denotes min and max of the two expressions in the first two cases. A necessary condition for local optimality is $0 \in \partial(I_1 + I_3)(r_U)$, that is,

$$0 \in \begin{cases} (1+\lambda_3)A_1(r_U) + (1-\lambda_3)A_2(r_U), & h(r_U) = E[g] > \frac{1}{2} \\ (1-\lambda_3)A_1(r_U) + (1+\lambda_3)A_2(r_U), & h(r_U) = E[g] < \frac{1}{2} \\ [\min, \max], & h(r_U) = E[g] = \frac{1}{2} \end{cases}$$

Let $T(r_U) = -\frac{A_1}{A_2} = \frac{\int_{r_U - \frac{1}{2L}}^{r_d} rf(r)dr}{\int_{r_a}^{r_U + \frac{1}{2L}} rf(r)dr}$, then the optimal r_U satisfies

$$T(r_U^*) \in \begin{cases} \frac{1-\lambda_3}{1+\lambda_3}, & h(r_U^*) > 1/2 \\ \frac{1+\lambda_3}{1-\lambda_3}, & h(r_U^*) < 1/2 \\ [\frac{1-\lambda_3}{1+\lambda_3}, \frac{1+\lambda_3}{1-\lambda_3}], & h(r_U^*) = 1/2 \end{cases} .$$

Claim: There is a unique r that satisfies the above condition, i.e., one and only one of the three cases can hold true.

Proof. Note that $h(r_U) = E[g]$, as r_U increases, g will decrease, so will $E[g]$, which implies $h(r_U)$ is monotone decreasing in r_U.

Since $T(r)$ is monotone decreasing in r, ranges from $[0, \infty]$, there exist r_1, r_2, r_3 s.t.

$$T(r_1) = \frac{1-\lambda_3}{1+\lambda_3}, T(r_2) = \frac{1+\lambda_3}{1-\lambda_3}, T(r_3) \in [\frac{1-\lambda_3}{1+\lambda_3}, \frac{1+\lambda_3}{1-\lambda_3}].$$

As $h(r)$ is monotone decreasing in r, we have

$$r_1 \geq r_3 \geq r_2, h(r_1) \leq h(r_3) \leq h(r_2).$$

Suppose case 1 in optimality condition holds, i.e., $h(r_1) > \frac{1}{2}$, then $\frac{1}{2} \leq h(r_3) \leq h(r_2)$, which implies the other two cases fail to hold. Similar for case 2. Suppose case 3 is true, i.e., $h(r_3) = \frac{1}{2}$, then $h(r_1) \leq \frac{1}{2}, h(r_2) \geq \frac{1}{2}$, which implies case 1 and 2 cannot hold. Uniqueness in case 3 follows from strict monotonicity of h (so that there is one and only one r such that $h(r) = \frac{1}{2}$). □

Finding optimal L

The analysis is similar. Fixing r_U,

$$I(g) = I(L) = I_1(L) + \lambda_2 L + I_3(L),$$

$$\begin{aligned}
I_1(L) &= \int_{y\in S_1, ||y||\geq r_U - \frac{1}{2L}} (\frac{1}{2} - L(r_U - ||y||))dP + \int_{y\in S_2, ||y||\leq r_U + \frac{1}{2L}} (\frac{1}{2} - L(||y|| - r_U))dP, \\
E[g] &= \int_{y\in S_1, ||y||\geq r_U - \frac{1}{2L}} (\frac{1}{2} - L(r_U - ||y||))dP + \int_{y\in S_2, ||y||\leq r_U + \frac{1}{2L}} (\frac{1}{2} - L(r_U - ||y||))dP \\
&\quad + \int_{y\in S_2, ||y||> r_U + \frac{1}{2L}} 1dP \\
&:= h(L),
\end{aligned}$$

$$I_3(L) = \max\{E[g], 1 - E[g]\} = \max\{h(L), 1 - h(L)\}.$$

Using polar coordinates,

$$\begin{aligned}
\int_{y\in S_1, ||y||\geq r_U - \frac{1}{2L}} (\frac{1}{2} - L(r_U - ||y||))dP &= \int_0^{2\pi} \int_{r_U - \frac{1}{2L}}^{r_d} (\frac{1}{2} - L(r_U - r))f(r)rdrd\theta \\
&= 2\pi \int_{r_U - \frac{1}{2L}}^{r_d} (\frac{1}{2} - L(r_U - r))f(r)rdr.
\end{aligned}$$

So far, these are the same expressions as before, but viewed as functions of L. Now taking derivative with respect to L, e.g.,

$$\frac{d}{dL} \int_{r_U - \frac{1}{2L}}^{r_d} (\frac{1}{2} - L(r_U - r))f(r)rdr = -\int_{r_U - \frac{1}{2L}}^{r_d} (r_U - r)f(r)rdr.$$

Therefore, let $A_1(L) = -2\pi \int_{r_U - \frac{1}{2L}}^{r_d} (r_U - r)f(r)rdr$, $A_2(L) = -2\pi \int_{r_a}^{r_U + \frac{1}{2L}} (r - r_U)f(r)rdr$,

similar to the analysis of r_U, we have

$$\partial(I_1+I_3)(L)=\begin{cases}(1+\lambda_3)A_1(L)+(1-\lambda_3)A_2(L), & h(L)=E[g]>\frac{1}{2}\\ (1-\lambda_3)A_1(L)+(1+\lambda_3)A_2(L), & h(L)=E[g]<\frac{1}{2}\\ [\min,\max], & h(L)=E[g]=\frac{1}{2}\end{cases}.$$

Combining $I_2=\lambda_2 L$, the first order optimality condition for L is

$$-\lambda_2\in\begin{cases}(1+\lambda_3)A_1(L)+(1-\lambda_3)A_2(L), & h(L)=E[g]>\frac{1}{2}\\ (1-\lambda_3)A_1(L)+(1+\lambda_3)A_2(L), & h(L)=E[g]<\frac{1}{2}\\ [\min,\max], & h(L)=E[g]=\frac{1}{2}\end{cases}.$$

Claim: $\partial(I_1+I_3)(L)$ (the R.H.S) is a monotone function of L, so that there is a unique L that satisfies this condition.

Proof. First, note that $A_1(L), A_2(L)$ are both monotone increasing in L. Denote

$$(1+\lambda_3)A_1(L)+(1-\lambda_3)A_2(L):=(1),\quad (1-\lambda_3)A_1(L)+(1+\lambda_3)A_2(L):=(2),$$

we have

$$(1)-(2)=2\lambda_3(A_1(L)-A_2(L)).$$

It suffices to prove that for any neighborhood of L where $A_1(L)>A_2(L)$, $\partial(I_1+I_3)(L)$ is monotone, and the same hold for neighborhoods where $A_1(L)<A_2(L)$.

Note that

$$\frac{d}{dL}h(L) = A_1(L) - A_2(L),$$

so in any neighborhood of L where $A_1(L) > A_2(L)$ (which means $(1) > (2)$), $h(L)$ is monotone increasing, so there is at most one "turning point" L_t on this neighborhood where

$$\partial(I_1 + I_3)(L) = \begin{cases} (2), & L < L_t \\ [(2),(1)], & L = L_t \\ (1), & L > L_t \end{cases}$$

which is monotone increasing since $(1) > (2)$ on this neighborhood. Otherwise if there is no turning point, then $\partial(I_1 + I_3)(L)$ coincides with either (1) or (2) on the entire neighborhood, and monotonicity follows from monotonicity of (1) and (2). □

Remark. We may denote the optimal r_U for a fixed L by $r_{U,L}$, and plug in to solve for L, but the difficulty is that since $r_{U,L}$ does not have closed form in general, the derivative with respect to L will not be explicit. The analysis carried out here basically gives a first order system for L and r, the optimal pair (L, r) (or (L, U), equivalently) can be found by solving this system.

So far, these uniqueness results are special cases. In section 2.5, we will prove a general uniqueness result using a convex combination trick. It shows that in general dimension, the solution is unique up to the part of g^* that comes into (1.5) – e.g., in example 2.1, we may "bend" the hyperplane U (which was proven to be an optimal) from some far-away place without

changing any part of (1.5).

2.5 Uniqueness

This section investigates uniqueness of solution under sharp cluster model (C1) and bipartite condition (C3). As explained in section 2.4, $I(g)$ is convex in $\mathcal{G} := \{g : g|_{S_1} < 1/2, g|_{S_2} > 1/2\}$. Therefore any locally optimal $g \in \mathcal{G}$ is globally optimal. Suppose $I(g)$ is strictly convex (i.e., for any $0 < t < 1, g_1 \neq g_2, I(tg_1 + (1-t)g_2) < tI(g_1) + (1-t)I(g_2)$), then we can conclude that the optimal g is unique. However, it is not clear that this property will hold here, so we instead work on an alternative idea, borrowing strength from our existing results on necessary conditions: to establish uniqueness in a subset containing only the functions that have the form in necessary conditions 1.1 and 2.1. This approach appears well-suited for this problem, leading to a short uniqueness proof in 1-d.

Let $g^* \in \arg\min I(g)$. Recall that we have the following necessary condition for g^* (Theorem 1.1) when $\lambda_3 < 1$ and support$(P_X) = \Omega$:

N.C.1 $g^*(x) = 0$ or $g^*(x) = 1$ or $||\nabla g^*(x)|| = L$, a.e. in Ω, where L is the Lipschitz constant of g^*.

A more precise result is (Theorem 2.1):

N.C.2 Let $U = \{x : g^*(x) = 1/2\}, U_1 = \{x : g^*(x) < 1/2\}, U_2 = \{x : g^*(x) > 1/2\}$, $L = L(g^*)$, then g^* must have the form:

$$g^*(x) = \max\{\frac{1}{2} - Ld(x, U), 0\}, \forall x \in U_1 \cap \Omega; \quad g^*(x) = \min\{\frac{1}{2} + Ld(x, U), 1\}, \forall x \in U_2 \cap \Omega.$$

In the sharp cluster model (C1), $\Omega = S_1 \cup S_2$. The two necessary conditions will be used heavily, denoted by N.C.1 and N.C.2, respectively.

2.5.1 Uniqueness in 1-d

Theorem 2.5. *Assume one-dimensional sharp cluster model C1 where S_1, S_2 are two disjoint intervals. Let $\mathcal{G} = \{g : g|_{S_1} < 1/2, g|_{S_2} > 1/2, g : \mathbb{R} \to [0,1]\}$. Suppose λ_2, λ_3 satisfy (C3), then g^* is unique in $\mathcal{G}$. In 1-d, this means the optimal Lipschitz constant L and optimal U are both unique (U reduces to a single point in 1-d).*

Proof. For $i = 0, 1$, let

$$g_i(x) = \begin{cases} 0, & x \in (-\infty, x_i - \frac{1}{2L_i}]; \\ L_i(x - x_i) + \frac{1}{2}, & x \in (x_i - \frac{1}{2L_i}, x_i + \frac{1}{2L_i}); \\ 1, & x \in [x_i + \frac{1}{2L_i}, \infty) \end{cases}$$

be two candidate functions that $I(g_0) = I(g_1) = I$, $g_0, g_1 \in \mathcal{G}$, x_i's are some points for which $g_i(x_i) = 1/2$. Let $g_t = tg_0 + (1-t)g_1, 0 < t < 1$, then $g_t \in \mathcal{G}$. Suppose g_0, g_1 are both optimal solutions and are distinct–i.e., either $x_0 \neq x_1$ or $L_0 \neq L_1$, then g_t won't have the above form any more (see Figure 2.4).

On the other hand, by convexity (more precisely, restricted convexity of $I(g)$ in $\mathcal{G}$),

$$I(g_t) \leq tI(g_0) + (1-t)I(g_1) = I.$$

This will imply $I(g_t)$ is another optimal, but it does not satisfy N.C.1 (applied to here–any optimal should have only one non-zero derivative value), a contradiction.

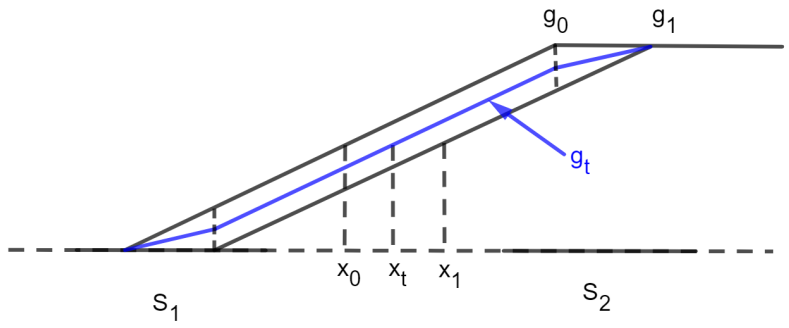


Figure 2.4: $g_t = tg_0 + (1-t)g_1$: g_t will be a "5-piece" function which violates N.C.1

In fact, let x_t be the point where $g_t(x_t) = \frac{1}{2}$, then using the same extension procedure as in (2.1) and (2.2) , we can produce a function g (such that $g(x_t) = \frac{1}{2}$, $L(g) = L(g_t)$, possesses the "3-piece" form above) with a strictly smaller $I(g)$ value, a further contradiction. □

Remark. We have not really used all $0 < t < 1$ to establish the contradiction above. Indeed, pick any $0 < t < 1$, the proof argument will still work. For the more difficult uniqueness proof in general dimension below, the strategy becomes: as long as there exists one t that leads to contradiction, it is a contradiction.

2.5.2 Uniqueness in general dimension

Theorem 2.6. *Assume sharp cluster model C1. Let* $\mathcal{G} = \{g : g|_{S_1} < 1/2, g|_{S_2} > 1/2\}, S = S_1 \cup S_2$, *where* S_1, S_2 *have nonempty interior in* $\mathbb{R}^d$, *and suppose* λ_2, λ_3 *satisfy (C3). Consider any*

$$g^* \in \arg\min_{g \in \mathcal{G}} I(g).$$

Then

1. *The optimal Lipschitz constant is unique.*
2. *The function value on the clusters* $g^*\Big|_S$ *is unique.*

3. Any function that agrees with an optimal g^* in the first two respects is optimal.

The proof is in section 2.7.9.

Remark. Statement 3. identifies the optimal solution as a unique equivalence class in $\mathcal{G}$ according to Lipschitz constant and function value on the clusters.

Remark (about condition on S_1, S_2). The extra condition on S_1, S_2 that they have nonempty interior in $\mathbb{R}^d$ is used in Step 2 of the proof. A remark at the end of Step 2 describes the possibility of generalizing the assumption to having interior in a k-dimensional subspace where $k < d$ and P has k-density, and what are the pieces needed to be modified in the proof.

The following corollary shows that, uniqueness property on clusters can be extended to "in-between" cluster regions, more or less by "rigidity" of our Lipschitz solution. For example, in the disk and annulus case (Example 2 in section 2.4), this implies uniqueness of solution also in the middle annulus which separates the two clusters.

Corollary 2.4. *Suppose the assumptions in Theorem 2.6 hold (so solution is unique on S by Theorem 2.6). Let g^* be any solution (possibly nonunique outside S) with $U = \{g^* = 1/2\}$ and $L = L(g^*)$. Let $A = B_{\frac{1}{2L}}(U)$. For each point x in $A \cap S$ (the union of $A \cap S_1$ and $A \cap S_2$), draw the line segment between x and $y_x = proj_U(x)$. If y_x is not unique, draw all such line segments. Consider the collection of points on U which are shared end points of a pair of line segments drawn respectively from the two clusters in this way, and the region formed by these pairs of line segments. Then g^* is also unique on this "swept over" region.*

The proof is in section 2.7.10.

Remark. The corollary offers an additional "in-between" region of uniqueness on top of Theorem

2.6: starting from any solution, we can form a further region of uniqueness shared by all solutions from these "sweeping normals".

Remark. By Theorem 2.6, $A \cap S$ is unique even though U may not be unique. Corollary 2.4 specifies the part of U that is necessarily unique.

Example of nonuniqueness

Figure 2.5 provides a counterexample where the solution is nonunique everywhere outside the region specified by Corollary 2.4, showing that the result is sharp.

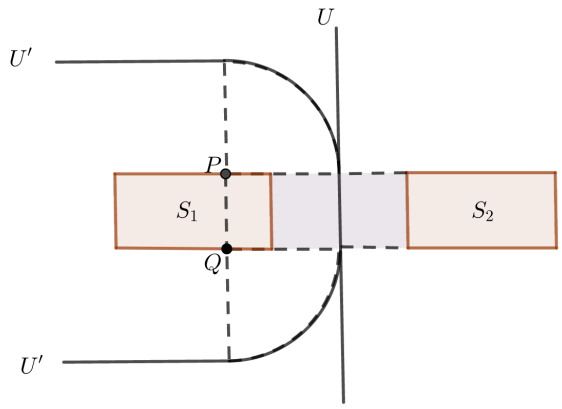


Figure 2.5: Shaded areas (including the two rectangles and the region "between" them) indicate region of uniqueness. The line PQ satisfies $d(x, U) = \frac{1}{2L(g)}$. U' is constructed such that it shares a common segment with U, and $g' = g$ everywhere exactly on $S_1 \cup S_2$. It can be checked (by elementary geometry) that g and g' have different values everywhere outside the shaded region.

2.6 Summary of uniqueness and consistency results

This section collects and clarifies several uniqueness and consistency results in the thesis.

Proved uniqueness

(1) (Major result) Uniqueness on well-separated clusters in general dimension (Chapter 2, Theorem 2.6) is proved under bipartite condition C3. Corollary 2.4 extends the uniqueness to certain "in-between" cluster regions.

(2) (Preliminary result for future direction) For well-separated clusters with noise (Chapter 4), uniqueness is proved in 1-d under a ratio condition on the density lower bound on clusters and density upper bound of noise density (Theorem 4.3). However, this condition requires a gap between the two bounds, and thus does not apply to examples like Gaussian mixtures. A result in general dimension is desired but not yet available.

Nonuniqueness

Figure 2.5 provides a counterexample that for well-separated clusters, the solution can be nonunique everywhere except on or in between clusters. This nonuniqueness is due to the noiseless feature of a well-separated cluster model.

The above rigorously established uniqueness/nonuniqueness results mostly cover the ideal case with well-separated clusters.

Assumed uniqueness

The Pollard-type consistency theorem (Theorem 1.3) states that: assume solution to the variational problem is unique, then the data-based solution is consistent to the (unique) ideal solution.

Remark. The possibly confusing situation lies for example in the implementation of a confidence

band (Figure 3.4) for a presumably unique U (there numerical study goes beyond what has been proved). In such cases (not necessarily Gaussian) where clusters are not well-separated, and which are more realistic situations for a real data set, uniqueness needs to be more or less assumed (to do further things like confidence band) as a general proof argument is not available. This may be confusing because, on the other end, it is also not clear whether there can be nonuniqueness, and we are still anticipating a proof for these noise cases.

Consistency

(1) Pollard-type consistency (Theorem 1.3): the data-based solution is consistent to the ideal solution if the latter is unique. When ideal solution is not unique, the statement becomes convergence to one of the solutions, or convergence to the set of solutions (see [45] Chapter 2 Problem 1). This theorem does not place any condition on P_X.

(2) Model-based consistency (Theorem 2.3): the clustering risk converges to 0 for the sharp cluster model, under some conditions on the tuning parameters.

(3) L_1 consistency vs. pointwise consistency: the consistency in Theorem 1.3 is in L_1. Corollary 1.2 says for any point $x \in supp(P)$, $g_n(x) \xrightarrow{p} g^*(x)$ given that g^* is unique. The pointwise consistency can be extended to some region outside the support by Corollary 2.4 for well-separated clusters.

Subsampling version

Later in Chapter 3, a subsampling version of the solution will be proposed where consistency is maintained for suitable choice of m, B (size of subsample and number of subsamples),

whenever consistency/uniqueness is justified or assumed in the several cases above. See Theorem 3.1 and remarks therein.

2.7 Proofs of chapter 2

2.7.1 Proof of Theorem 2.1

Since g is continuous, and U is the preimage of $\{1/2\}$ under g, then U must be a closed set. Therefore for any x, its distance to U, $d(x,U) = \inf_{u\in U} ||x-u||$ is always achieved at some point $u \in U$, and $d(x,U) = 0$ iff $x \in U$.

Suppose a Lipschitz function g does not have the form specified in Theorem 2.1. Let $U = \{g = 1/2\}, U_1 = \{g < 1/2\}, U_2 = \{g > 1/2\}, L = L(g)$. Let g^* be a modification of g such that

$$g^* = g, \forall x \in U;$$

$$g^*(x) = \max\{\frac{1}{2} - Ld(x,U), 0\}, \forall x \in U_1; g^*(x) = \min\{\frac{1}{2} + Ld(x,U), 1\}, \forall x \in U_2.$$

We will show that $I(g^*) < I(g)$.

First let us show $L(g^*) = L(g)$ (so g^* is Lipschitz continuous).

Claim: For any two points x_1, x_2, $\frac{|g^*(x_1)-g^*(x_2)|}{d(x_1,x_2)} \leq L$.

- Suppose $x_1, x_2 \in U_1$, let $u_1, u_2 \in U$ satisfy $d(x_1,u_1) = d(x_1,U), d(x_2,u_2) = d(x_2,U)$,

$$\frac{g^*(x_1) - g^*(x_2)}{d(x_1,x_2)} = \frac{\max\{1/2 - Ld(x_1,U), 0\} - \max\{1/2 - Ld(x_2,U), 0\}}{d(x_1,x_2)}.$$

When $g^*(x_1) = \max\{1/2 - Ld(x_1,U), 0\} = 0$, $\frac{g^*(x_1)-g^*(x_2)}{d(x_1,x_2)} \leq 0$; when $g^*(x_1) = 1/2 -$

$Ld(x_1, U)$,

$$\begin{aligned}
\frac{g^*(x_1) - g^*(x_2)}{d(x_1, x_2)} &= \frac{1/2 - Ld(x_1, U) - \max\{1/2 - Ld(x_2, U), 0\}}{d(x_1, x_2)} \\
&\leq \frac{1/2 - Ld(x_1, U) - (1/2 - Ld(x_2, U))}{d(x_1, x_2)} \\
&= \frac{L(d(x_2, U) - d(x_1, U))}{d(x_1, x_2)} \\
&\leq L\frac{d(x_2, u_1) - d(x_1, u_1)}{d(x_1, x_2)} \\
&\quad \text{(since } d(x_2, U) \leq d(x_2, u_1), d(x_1, U) = d(x_1, u_1)\text{)} \\
&\leq L\frac{d(x_1, x_2)}{d(x_1, x_2)} = L.
\end{aligned}$$

Therefore $\frac{g^*(x_1)-g^*(x_2)}{d(x_1,x_2)} \leq L$. Similarly, $\frac{g^*(x_2)-g^*(x_1)}{d(x_1,x_2)} \leq L$, so $\frac{|g^*(x_1)-g^*(x_2)|}{d(x_1,x_2)} \leq L, \forall x_1, x_2 \in U_1$.

- Suppose $x_1, x_2 \in U_2$, again let $u_1, u_2 \in U$ satisfy $d(x_1, u_1) = d(x_1, U), d(x_2, u_2) = d(x_2, U)$,

$$\frac{g^*(x_1) - g^*(x_2)}{d(x_1, x_2)} = \frac{\min\{1/2 + Ld(x_1, U), 1\} - \min\{1/2 + Ld(x_2, U), 1\}}{d(x_1, x_2)}.$$

When $g^*(x_2) = \min\{1/2 + Ld(x_2, U), 1\} = 1$, $\frac{g^*(x_1)-g^*(x_2)}{d(x_1,x_2)} \leq 0$. When $g^*(x_2) =$

$\min\{1/2 + Ld(x_2, U), 1\} = 1/2 + Ld(x_2, U)$,

$$\begin{aligned}\frac{g^*(x_1) - g^*(x_2)}{d(x_1, x_2)} &= \frac{\min\{1/2 + Ld(x_1, U), 1\} - (1/2 + Ld(x_2, U))}{d(x_1, x_2)} \\ &\leq \frac{(1/2 + Ld(x_1, U)) - (1/2 + Ld(x_2, U))}{d(x_1, x_2)} \\ &= L\frac{d(x_1, U) - d(x_2, U)}{d(x_1, x_2)} \\ &\leq L\frac{d(x_1, u_2) - d(x_2, u_2)}{d(x_1, x_2)} \\ &\leq L\frac{d(x_1, x_2)}{d(x_1, x_2)} = L.\end{aligned}$$

Therefore $\frac{g^*(x_1)-g^*(x_2)}{d(x_1,x_2)} \leq L$. Similarly, $\frac{g^*(x_2)-g^*(x_1)}{d(x_1,x_2)} \leq L$, so $\frac{|g^*(x_1)-g^*(x_2)|}{d(x_1,x_2)} \leq L, \forall x_1, x_2 \in U_2$.

- Suppose $x_1 \in U_1, x_2 \in U_2$. Note that since $g(x_1) < 1/2, g(x_2) > 1/2$, we have

$$\frac{g^*(x_2) - g^*(x_1)}{d(x_1, x_2)} > 0.$$

$$\begin{aligned}\frac{g^*(x_2) - g^*(x_1)}{d(x_1, x_2)} &= \frac{\min\{\frac{1}{2} + Ld(x_2, U), 1\} - \max\{\frac{1}{2} - Ld(x_1, U), 0\}}{d(x_1, x_2)} \\ &\leq \frac{\frac{1}{2} + Ld(x_2, U) - (\frac{1}{2} - Ld(x_1, U))}{d(x_1, x_2)} \\ &= L\frac{d(x_1, U) + d(x_2, U)}{d(x_1, x_2)}.\end{aligned}$$

Consider the line segment between x_1, x_2. Since $g(x_1) < 1/2, g(x_2) > 1/2$, by continuity

of g this line segment must intersect U at some point x_u. We have

$$d(x_1, x_2) = d(x_1, x_u) + d(x_u, x_2) \geq d(x_1, U) + d(x_2, U).$$

Therefore

$$\frac{g^*(x_2) - g^*(x_1)}{d(x_1, x_2)} \leq L \frac{d(x_1, x_u) + d(x_u, x_2)}{d(x_1, x_2)} = L,$$

This proves $\frac{|g^*(x_2) - g^*(x_1)|}{d(x_1, x_2)} \leq L, \forall x_1 \in U_1, x_2 \in U_2$.

Finally, note that in all of the above arguments, we can extend either U_1 or U_2 to $U_1 \cup U$ and $U_2 \cup U$. This is because for any $x \in U, \max\{\frac{1}{2} - Ld(x, U), 0\} = \min\{\frac{1}{2} + Ld(x, U), 1\} = 1/2$. We have proved for any two points $x_1, x_2, \frac{|g^*(x_1) - g^*(x_2)|}{d(x_1, x_2)} \leq L$, so $L(g^*) \leq L(g)$.

For the other direction of inequality, note that if $U = \mathbb{R}^d$, then $g \equiv 1/2$, so $L(g^*) = L(g) = 0$. When $U \subsetneq \mathbb{R}^d$, one of U_1 or U_2 must be nonempty, let us assume $U_1 \neq \emptyset$. Take any point $x_1 \in U_1$ such that $g^*(x_1) = 1/2 - Ld(x_1, U)$. Such a point must exist, otherwise $g^*(x) = 0$ whenever $g^*(x) < 1/2$, which violates continuity of g^*. Now let $u_1 \in U$ satisfy $d(x_1, u_1) = d(x_1, U)$, then

$$\frac{|g^*(x_1) - g^*(u_1)|}{d(x_1, u_1)} = \frac{|1/2 - Ld(x_1, U) - 1/2|}{d(x_1, u_1)} = \frac{Ld(x_1, u_1)}{d(x_1, u_1)} = L.$$

Therefore the Lipschitz constant L is achieved by g^*, so $L(g^*) \geq L = L(g)$. Putting these together, we can conclude $L(g^*) = L(g)$.

Now we are ready to show $I(g^*) < I(g)$.

For any point $x \in U_1, x_u \in U$, note that

$$\frac{g(x_u) - g(x)}{d(x, x_u)} \leq L,$$

$$\frac{1}{2} - g(x) \leq Ld(x, x_u),$$

$$g(x) \geq \frac{1}{2} - Ld(x, x_u).$$

Taking supremum over all $x_u \in U$ on the R.H.S of the above gives

$$g(x) \geq \frac{1}{2} - L \inf_{x_u \in U} d(x, x_u) = \frac{1}{2} - Ld(x, U).$$

Since $g(x) \in [0, 1]$, we get

$$g(x) \geq \max\{\frac{1}{2} - Ld(x, U), 0\} = g^*(x).$$

Apply similar argument on any point $x \in U_2$, together this shows

$$g^*(x) \leq g(x) < \frac{1}{2}, \forall x \in U_1;\ g^*(x) \geq g(x) > \frac{1}{2}, \forall x \in U_2.$$

Therefore, combining with $L(g^*) = L(g)$,

$$
\begin{aligned}
I(g^*) - I(g) &= I_1(g^*) - I_1(g) + I_3(g^*) - I_3(g) \\
&= E[g^* \wedge (1 - g^*)] - E[g \wedge (1 - g)] + \max\{E[g^*], 1 - E[g^*]\} \\
&\quad - \max\{E[g], 1 - E[g]\} \\
&\leq E[g^* I_{U_1}] + E[(1 - g^*) I_{U_2}] - (E[g I_{U_1}] + E[(1 - g) I_{U_2}]) + \lambda_3 |E[g^* - g]| \\
&\quad \text{(by [A.14])} \\
&\leq E[(g^* - g) I_{U_1}] + E[(g - g^*) I_{U_2}] + \lambda_3 (|E[(g^* - g) I_{U_1}]| + |E[(g^* - g) I_{U_2}]|) \\
&= (1 - \lambda_3) E[(g^* - g) I_{U_1}] + (1 - \lambda_3) E[(g - g^*) I_{U_2}] \\
&\quad \text{(since } g^* - g \leq 0 \text{ on } U_1, g - g^* \leq 0 \text{ on } U_2) \\
&\leq 0,
\end{aligned}
$$

when $\lambda_3 < 1$. Suppose g disagrees with g^* in either U_1 or U_2 at some point x in the support Ω, then by continuity, g will disagree with g^* at least on some neighborhood $B_\epsilon(x)$, $\epsilon > 0$. By definition of support, $P_X(B_\epsilon(x)) > 0$, then the above inequality is strict.

2.7.2 Proof of Corollary 2.1

Since M is closed, and g^{**} is well defined on entire $\mathbb{R}^d$, $L(g^{**}) = L$ can be proved similarly as in Theorem 2.1. Since $M \subset U$, we have $d(x, M) \geq d(x, U)$, so

$$
g^{**} \leq g^* < 1/2, \forall x \in U_1 \cap M_1; \quad g^{**} \geq g^* > 1/2, \forall x \in U_2 \cap M_2.
$$

Note that

$$S_1 \subset U_1 \cap M_1, \quad S_2 \subset U_2 \cap M_2,$$

$$\begin{aligned}
I_1(g^{**}) - I_1(g^*) &= E[(g^{**} - g^*)I_{S_1}] + E[(g^* - g^{**})I_{S_2}] \\
&= E[(g^{**} - g^*)I_{U_1 \cap M_1}] + E[(g^* - g^{**})I_{U_2 \cap M_2}], \\
I_3(g^{**}) - I_3(g^*) &\leq \lambda_3 |E[g^{**} - g^*]| \\
&= \lambda_3 |E[(g^{**} - g^*)I_{U_1 \cap M_1}] + E[(g^{**} - g^*)I_{U_2 \cap M_2}]| \\
&= \lambda_3 (-E[(g^{**} - g^*)I_{U_1 \cap M_1}] - E[(g^* - g^{**})I_{U_1 \cap M_1}]), \\
I(g^{**}) - I(g^*) &= I_1(g^{**}) - I_1(g^*) + I_3(g^{**}) - I_3(g^*) \\
&\leq (1 - \lambda_3)E[(g^{**} - g^*)I_{U_1 \cap M_1}] + (1 - \lambda_3)E[(g^* - g^{**})I_{U_1 \cap M_1}] \\
&\leq 0,
\end{aligned}$$

because the two expectations are both nonpositive.

2.7.3 Proof of Corollary 2.2

First define g_α as

$$g_\alpha(x) = g^*(x), \ \forall x \in U_{1,\alpha}^C;$$

$$g_\alpha(x) = \max\{\alpha - Ld(x, U_\alpha), 0\}, \ \forall x \in U_{1,\alpha}.$$

The goal is to show for any $x \in U_{1,\alpha}$, we have both $g_\alpha(x) \geq g^*(x)$ and $g^*(x) \geq g_\alpha(x)$. One technical point is to justify $L(g_\alpha) = L$ (specifically, in the neighborhood of U_α), this part is proved last.

Part 1. $g_\alpha(x) \geq g^*(x)$ comes from optimality of g^* proved in Theorem 2.1, we go through the derivation again because the other case is similar:

For any $x \in U_1$ and $y \in U$, since

$$\frac{g_\alpha(y) - g_\alpha(x)}{d(x,y)} \leq L,$$

we have

$$\begin{aligned} g_\alpha(x) &\geq g_\alpha(y) - Ld(x,y) \\ &= \frac{1}{2} - Ld(x,y), \quad \text{(by definition of } g_\alpha, \{g_\alpha = 1/2\} = \{g^* = 1/2\} = U) \end{aligned}$$

taking supremum over all $y \in U$ on R.H.S.:

$$\begin{aligned} g_\alpha(x) &\geq \frac{1}{2} - L \inf_{y \in U} d(x,y) \\ &= \frac{1}{2} - Ld(x,U). \end{aligned}$$

Since $g_\alpha(x) \in [0,1]$, we get

$$g_\alpha(x) \geq \max\{\frac{1}{2} - Ld(x,U), 0\} = g^*(x).$$

Part 2. For any $x_\alpha \in U_\alpha, x \in U_{1,\alpha}$, since

$$\frac{g^*(x_\alpha) - g^*(x)}{d(x,x_\alpha)} \leq L,$$

we have

$$g^*(x) \geq g^*(x_\alpha) - Ld(x, x_\alpha)$$

$$= \alpha - Ld(x, x_\alpha),$$

taking supremum over all $x_\alpha \in U_\alpha$:

$$g^*(x) \geq \alpha - L \inf_{x_\alpha \in U_\alpha} d(x, x_\alpha)$$

$$= \alpha - Ld(x, U_\alpha).$$

Since $g^*(x) \in [0, 1]$, we get

$$g^*(x) \geq \max\{\alpha - Ld(x, U_\alpha), 0\} = g_\alpha(x).$$

Part 3. Now, for any $x \in U_{1,\alpha} \cap A$ where $A := \{x : d(x, U) \leq \frac{1}{2L}\}$,

$$g^*(x) = 1/2 - Ld(x, U),$$

$$g_\alpha(x) = \alpha - Ld(x, U_\alpha)$$

$$g^*(x) = g_\alpha(x),$$

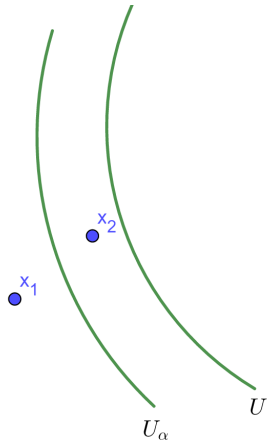
thus $d(x, U) - d(x, U_\alpha) = \frac{1/2-\alpha}{L}, \forall x \in U_{1,\alpha} \cap A$.

Instead of the truncated functions, applying the proof argument to $g^*(x) = 1/2 - Ld(x, U)$

and $g_\alpha(x) = \alpha - Ld(x, U_\alpha)$ for any $x \in U_{1,\alpha}$ yields

$$d(x, U) - d(x, U_\alpha) = \frac{1/2 - \alpha}{L}, \forall x \in U_{1,\alpha}.$$

Part 4. Lastly, we confirm that $L(g_\alpha) = L$. Since the Lipschitz constant of g_α is L when restricted to either the region $U_{1,\alpha}$ or $U_{1,\alpha}^C$, it suffices to look at the case $x_1 \in U_{1,\alpha}, x_2 \in U_1 \cap U_{1,\alpha}^C$ and check $\frac{g_\alpha(x_2) - g_\alpha(x_1)}{d(x_1, x_2)} \leq L$. First,



$$\begin{aligned} g_\alpha(x_2) - g_\alpha(x_1) &= \max\{1/2 - Ld(x_2, U), 0\} - \max\{\alpha - Ld(x_1, U_\alpha), 0\} \\ &\leq \max\{1/2 - Ld(x_2, U) - \alpha + Ld(x_1, U_\alpha), 0\} \\ &\quad (\text{since } \max\{a_1, b_1\} - \max\{a_2, b_2\} \leq \max\{a_1 - a_2, b_1 - b_2\}) \\ &= \max\{1/2 - \alpha - L(d(x_2, U) - d(x_1, U_\alpha)), 0\}. \end{aligned} \tag{2.8}$$

It remains to obtain a lower bound for $d(x_2, U) - d(x_1, U_\alpha)$. For any $y_\alpha \in U_\alpha, y \in U$, we have

$$d(U, U_\alpha) \leq d(y, y_\alpha) \leq d(x_2, y) + d(x_2, y_\alpha),$$

taking infimum over $y \in U$:

$$\begin{aligned} d(U, U_\alpha) &\leq \inf_{y \in U} d(x_2, y) + d(x_2, y_\alpha) \\ &= d(x_2, U) + d(x_2, y_\alpha). \end{aligned}$$

Also, we have

$$\begin{aligned} d(U, U_\alpha) &= \inf_{y \in U, y_\alpha \in U_\alpha} d(y, y_\alpha) \\ &\geq \inf_{y \in U, y_\alpha \in U_\alpha} \frac{g^*(y) - g^*(y_\alpha)}{L} \text{ (since } \frac{g^*(y) - g^*(y_\alpha)}{d(y, y_\alpha)} \leq L) \\ &= \frac{1/2 - \alpha}{L}. \end{aligned}$$

Consider the line segment between x_1, x_2. By continuity of g^*, there is a point on the line segment where $g^* = \alpha$. Now take y_α to be this point (or one of these points), so that $d(x_1, y_\alpha) + d(y_\alpha, x_2) = d(x_1, x_2)$, then

$$\begin{aligned} d(x_2, U) - d(x_1, U_\alpha) &= d(x_2, U) - d(U, U_\alpha) + d(U, U_\alpha) - d(x_1, U_\alpha) \\ &\geq -d(x_2, y_\alpha) + d(U, U_\alpha) - d(x_1, U_\alpha) \\ &\geq -d(x_2, y_\alpha) + d(U, U_\alpha) - d(x_1, y_\alpha) \\ &= d(U, U_\alpha) - d(x_1, x_2) \\ &\geq \frac{1/2 - \alpha}{L} - d(x_1, x_2). \end{aligned}$$

Plug this back to (2.8) to obtain

$$g_\alpha(x_2) - g_\alpha(x_1) \leq 1/2 - \alpha - L(\frac{1/2-\alpha}{L} - d(x_1, x_2))$$
$$= L \cdot d(x_1, x_2),$$

so $\frac{g_\alpha(x_2)-g_\alpha(x_1)}{d(x_1,x_2)} \leq L$. This proves the inequality needed to check Part 4 and completes the proof.

2.7.4 Proof of Theorem 2.2

Denote $g^* = g^*(\cdot, \lambda_2, \lambda_3)$ for convenience. We have

$$I(\tilde{g}) = \lambda_2 L_0 + \lambda_3 \alpha_0 := \epsilon,$$

$$I(g^*) = \arg\min_g I(g) \leq I(\tilde{g}) = \epsilon,$$

and $\epsilon \to 0$ as $\lambda_2 \to 0, \lambda_3 \to 0$. The proof is divided into 3 parts:

When ϵ is small enough,

1. There exists a point x in S_i such that $g^* \wedge (1-g^*)(x) < \epsilon/\pi_i$.
2. $\frac{1}{2} - g^*$ does not change sign within each cluster.
3. $\frac{1}{2} - g^*$ has different signs on the two clusters.

By 1,2,3, we can assume w.l.o.g that $g^*|_{S_1} \leq 1/2, g^*|_{S_2} \geq 1/2$, when ϵ is small enough. Therefore $R(g^*) = E[g^* I_{S_1}] + E[(1-g^*)I_{S_2}] = E[g^* \wedge (1-g^*)] \leq I(g^*) \leq I(\tilde{g}) = \epsilon \to 0$, proving the theorem.

1. Suppose $g^* \wedge (1-g^*) \geq \epsilon/\pi_1$ on S_1, then $E[g^* \wedge (1-g^*)I_{S_1}] \geq \epsilon/\pi_1 \cdot P(S_1) = \epsilon$, so $I(g^*) > I(\tilde{g})$, a contradiction. Similarly for S_2.

The Lipschitz constant of g^* is bounded. When $\lambda_3 < C\lambda_2$, we have

$$\lambda_2 L(g^*) \leq I(g^*) \leq I(\tilde{g}) = \lambda_2 L_0 + \lambda_3 \alpha_0 < \lambda_2 L_0 + \lambda_3 < \lambda_2 L_0 + C\lambda_2,$$

so $L(g^*) \leq L_0 + C$.

2. Define a smoothness parameter on S_1 as $C(L,a,b) = \inf\{\int_{S_1} f dP : f(x_1) = a, f(x_2) = b \text{ for some } x_1, x_2 \in S_1, L(f) \leq L, f : \mathcal{X} \to [0,1]\}$. By sharpness of S_1, $C(L,a,b) = 0$ iff $a = b = 0$. Suppose $\frac{1}{2} - g^*$ changes sign on S_1, then by conclusion of 1, continuity of g^* and connectedness of S_1, there exist two points $x_1, x_2 \in S_1$ such that $g^* \wedge (1-g^*)(x_1) = \epsilon_1/\pi_1, g^*(x_2) = 1/2 = g^* \wedge (1-g^*)(x_2)$. Let ϵ be small enough such that $\epsilon < \min\{C(L_0 + C, 1/4, 1/2), \frac{1}{4}\pi_1\}$. By [A.2], $L(g^* \wedge (1-g^*)) \leq L(g^*) \leq L_0 + C$, so

$$\begin{aligned} I(g^*) \geq E[g^* \wedge (1-g^*)I_{S_1}] &= \int_{S_1} g^* \wedge (1-g^*) dP \\ &\geq C(L_0 + C, \epsilon/\pi_1, 1/2) \geq C(L_0 + C, 1/4, 1/2) > \epsilon. \end{aligned}$$

Therefore $I(g^*) > I(\tilde{g})$, a contradiction. Similarly for S_2.

3. Suppose $\frac{1}{2} - g^*$ have the same sign on S_1, S_2, assume w.l.o.g that $\frac{1}{2} - g^* > 0$, then

$$I(g^*) = E[g^* \wedge (1-g^*)] + \lambda_2 L(g^*) + \lambda_3 \max\{E[g^*], 1 - E[g^*]\} = E[g^*] + \lambda_2 L(g^*) + \lambda_3 (1 - E[g^*]).$$

Therefore we have $E[g^*] < I(g^*) \leq I(\tilde{g}) = \epsilon$. Let $C_2 = \frac{1+\alpha_0}{2}$. In fact, choose any $0 < C_2 < 1$

that satisfies

$$C_2\lambda_3 > \lambda_2 L_0 + \lambda_3\alpha_0.$$

Since $0 < C_2 < 1$, we can let ϵ be small enough so that $1 - \epsilon > C_2$. It follows that

$$I(g^*) \geq \lambda_3(1 - E[g^*]) > \lambda_3(1 - \epsilon) > C_2\lambda_3 > \lambda_2 L + \lambda_3\alpha_0 = I(\tilde{g}),$$

a contradiction. When $C_2 = \frac{1+\alpha_0}{2}$, by rearranging $C_2\lambda_3 > \lambda_2 L_0 + \lambda_3\alpha_0$ we obtain the lower bound assumption on the ratio (which appears in the theorem)

$$\frac{\lambda_3}{\lambda_2} > \frac{2L_0}{1 - \alpha_0}.$$

2.7.5 Proof of Theorem 2.3

$$\begin{aligned} I_n(\tilde{g}) &= \lambda_{2,n}L_0 + \lambda_{3,n}\frac{\max\{\text{number of pts in } S_1, \text{number of pts in } S_2\}}{n} \\ &\leq \lambda_{2,n}L_0 + \lambda_{3,n} := \epsilon_n. \end{aligned}$$

To simplify the notation, we denote $g_n(\cdot, \lambda_{2,n}, \lambda_{3,n})$ by g_n.

Since $g_n \in \arg\min_g I_n(g)$, we have $I_n(g_n) \leq I_n(\tilde{g})$.

The proof is divided into 3 parts:

1. There exists a data point x_n in S_k such that $g_n \wedge (1 - g_n)(x_n) < 2I_n(\tilde{g})/\pi_k$, when n large enough, a.s., and $\lim_n P_n[g_n \wedge (1 - g_n)] = \lim_n P[g_n \wedge (1 - g_n)] = 0$.

2. $\frac{1}{2} - g_n$ does not change sign within each cluster, when n large enough, a.s.

3. $\frac{1}{2} - g_n$ has different signs on the two clusters, when n large enough, a.s.

By 1,2,3, we can assume w.l.o.g that $g_n|_{S_1} \leq 1/2, g_n|_{S_2} \geq 1/2$, when n large enough, a.s.. Therefore $R(g_n) = P[g_n I_{S_1}] + P[(1-g_n)I_{S_2}] = P[g_n \wedge (1-g_n)] \to 0$, a.s., proving the claim.

1. By law of large numbers,

$$\frac{\sum_{i=1}^{n} I_{\{X_i \in S_k\}}}{n} \xrightarrow{a.s.} \pi_k.$$

For any $\delta > 0$, let n be large enough that $\frac{\sum_{i=1}^{n} I_{\{X_i \in S_k\}}}{n} \geq \pi_k - \delta$. Suppose for every data point $X_i \in S_k$, $g_n \wedge (1-g_n)(X_i) \geq 2I_n(\tilde{g})/\pi_k$, then

$$I_n(g_n) \geq \frac{1}{n}\sum_{i=1}^{n} g_n \wedge (1-g_n)(X_i) I_{\{X_i \in S_k\}} \geq \frac{2I_n(\tilde{g})}{\pi_k} \cdot \frac{\sum_{i=1}^{n} I_{\{X_i \in S_k\}}}{n} \geq \frac{2I_n(\tilde{g})}{\pi_k} \cdot (\pi_k - \delta) > I_n(\tilde{g}),$$

by choosing any $\delta < \pi_k/2$. This is a contradiction since g_n is a minimizer of I_n.

Since $I_n(\tilde{g}) \to 0$, and $P_n[g_n \wedge (1-g_n)] \leq I_n(g_n) \leq I_n(\tilde{g})$, we get $P_n[g_n \wedge (1-g_n)] \to 0$.

The Lipschitz constant of g_n is bounded. In fact, when $\lambda_{3,n} \leq C\lambda_{2,n}$, we have

$$\lambda_{2,n} L(g_n) \leq I_n(g_n) \leq I_n(\tilde{g}) \leq \lambda_{2,n} L_0 + \lambda_{3,n} \leq \lambda_{2,n}(L_0 + C),$$

$$L(g_n) \leq L_0 + C,$$

so $L(g_n \wedge (1-g_n)) \leq L(g_n) \leq L_0 + C$ by [A.2].

Let $\mathcal{F} = \{f : \mathcal{X} \to [0,1], L(f) \leq L_0 + C\}$, where $\mathcal{X}$ is some bounded domain on $\mathbb{R}^d$ such that $\cup_k S_k \subset \mathcal{X}$. Then by Lemma 1.2,

$$\sup_{f \in \mathcal{F}} (P_n - P)[f] \to 0 \text{ a.s..}$$

Therefore $P_n[g_n \wedge (1-g_n))] - P[g_n \wedge (1-g_n))] \to 0$, a.s., it follows that $P[g_n \wedge (1-g_n))] \to 0$.

2. If any function f (in particular, $g_n \wedge (1-g_n)$) with a bounded Lipschitz constant takes on two different values (one close to 0 by argument 1, one being $1/2$ suppose $1/2 - f$ changes sign) within a sharp cluster S_k, then its integral $\int f I_{S_k} dP$ will be lower bounded (lower bound depends only on the Lipschitz constant, and $P|_{S_k}$), contradictory to $P[g_n \wedge (1-g_n)] \to 0$.

Specifically, for any $a \in [0,1], b \in [0,1], L > 0$, define

$$C_{S_k}(L,a,b) := \inf\{\int_{S_k} f dP : f(x_1) = a, f(x_2) = b \text{ for some } x_1, x_2 \in S_k, \tag{2.9}$$

$$L(f) \leq L, f : \mathcal{X} \to [0,1]\}.$$

This quantity measures regularity of P, it is decreasing in L, increasing in a and b. By sharpness of S_k, $C_{S_k}(L,a,b) = 0$ iff $a = b = 0$.

Suppose $\frac{1}{2} - g_n$ changes sign within S_k, then by continuity of g_n and connectedness of S_k, there exists a point $x_1 \in S_k$ such that $g_n(x_1) = \frac{1}{2}$. By argument 1, there exists another point $x_2 \in S_k$ such that $g_n \wedge (1-g_n)(x_2) < 2I_n(\tilde{g})/\pi_k$. For any $\epsilon < \frac{1}{2}$, let n be large enough such that $2I_n(\tilde{g})/\pi_k < \epsilon$, we have, by definition of C_{S_k}, $P[g_n \wedge (1-g_n) I_{S_k}] \geq C_{S_k}(L_0 + C, \epsilon, \frac{1}{2}) > 0$, a

contradiction to $P[g_n \wedge (1-g_n)] \to 0$.

3. Proof is shown by contradiction. Suppose $\frac{1}{2} - g_n$ have the same sign on S_1, S_2, note that we can always switch the role of $k=1, k=2$ and in turn switch $g, 1-g$ accordingly, so we can assume w.l.o.g that $\frac{1}{2} - g_n > 0$. Then

$$\begin{aligned} I_n(g_n) &= P_n[g_n \wedge (1-g_n)] + \lambda_{2,n} L(g_n) + \lambda_{3,n} \max\{P_n[g_n], 1 - P_n[g_n]\} \\ &= P_n[g_n] + \lambda_{2,n} L(g_n) + \lambda_{3,n}(1 - P_n[g_n]). \end{aligned}$$

Therefore we have $P_n[g_n] < I_n(g_n) \leq I_n(\tilde{g}) \leq \epsilon_n = \lambda_{2,n} L_0 + \lambda_{3,n}$.

For the rest of proof, note that when $P_n[g_n]$ goes to 0, the third term $\lambda_{3,n}(1 - P_n[g_n]) \approx \lambda_{3,n}$, so this term is much larger than the corresponding term for $\tilde{g}$ ($\approx \lambda_{3,n}\alpha_0$ a.s., where α_0 is the true proportion). When $\lambda_{2,n}$ is controlled by $\lambda_{3,n}$, this indicates $I_n(g_n) > I_n(\tilde{g})$, a contradiction.

Specifically, for some $\delta > 0$, let n be large enough that $P_n[g_n] < \delta$, and

$$\left| \frac{\max\{\text{number of pts in } S_1, \text{number of pts in } S_2\}}{n} - \alpha_0 \right| < \delta.$$

We have

$$\begin{aligned} \lambda_{2,n} L_0 + \lambda_{3,n}(\alpha_0 + \delta) &\geq \lambda_{2,n} L_0 + \lambda_{3,n} \frac{\max\{\text{number of pts in } S_1, \text{number of pts in } S_2\}}{n} \\ &= I_n(\tilde{g}) \geq I_n(g_n) \geq \lambda_{3,n}(1 - P_n[g_n]). \end{aligned}$$

On the other hand,

$$\lambda_{3,n}(1 - P_n[g_n]) - (\lambda_{2,n}L_0 + \lambda_{3,n}(\alpha_0 + \delta)) = \lambda_{3,n}(1 - P_n[g_n] - \alpha_0 - \delta) - \lambda_{2,n}L_0$$
$$> \lambda_{3,n}(1 - 2\delta - \alpha_0) - \lambda_{2,n}L_0,$$

where the last line is non-negative as long as $\frac{\lambda_{3,n}}{\lambda_{2,n}} \geq \frac{L_0}{1-2\delta-\alpha_0}$, and a contradiction will follow. Choose $\delta = \frac{1-\alpha_0}{4}$ (which is quite arbitrary) to obtain the constant $\frac{2L_0}{1-\alpha_0}$ in the theorem.

Remark. The continuous nature of P in model C1 is used in proving the claim in Step 2, see appendix 2.8 for related discussion, and why the same proof may not be applied directly to a discrete P. Here although the data are discrete, but in the limit we are concerned with $P[g_n \wedge (1 - g_n)]$, a quantity involving the continuous P. This is also relevant because the true risk $R(g_n)$ in the statement of the theorem is evaluated on P.

2.7.6 Corollary 2.5 and proof

Corollary 2.5. *For any sequence of $(\lambda_{2,n}, \lambda_{3,n})$ satisfying condition (C2),*

$$\limsup_n L(g^*(\cdot, \lambda_{2,n}, \lambda_{3,n})) \leq L_0,$$

$$\limsup_n L(g_n(\cdot, \lambda_{2,n}, \lambda_{3,n})) \leq L_0.$$

Assume further that there exists $0 < \delta < 1, M > 0$ such that

$$\frac{P(B(x,h) \cap S_i)}{M} \geq \mu(B(x,h) \cap S_i) \geq \delta\mu(B(x,h)), \quad \forall x \in S_i, 0 < h < diam(S_i),\ i = 1, 2, \tag{2.10}$$

then

$$\lim_n L(g^*(\cdot, \lambda_{2,n}, \lambda_{3,n})) = L_0,$$

$$\lim_n L(g_n(\cdot, \lambda_{2,n}, \lambda_{3,n})) = L_0.$$

Remark. The two inequalities in (2.10) holds if we "thicken" any clusters in the following two ways.

The first inequality holds by taking $P_\delta = (1-\delta)P + \delta Unif(S_1 \cup S_2)$, so for any $E \subset S_i$,

$$P_\delta(E) \geq \frac{\delta}{\mu(S_1 \cup S_2)}\mu(E).$$

The second inequality holds by taking $P_\delta = P * Unif(B(0,\delta))$ (where $*$ denotes convolution), supported on $S_1^\delta \cup S_2^\delta$. To see this, for any $x \in S_i^\delta$, there exists a ball $B(x',\delta)$ such that $x \in B(x',\delta)$ and $B(x',\delta) \subset S_i^\delta$. We then have two cases:

1. If $B(x',\delta) \subsetneqq B(x,h)$, then

$$\mu(B(x,h) \cap S_i^\delta) \geq \mu(B(x',\delta)) = (\frac{\delta}{h})^d \mu(B(x',h)) \geq (\frac{\delta}{diam(S_i^\delta)})^d \mu(B(x',h)).$$

2. If the two spheres $S(x',\delta) \cap S(x,h) \neq \emptyset$, take $x'' \in S(x',\delta) \cap S(x,h)$, so $d(x'',x) = h$. Then, the ball with diameter $\overline{xx''}$ is contained in both $B(x',\delta)$ and $B(x,h)$, therefore contained in $B(x,h) \cap S$. The radius of this ball is $\frac{h}{2}$, and

$$\mu(B(x,h) \cap S) \geq \mu(B(0,\frac{h}{2})) = (\frac{1}{2})^d \mu(B(0,h)).$$

Proof of Corollary 2.5.

Step 1. We first show that $L(g^*) \leq L(\tilde{g}) + O(\epsilon)$, where $\epsilon = \epsilon_1 L_0 + \epsilon_2 \alpha_0 = \lambda_2 L_0 + \lambda_3 \alpha_0$, when $\frac{\lambda_3}{\lambda_2} = O(1)$.

We cannot show the opposite direction in general without further smoothness assumptions. To see a counterexample, consider two disjoint, closed disks D_1, D_2, and extend one of them by a line-segment spike on its boundary which points to the in-between area of the two disks. In this case $L(\tilde{g})$ will be much larger than $\frac{1}{d(D_1,D_2)}$, because of the spike. On the other hand, $L(g^*)$ can be close to $\frac{1}{d(D_1,D_2)}$, by allowing g^* to take positive value on the spike which does not change the value of I_1 and I_3.

By Theorem 2.2, when ϵ small enough, we can assume w.l.o.g. that $g^*|_{S_1} < 1/2, g^*|_{S_2} > 1/2$. We have

$$I_1(g^*) = E[g^* I_{S_1}] + E[(1 - g^*) I_{S_2}] < \epsilon,$$

$$0 \leq E[g^* I_{S_1}] < \epsilon, \quad \pi_2 - \epsilon < E[g^* I_{S_2}] \leq \pi_2,$$

$$\pi_2 - \epsilon < E[g^*] = E[g^* I_{S_1}] + E[g^* I_{S_2}] < \pi_2 + \epsilon,$$

$$\pi_1 - \epsilon < 1 - E[g^*] < \pi_1 + \epsilon.$$

Therefore

$$\max\{\pi_1, \pi_2\} - \epsilon < I_3(g^*) = \max\{E[g^*], 1 - E[g^*]\} < \max\{\pi_1, \pi_2\} + \epsilon,$$

$$|I_3(g^*) - I_3(\tilde{g})| = |I_3(g^*) - \alpha_0| < \epsilon.$$

Since $I(g^*) = I_1(g^*) + \lambda_2 L(g^*) + \lambda_3 I_3(g^*) \leq I(\tilde{g}) = \lambda_2 L_0 + \lambda_3 \alpha_0$, we have

$$\begin{aligned}
\lambda_2 L(g^*) + \lambda_3 I_3(g^*) &\leq \lambda_2 L_0 + \lambda_3 \alpha_0, \\
\lambda_2 L(g^*) &\leq \lambda_2 L_0 + \lambda_3 (\alpha_0 - I_3(g^*)), \\
L(g^*) &\leq L_0 + \frac{\lambda_3 (\alpha_0 - I_3(g^*))}{\lambda_2} \\
&\leq L_0 + \frac{\epsilon_2 \cdot \epsilon}{\epsilon_1} \\
&= L_0 + O(\epsilon) \ \ (\text{since } \frac{\epsilon_2}{\epsilon_1} = \frac{\lambda_3}{\lambda_2} = O(1))
\end{aligned}$$

Step 2. Under assumption (2.10), $g^* \wedge (1 - g^*)(x) = O(\epsilon^{1/(d+1)})$ for any $x \in S_i$, and $L(g^*) \geq L_0 - O(\epsilon^{1/(d+1)})$, where d is the dimension.

Subproof of Step 2: Suppose there exists a point $x_i \in S_1$ such that $g^*(x_i) > \epsilon'$, let $h = \frac{\epsilon'}{3L_0}$, then $h < \frac{\epsilon'}{2L(g^*)}$ since $L(g^*) \leq L_0 + O(\epsilon)$. By Lipschitzness of g^*, for any $x \in B(x_i, h) \cap S_i, g^*(x) \geq g^*(x_i) - L(g^*) \cdot h > \epsilon' - L(g^*) \frac{\epsilon'}{2L(g^*)} = \frac{\epsilon'}{2}$, so we have

$$\begin{aligned}
I(g^*) \geq E[g^* I_{S_1}] &\geq \frac{\epsilon'}{2} P(B(x_i, h) \cap S_i) \\
&\geq \frac{\epsilon'}{2} M \mu(B(x_i, h) \cap S_i) \\
&\geq \frac{\delta M \epsilon'}{2} \mu(B(x_i, h)) \\
&= O(\epsilon' h^d) \\
&= O(\epsilon'^{d+1}).
\end{aligned}$$

Therefore choosing $\epsilon' = C\epsilon^{1/(d+1)}$, for some constant C depending only on δ, M, d, will lead to $I(g^*) > \epsilon = I(\tilde{g})$, a contradiction. Apply the same argument to S_2, we have for every point

$x \in S_1, g^*(x) < C\epsilon^{1/(d+1)}$ and for every point $x \in S_2, g^*(x) > 1 - C\epsilon^{1/(d+1)}$, so

$$L(g^*) \geq \frac{1 - O(\epsilon^{1/(d+1)}) - O(\epsilon^{1/(d+1)})}{d(S_1, S_2)} = L_0 - O(\epsilon^{1/(d+1)}).$$

This proves the statement in Step 2.

The corollary is proved by letting $\epsilon \to 0^+$ in Step 1 and Step 2.

□

2.7.7 Proof of Theorem 2.4

The proof arguments are mostly adapted from Theorem 2.2 and 2.3.

a. There exists a point x in S_k such that $g^* \wedge (1 - g^*)(x) < I(\tilde{g})/\pi_k$. Otherwise $I(g^*) \geq E[g^* \wedge (1 - g^*)I_{S_1}] \geq I(\tilde{g})$, a contradiction.

b. Note that $\lambda_2 L(g^*) \leq I(g^*) \leq I(\tilde{g}) = \lambda_2 L_0 + \lambda_3 \alpha_0$, so $L(g^*) \leq L_0 + \frac{\lambda_3}{\lambda_2}\alpha_0 \leq L_0 + C\alpha_0$, and also $L(g^* \wedge (1 - g^*)) \leq L(g^*) \leq L_0 + C\alpha_0$.

c. Suppose $\frac{1}{2} - g^*$ changes sign on S_k, then by a., b. and definition of $C_{S_k}(L, a, b)$ (2.9), we have

$$I(g^*) > E[g^* \wedge (1 - g^*)] \geq C_{S_k}(L_0 + C\alpha_0, \frac{I(\tilde{g})}{\pi_k}, \frac{1}{2}).$$

Define a "normalized" version of this constant

$$\bar{C}_{S_k}(L, a, b) := \inf\{\int_{S_k} f dP, f(x) = |b - a| \text{ for some } x \in S_k, L(f) \leq L, f : \mathcal{X} \to [0, 1]\},$$

(2.11)

for any $a, b \geq 0$. We have $\bar{C}_{S_k}(L, a, b) \leq C_{S_k}(L, a, b)$. Now consider the two functions

$$h_1(x) = \bar{C}_{S_k}(L_0 + C\alpha_0, \frac{x}{\pi_k}, \frac{1}{2}), h_2(x) = x, x \in [0, \frac{\pi_k}{2}].$$

Since $h_1(0) > 0$, decreasing in x and continuous, $h_1(\frac{\pi_k}{2}) = 0$; $h_2(0) = 0$, increasing in x, there is a point where $h_1(x) = h_2(x)$, **denote that point by** c_k. For any $x < c_k$, $h_1(x) > h_2(x)$. Therefore as long as λ_2, λ_3 are small enough that

$$\lambda_2 L_0 + \lambda_3 \alpha_0 = I(\tilde{g}) < c_k,$$

we have

$$I(g^*) \geq C_{S_k}(L_0 + C\alpha_0, \frac{I(\tilde{g})}{\pi_k}, \frac{1}{2}) \geq \bar{C}_{S_k}(L_0 + C\alpha_0, \frac{I(\tilde{g})}{\pi_k}, \frac{1}{2}) = h_1(I(\tilde{g})) > h_2(I(\tilde{g})) = I(\tilde{g}),$$

a contradiction.

Let $c = \min\{c_1, c_2\}$ be the constant that appears in the theorem.

d. Suppose $\frac{1}{2} - g^*$ have the same sign on S_1, S_2, assume w.l.o.g that $g^* < \frac{1}{2}$.

$$I(\tilde{g}) \geq I(g^*) = E[g^*] + \lambda_2 L(g^*) + \lambda_3(1 - E[g^*]) > E[g^*],$$

so $E[g^*] < I(\tilde{g}) \leq c$. On one hand,

$$\lambda_2 L_0 + \lambda_3 \alpha_0 \geq I(g^*) \geq \lambda_3(1 - E[g^*]).$$

On the other hand,

$$\begin{aligned}\lambda_3(1-E[g^*])-(\lambda_2 L_0+\lambda_3\alpha_0) &= \lambda_3(1-E[g^*]-\alpha_0)-\lambda_2 L_0\\ &\geq \lambda_3(1-c-\alpha_0)-\lambda_2 L_0\\ &> 0,\end{aligned}$$

when $\frac{\lambda_3}{\lambda_2} > \frac{L_0}{1-c-\alpha_0}$. This is a contradiction.

c. and d. together shows when $\frac{\lambda_3}{\lambda_2}$ is bounded and λ_2, λ_3 small enough, it must be that $g < \frac{1}{2}$ on one cluster and $g > \frac{1}{2}$ on the other.

2.7.8 Proof of Corollary 2.3

From the form of g^*, we may write $I(g^*) = I(x_0, L)$,

$$\begin{aligned}I(x_0,L) = &\int_{x_0-\frac{1}{2L}}^{a}[L(x-x_0)+\frac{1}{2}]f(x)dx+\int_{b}^{x_0+\frac{1}{2L}}[\frac{1}{2}-L(x-x_0)]f(x)dx+\lambda_2 L\\ &+\lambda_3\max\{H(x_0),1-H(x_0)\},\end{aligned}$$

where $H(x_0) = \int_{x_0-\frac{1}{2L}}^{a}\left(L(x-x_0)+\frac{1}{2}\right)f(x)dx+\int_{b}^{x_0+\frac{1}{2L}}\left(L(x-x_0)+\frac{1}{2}\right)f(x)dx+\int_{x_0+\frac{1}{2L}}^{\infty}1\cdot f(x)dx$.

Fix L and take partial derivative with respect to x_0, using Leibniz integral rule,

$$\begin{aligned}\frac{\partial I}{\partial x_0} &= \left(L \cdot (-\frac{1}{2L}) + \frac{1}{2}\right) f(x_0 - \frac{1}{2L}) + \int_{x_0-\frac{1}{2L}}^{a} -Lf(x)dx \\ &\quad + (\frac{1}{2} - L \cdot \frac{1}{2L}) f(x_0 + \frac{1}{2L}) + \int_{b}^{x_0+\frac{1}{2L}} Lf(x)dx + \lambda_3 h(x_0) \\ &= -L \int_{x_0-\frac{1}{2L}}^{a} f(x)dx + L \int_{b}^{x_0+\frac{1}{2L}} f(x)dx + \lambda_3 h(x_0),\end{aligned}$$

$$\text{where } h(x_0) = \begin{cases} H'(x_0), & H(x_0) > 1/2 \\ -H'(x_0), & H(x_0) < 1/2 \\ [H'(x_0), -H'(x_0)], & H(x_0) = 1/2 \end{cases},$$

$$\begin{aligned}H'(x_0) &= \int_{x_0-\frac{1}{2L}}^{a} (-L)f(x)dx + (L \cdot \frac{1}{2L} + \frac{1}{2}) f(x_0 + \frac{1}{2L}) + \int_{b}^{x_0+\frac{1}{2L}} (-L)f(x)dx \\ &\quad + f(x_0 + \frac{1}{2L}) \cdot (-1) \\ &= -L(\int_{x_0-\frac{1}{2L}}^{a} f(x)dx + \int_{b}^{x_0+\frac{1}{2L}} f(x)dx) < 0.\end{aligned}$$

$$\text{We obtain } \frac{\partial I}{\partial x_0} = \begin{cases} (1+\lambda_3) \int_{x_0-\frac{1}{2L}}^{a} f(x)dx - (1-\lambda_3) \int_{b}^{x_0+\frac{1}{2L}} f(x)dx, & H(x_0) > 1/2 \\ (1-\lambda_3) \int_{x_0-\frac{1}{2L}}^{a} f(x)dx - (1+\lambda_3) \int_{b}^{x_0+\frac{1}{2L}} f(x)dx, & H(x_0) < 1/2 \\ [(1-\lambda_3) \int_{x_0-\frac{1}{2L}}^{a} f(x)dx - (1+\lambda_3) \int_{b}^{x_0+\frac{1}{2L}} f(x)dx, & \\ (1+\lambda_3) \int_{x_0-\frac{1}{2L}}^{a} f(x)dx - (1-\lambda_3) \int_{b}^{x_0+\frac{1}{2L}} f(x)dx], & H(x_0) = 1/2 \end{cases}$$

$$= \begin{cases} (1+\lambda_3) P([x_0 - \frac{1}{2L}, a]) - (1-\lambda_3) P([b, x_0 + \frac{1}{2L}]), & H(x_0) > 1/2 \\ (1-\lambda_3) P([x_0 - \frac{1}{2L}, a]) - (1+\lambda_3) P([b, x_0 + \frac{1}{2L}]), & H(x_0) < 1/2 \\ [(1-\lambda_3) P([x_0 - \frac{1}{2L}, a]) - (1+\lambda_3) P([b, x_0 + \frac{1}{2L}]), & \\ (1+\lambda_3) P([x_0 - \frac{1}{2L}, a]) - (1-\lambda_3) P([b, x_0 + \frac{1}{2L}])], & H(x_0) = 1/2. \end{cases}$$

Denote the ratio $\frac{P([x_0-\frac{1}{2L},a])}{P([b,x_0+\frac{1}{2L}])} := r(x_0)$, then the optimality condition is

$$r(x_0) = \begin{cases} \frac{1-\lambda_3}{1+\lambda_3}, & H(x_0) > \frac{1}{2} \\ \frac{1+\lambda_3}{1-\lambda_3}, & H(x_0) < \frac{1}{2} \\ [\frac{1-\lambda_3}{1+\lambda_3}, \frac{1+\lambda_3}{1-\lambda_3}], & H(x_0) = \frac{1}{2} \end{cases}.$$

Claim: There is a unique x_0^* that satisfies the above condition, i.e., one and only one of the three cases can hold true.

Proof of claim. Since $r(x_0)$ is monotone decreasing in x_0, ranges from $[0, \infty]$, $H(x_0)$ is monotone decreasing in x_0 (as x_0 increases, g as a function of x_0 will decrease, thus $H(x_0) = E[g]$ will decrease), let $r(x_0^1) = \frac{1-\lambda_3}{1+\lambda_3}, r(x_0^2) = \frac{1+\lambda_3}{1-\lambda_3}, r(x_0^3) \in [\frac{1-\lambda_3}{1+\lambda_3}, \frac{1+\lambda_3}{1-\lambda_3}]$, then

$$x_0^1 \geq x_0^3 \geq x_0^2, H(x_0^1) \leq H(x_0^3) \leq H(x_0^2).$$

Suppose case 1 in optimality condition holds, i.e., $H(x_0^1) > \frac{1}{2}$, then $\frac{1}{2} \leq H(x_0^3) \leq H(x_0^2)$, which implies the other two cases fail to hold. Similar for case 2. Suppose case 3 is true, i.e., $H(x_0^3) = \frac{1}{2}$, then $H(x_0^1) \leq \frac{1}{2}, H(x_0^2) \geq \frac{1}{2}$, which implies case 1 and 2 cannot hold. Uniqueness in case 3 follows from strict mononicity of H (so that there is one and only one x_0 such that $H(x_0) = \frac{1}{2}$). □

2.7.9 Proof of Theorem 2.6

For $i = 0, 1$, let g_0, g_1 be two candidate functions that have the form in N.C.2 with $U^{(0)}, U^{(1)}$ be their corresponding level sets at $1/2$, and L_0, L_1 be their Lipschitz constants, respectively. Let

$g_t = tg_0 + (1-t)g_1, 0 < t < 1$. By convexity, for any $0 < t < 1$, g_t is optimal for $I(g)$. We prove uniqueness by contradiction: if $U^{(0)}, U^{(1)}$ do not "agree to some extent", then there exists $0 < t < 1$ such that g_t violates N.C.2. The proof is divided into 4 steps, which are successively stronger statements that make "agree to some extent" more precise. The final conclusion is made after step 4.

To get an analogue of the "5-piece" argument in the 1-d proof, we are going to use the property of the gradient of distance functions given by [A.10], along with N.C.2.

For $r > 0$, denote $B_r(U) := \{x : d(x, U) \leq r\}$. Let $A = B_{\frac{1}{2L_0}}(U^{(0)}), B = B_{\frac{1}{2L_1}}(U^{(1)})$. Let $\triangle$ denote symmetric difference between sets, and $^\circ$ denote the interior of a set. By N.C.2, $0 < g_0(x) < 1$ iff $x \in A^\circ$, $0 < g_1(x) < 1$ iff $x \in B^\circ$. Let $S := S_1 \cup S_2$.

Step 1: $P(A \triangle B) = 0$.

(Sketch of proof of Step 1) Divide $\mathbb{R}^d$ into the following regions $A \backslash B, A \cap B, B \backslash A, A^C \cap B^C$. It can be shown that

$$||\nabla g_t(x)|| = \begin{cases} tL_0, & a.e. \text{ in } A \backslash B \\ ||tL_0 \nabla d(x, U^{(0)}) \pm (1-t)L_1 \nabla d(x, U^{(1)})||, & a.e. \text{ in } A \cap B \\ tL_1, & a.e. \text{ in } B \backslash A \end{cases},$$

and $g_t(x) = 0$ or 1 on $A^C \cap B^C$. By N.C.1, g_t cannot be optimal for every t in $[0, 1]$ unless A and B coincide within the support of P_X. This extends the proof argument in 1-d (Figure 2.4)

to general dimension. The formal proof is given below, taking into account that some of these regions may have zero measure.

(Formal proof of Step 1) Suppose $P(A \triangle B) > 0$, assume w.l.o.g that $P(A \backslash B) > 0$. We deal separately with three cases (at least one of which must hold): $P(A \cap B) > 0$ or $P(B \backslash A) > 0$ or $P(B) = 0$.

When $x \in (A \backslash B) \cap S_1$,

$$g_0(x) = \frac{1}{2} - L_0 d(x, U^{(0)}), g_1(x) = 0, g_t(x) = t g_0(x) = \frac{1}{2}t - t L_0 d(x, U^{(0)});$$

in general, for any $x \in A \backslash B$,

$$g_0(x) = \frac{1}{2} \pm L_0 d(x, U^{(0)}), g_1(x) = 0 \text{ or } 1,$$

$$g_t(x) = t g_0(x) + (1-t) g_1(x) = \frac{1}{2}t \pm t L_0 d(x, U^{(0)}) \text{ or } 1 - \frac{1}{2}t \pm t L_0 d(x, U^{(0)}),$$

so in either case we have, by [A.10],

$$||\nabla g_t(x)|| = t L_0, a.e. \text{ in } A \backslash B. \tag{2.12}$$

We argue in the following paragraphs that when either $P(A \cap B) > 0$ or $P(B \backslash A) > 0$, there is a contradiction to N.C.1 (that non-zero gradient norms must be equal almost everywhere) because of different gradient norms in different regions.

For any $x \in B \backslash A$, similar to $A \backslash B$, we have

$$||\nabla g_t(x)|| = (1-t)L_1, a.e. \text{ in } B \backslash A. \tag{2.13}$$

Since tL_0 and $(1-t)L_1$ cannot be equal for all $0 < t < 1$, it follows that when $P(A \backslash B) > 0$, there exists t such that N.C.1 is violated.

When $x \in A \cap B \cap S_1$,

$$g_0(x) = \frac{1}{2} - L_0 d(x, U^{(0)}), g_1(x) = \frac{1}{2} - L_1 d(x, U^{(1)}),$$

$$g_t(x) = \frac{1}{2} - \{tL_0 d(x, U^{(0)}) + (1-t)L_1 d(x, U^{(1)})\};$$

in general, for any $x \in A \cap B$,

$$g_0(x) = \frac{1}{2} \pm L_0 d(x, U^{(0)}), g_1(x) = \frac{1}{2} \pm L_1 d(x, U^{(1)}),$$

$$g_t(x) = \frac{1}{2} \pm \{tL_0 d(x, U^{(0)}) \pm (1-t)L_1 d(x, U^{(1)})\},$$

so

$$||\nabla g_t(x)|| = ||tL_0 \nabla d(x, U^{(0)}) \pm (1-t)L_1 \nabla d(x, U^{(1)})||, a.e. \text{ in } A \cap B. \tag{2.14}$$

When $P(A \cap B) > 0$, by N.C.1, the gradient norm in (2.12) and (2.14) should be equal. That is, $||\nabla g_t(x)|| = tL_0, a.e.$ in $A \cap B$, so for each t,

$$||tL_0 \nabla d(x, U^{(0)}) \pm (1-t)L_1 \nabla d(x, U^{(1)})|| = tL_0, a.e. \text{ in } A \cap B.$$

Letting $t \to 0$ on both sides of the equality to obtain

$$||L_1 \nabla d(x, U^{(1)})|| = 0, a.e. \text{ in } A \cap B, \tag{2.15}$$

a contradiction to a property of the distance function [A.10] since $(U^{(1)})^C$ has positive probability.

Lastly, in the case $P(B) = 0$, contradiction follows from comparing local gradient norms with the global Lipschitz constant. Note that $L_0, L_1 < \infty$, so A and B have positive Lebesgue measure. Therefore either $A \cap B$ or $B \backslash A$ has positive Lebesgue measure, even though $P(A \cap B) = P(B \backslash A) = 0$. Since $P(A \backslash B) > 0$, by N.C.1,

$$tL_0 = L(g_t).$$

Now because $B \backslash A$ and $A \cap B$ are no longer in the support of P_X, we cannot deduce from N.C.1 that the gradient norms in these two regions needs to equal $L(g_t)$. Nevertheless, we always have that the local gradient norms are bounded by the global Lipschitz constant:

(1) If $m(B \backslash A) > 0$, then

$$tL_0 = L(g_t) \geq ||\nabla g_t(x)|| = (1-t)L_1, a.e. \text{ in } B \backslash A,$$

which cannot be ture for every t.

(2) If $m(A \cap B) > 0$, then

$$tL_0 = L(g_t) \geq ||\nabla g_t(x)|| = ||tL_0 \nabla d(x, U^{(0)}) \pm (1-t)L_1 \nabla d(x, U^{(1)})||, a.e. \text{ in } A \cap B,$$

from which we deduce, by letting $t \to 0$,

$$0 = ||L_1 \nabla d(x, U^{(1)})||, a.e. \text{ in } A \cap B.$$

As in (2.15), this contradicts [A.10].

This completes the proof of Step 1.

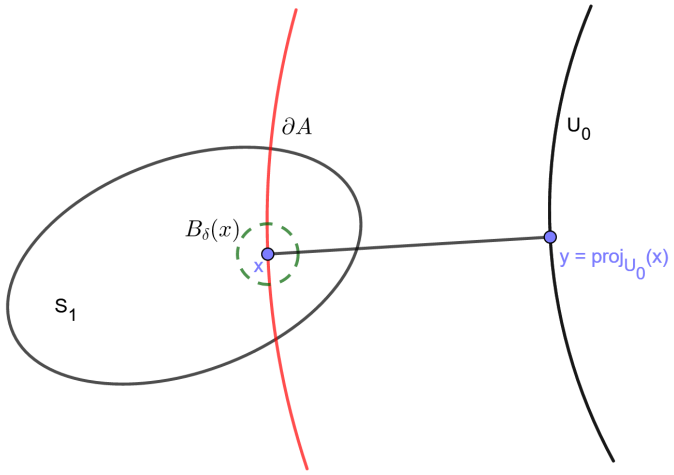
It follows from Step 1 that, for $P(A \cap B) = 0$, both g_0, g_1 satisfy $g\Big|_{S_1} = 0, g\Big|_{S_2} = 1, L(g) = \frac{1}{d(S_1,S_2)}$, and any g that satisfies these is optimal. Therefore, g is optimal iff $g\Big|_{S_1} = 0, g\Big|_{S_2} = 1, L(g) = \frac{1}{d(S_1,S_2)}$. The theorem is proved for the case $P(A \cap B) = 0$.

When $P(A \cap B) \neq 0$ (the two balls "peek into the clusters"), we obtain further information connecting $U^{(0)}, U^{(1)}$ in Step 2, 3, 4.

Step 2: Suppose $P(A \cap B) \neq 0$, then $\nabla d(x, U^{(0)}) = \nabla d(x, U^{(1)})$ a.e. in $A \cap S$.

Our proof of Step 2 uses the assumption about S_1, S_2 in the statement of the theorem: S_1, S_2 have nonempty interior in $\mathbb{R}^d$ (e.g., S_1, S_2 are closures of open sets in $\mathbb{R}^d$). See remark at the end of the proof of Step 2 for how it may be modified to extend to S_1, S_2 having nonempty interior in an affine subspace in $\mathbb{R}^d$ with dimension $k < d$.

By Step 1, $P((A^\circ \cap S_1^\circ) \triangle (B^\circ \cap S_1^\circ)) = P((A^\circ \triangle B^\circ) \cap S_1^\circ) \leq P(A \triangle B) = 0$, which, since P has support on $S_1 \cup S_2$ with density lower bounded, implies $m((A^\circ \cap S_1^\circ) \triangle (B^\circ \cap S_1^\circ)) = 0$ where m is the Lebesgue measure. Since two open sets with Lebesgue-null symmetric difference must be equal, we actually have $A^\circ \cap S_1^\circ = B^\circ \cap S_1^\circ$, so their closures are also the same. Therefore, $A \cap S_1 = B \cap S_1$, $\partial A \cap S_1 = \partial B \cap S_1$. The same assertion can be made on S_2.



The organization in the remaining proof of step 2 will be: In Lemma 2.1, it is shown that we can choose some point from $\partial A \cap S^\circ$ (which justifies the figure above). Further, in Corollary 2.8, it is shown we are able to choose such a point from $\partial A \cap S^\circ$ at which $d_{U^{(0)}}(\cdot)$ is differentiable, while Corollary 2.6 and 2.7 gives the intermediate construction, that there is a ball around this point with certain good properties (which will also be used later in Step 3). The desired equality in the statement of Step 2 will first be shown to hold at the differentiable point we pick on $\partial A \cap S^\circ$, and then extended to $A \cap S$ using N.C.1.

Lemma 2.1. *Suppose g_0 is an optimal solution, and $P(A) \neq 0$, then $\partial A \cap S^\circ$ is nonempty.*

The proof of Lemma 2.1 is long and a bit technical. It will be postponed to section 2.7.11.

By Lemma 2.1, assume w.l.o.g below that there exists a point $x \in \partial A \cap S_1^\circ$. For some $\delta > 0, B_\delta(x) \subset S_1^\circ$, we have $A \cap B_\delta(x) = B \cap B_\delta(x), \partial A \cap B_\delta(x) = \partial B \cap B_\delta(x)$.

The following corollary is a preparation for the ball construction corollary that follows:

Corollary 2.6. *For any point in A^C such that $d_{U^{(0)}}(\cdot)$ is differentiable, we have $\nabla d(\cdot, U^{(0)}) = \nabla d(\cdot, \partial A)$.*

Proof of Corollary 2.6. Apply [A.11] with $x \in A^C, S = U^{(0)}, \partial B_r(S) = \partial A$ $(r = \frac{1}{2L_0}), y = proj_{U^{(0)}}(x)$, and with z defined as the point of intersection of line segment xy with ∂A. We have $\nabla d(x, U^{(0)}) = \nabla d(x, \partial A) = \frac{x-y}{||x-y||}$. □

In the next corollary we introduce our ball construction around the point x chosen above. It is mainly a consequence of Lemma 2.1 and [A.11]:

Corollary 2.7 (ball construction)**.** *Pick any point $x \in \partial A \cap S_1^\circ$, and $\delta > 0$ such that $B_\delta(x) \subset S_1^\circ$. Consider the ball $B_{\delta/3}(x)$. For any point in $B_{\delta/3}(x) \cap A^C$, we have*

- *its closest point to A (or equivalently, ∂A) must lie inside $B_\delta(x)$, so the closest point is inside S_1.*
- *if the point is differentiable, then $\nabla d(\cdot, U^{(0)}) = \nabla d(\cdot, U^{(1)})$.*

Proof of Corollary 2.7. Any two points in $B_{\delta/3}(x)$ have distance at most $2\delta/3$. On the other hand, for a point in $B_{\delta/3}(x)$, its distance to any point outside $B_\delta(x)$ is larger than $2\delta/3$. Therefore, for points in $B_{\delta/3}(x) \cap A^C$, its closest point to A cannot be outside $B_\delta(x)$. See Figure 2.6 for the picture.

For any differentiable point (at which both $d_{U^{(0)}}$ and $d_{U^{(1)}}$ are differentiable) in $B_{\delta/3}(x) \cap A^C$, by Corollary 2.6, $\nabla d(\cdot, U^{(0)}) = \nabla d(\cdot, \partial A)$. Since $B_{\delta/3}(x) \cap A^C \subset S_1 \cap A^C = S_1 \cap B^C \subset B^C$, we have, similarly, $\nabla d(\cdot, U^{(1)}) = \nabla d(\cdot, \partial B)$. Note that $\partial A \cap S = \partial B \cap S$, and $\nabla d(\cdot, \partial A)$ (or $\nabla d(\cdot, \partial B)$) only depends on the closest point to A (or B), which, by the first part of corollary, is inside S. We therefore have $\nabla d(\cdot, \partial A) = \nabla d(\cdot, \partial B)$, so $\nabla d(\cdot, U^{(0)}) = \nabla d(\cdot, U^{(1)})$, for any $x \in B_{\delta/3} \cap A^C$. □

Remark. The construction around the set $B_{\delta/3} \cap A^C$ here will also be used later in Step 3.

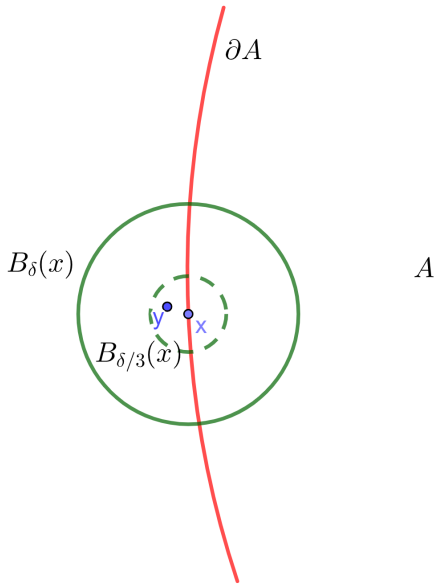


Figure 2.6: Idea behind construction of the ball with radius $\delta/3$: for any point $y \in B_{\delta/3}(x) \cap A^C$, the closest point of y to A must be in $B_\delta(x)$.

Now we are ready to say that we are able to choose a differentiable point in $\partial A \cap S_1^\circ$:

Corollary 2.8. *There exists a point in $\partial A \cap S_1^\circ$ at which $d_{U^{(0)}}(\cdot)$ is differentiable.*

Proof of Corollary 2.8. Since $x \in \partial A$, both $B_\delta(x) \cap A$ and $B_\delta(x) \cap A^C$ have nonempty interior. Note that $d_{U^{(0)}}(\cdot)$ is differentiable a.e. on $B_\delta(x) \cap A^C$, this implies $d_{U^{(0)}}(\cdot)$ is differentiable on all the line segments connecting these differentiable points and their unique projection on $U^{(0)}$ (see [A.11]), which will include some points on $\partial A \cap B_\delta(x)$. To see this, look at the restriction of the function $d(\cdot, U^{(0)})$ to any of these line segments, by its continuity, we have $d(\cdot, U^{(0)}) = \frac{1}{2L_0}$ at some point on the line segment, so these line segments intersect ∂A. It suffices to show that some of these intersections lie in $\partial A \cap B_\delta(x)$. This follows from Corollary 2.7. □

To summarize, we are able to first pick a differentiable (for both $d_{U^{(0)}}$ and $d_{U^{(1)}}$) point w in $A^C \cap S$, whose characteristic (line segment) to $U^{(0)}$ will intersect $\partial A \cap S$ at some point x. By

[A.11], all points on this line segment are differentiable, with the same gradient. Since the two gradients $\nabla d(\cdot, U^{(0)})$ and $\nabla d(\cdot, U^{(1)})$ are equal for w, they are also equal for x. Therefore, we have

$$\nabla d(x, U^{(0)}) = \nabla d(x, U^{(1)}), \tag{2.16}$$

and since $x \in A \cap B$ (A, B are defined as closed balls, and we have shown $\partial A \cap S = \partial B \cap S$), by analysis in Step 1 and (2.16),

$$||\nabla g_t(x)|| = ||tL_0 \nabla d(x, U^{(0)}) + (1-t)L_1 \nabla d(x, U^{(1)})|| = tL_0 + (1-t)L_1,$$

$$L(g_t) \geq ||\nabla g_t(x)|| = tL_0 + (1-t)L_1. \tag{2.17}$$

We are ready to extend (2.16) from this particular point x to almost everywhere in $A \cap S_1$. Denote $\vec{n}_0(\cdot) := \nabla d(\cdot, U^{(0)}), \vec{n}_1(\cdot) := \nabla d(\cdot, U^{(1)})$. For almost every point $x' \in A \cap S_1$,

$$\begin{aligned} ||\nabla g_t(x')|| &= ||tL_0 \vec{n}_0(x') + (1-t)L_1 \vec{n}_1(x')|| \leq tL_0 ||\vec{n}_0(x')|| + (1-t)L_1 ||\vec{n}_1(x')|| \\ &= tL_0 + (1-t)L_1 \\ &= ||\nabla g_t(x)|| \leq L(g_t) \ \text{(by (2.17))}, \end{aligned}$$

where the first inequality holds equal iff $\vec{n}_0(x) = \vec{n}_1(x)$. By N.C.1, for any optimal solution, the nonzero gradient norms should be equal to the Lipschitz constant almost everywhere. Thus, equality holds almost everywhere. Similar argument holds for points in $A \cap S_2$.

This concludes the proof of Step 2.

Remark (lower dimesional clusters). Our proof of Step 2 relies on S_1, S_2 to have interior in the

ambient space $\mathbb{R}^d$. This may be extended to S_1, S_2 being manifolds of dimension $k < d$ lying in $\mathbb{R}^d$. In particular, S_1, S_2 can be closures of open regions in an affine subspace, while P_X is absolutely continuous in that subspace. E.g., S_1, S_2 are two line segments living in $\mathbb{R}^2$, each has interior in a one dimensional affine subspace, while P_X has one-dimensional support on S_1, S_2.

This could be done formally by replacing all interiors appearing in the proof, such as in Lemma 2.1, by interiors in a k dimensional affine subspace; replacing all balls $B_\delta(x)$, such as the ball construction in Corollary 2.7 by k-balls; and replacing Lebesgue measure m by Hausdorff measure H^k, e.g., the statement of Step 2 should be changed to "H^k a.e. in $A \cap S$" accordingly.

Finally, we remark that from the point of view of modeling, one can think of "fattening" the lower dimensional clusters to a narrow tube around it, or consider clusters with a little noise in the ambient space, to circumvent the above mathematical technicality.

Step 3: Suppose $P(A \cap B) \neq 0$. For any $x \in A \cap S, d(x, U^{(0)}) - d(x, U^{(1)}) \equiv \frac{1}{L_0} - \frac{1}{L_1}$.

By Corollary 2.2, for any optimal g with the form in N.C.2 and its corresponding U, we have, for any $\alpha > 0$,

$$d(x, U) - d(x, U_\alpha) = \frac{1/2 - \alpha}{L}, \forall x \in U_{1,\alpha},$$

where $U_\alpha = \{g = \alpha\}, U_{1,\alpha} = \{g < \alpha\}$. Let $\alpha \to 0$, and note that $\lim_{\alpha \to 0} d(x, U_\alpha) = d(x, \partial B_{\frac{1}{2L}}(U))$, we obtain

$$d(x, U) - d(x, \partial B_{\frac{1}{2L}}(U)) = \frac{1}{2L}, \text{ for any } x \in B_{\frac{1}{2L}}(U)^C.$$

Apply this to both g_0 and g_1 (or equivalently, $U^{(0)}$ and $U^{(1)}$):

$$d(x, U^{(0)}) - d(x, \partial A) = \frac{1}{2L_0}, \text{ for any } x \in A^C,$$

$$d(x, U^{(1)}) - d(x, \partial B) = \frac{1}{2L_1}, \text{ for any } x \in B^C,$$

so

$$d(x, U^{(0)}) - d(x, U^{(1)}) = \frac{1}{2L_0} - \frac{1}{2L_1}, \text{ for any } x \in A^C \cap B^C \text{ } s.t. \text{ } d(x, \partial A) = d(x, \partial B). \quad (2.18)$$

Recall the set $B_{\delta/3}(x) \cap A^C$ in Corollary 2.7 within Step 2, where every point satisfies (2.18) (because by Corollary 2.7, their closest points to ∂A or ∂B are inside S_1, and $\partial A \cap S_1 = \partial B \cap S_1$):

$$d(x, U^{(0)}) - d(x, U^{(1)}) = \frac{1}{2L_0} - \frac{1}{2L_1}, \text{ for any } x \in B_{\delta/3}(x) \cap A^C.$$

Now it suffices to extend the property from this small open region to $A \cap S_1$. This is justified by the following lemma:

Lemma 2.2. *Let Ω be an open connected set (or its closure), suppose g is Lipschitz, equal to zero a.e. on an open subset D of Ω, and satisfies $\nabla g = 0$ a.e. in Ω. Then $g \equiv 0$ on Ω.*

The proof will be postponed to section 2.7.12.

Applying Lemma 2.2 with $\Omega = (B_{\delta/3}(x) \cup A) \cap S_1$, $D = B_{\delta/3}(x) \cap A^C$, $g = d(x, U^{(0)}) - d(x, U^{(1)}) - (\frac{1}{L_0} - \frac{1}{L_1})$ yields $g|_\Omega = 0$.

This concludes the proof of Step 3.

Step 4: Suppose $P(A \cap B) \neq 0$. We can show $L_0 = L_1$, and thus for any $x \in A \cap S$, $d(x, U^{(0)}) = d(x, U^{(1)})$.

For any $x \in A \cap B \cap S_1$, $g_0(x) = \frac{1}{2} - L_0 d(x, U^{(0)}), g_1(x) = \frac{1}{2} - L_1 d(x, U^{(1)})$, so

$$\begin{aligned} g_0(x) - g_1(x) &= L_1 d(x, U^{(1)}) - L_0 d(x, U^{(0)}) \\ &= L_1 d(x, U^{(1)}) - L_1 d(x, U^{(0)}) + L_1 d(x, U^{(0)}) - L_0 d(x, U^{(0)}) \\ &= L_1 (d(x, U^{(1)}) - d(x, U^{(0)})) + (L_1 - L_0) d(x, U^{(0)}) \\ &= L_1 (\frac{1}{L_1} - \frac{1}{L_0}) + (L_1 - L_0) d(x, U^{(0)}) \text{ (by Step 3)} \\ &= (L_0 - L_1)(\frac{1}{L_0} - d(x, U^{(0)})). \end{aligned}$$

Since $d(x, U^{(0)}) \leq \frac{1}{2L_0}$, we have $g_0 \geq g_1$ on $S_1 \iff L_0 \geq L_1$. Similarly, $g_0 \leq g_1$ on $S_2 \iff L_0 \geq L_1$. Suppose $L_0 > L_1$ (which means $I_2(g_0) > I_2(g_1)$), then $I_1(g_0) > I_1(g_1)$. By Lemma 1.1, $I(g_0) > I(g_1)$, so g_0 and g_1 cannot be both optimal. We conclude that $L_0 = L_1$, and again by Step 3, $d(x, U^{(0)}) = d(x, U^{(1)}), \forall x \in A \cap S$.

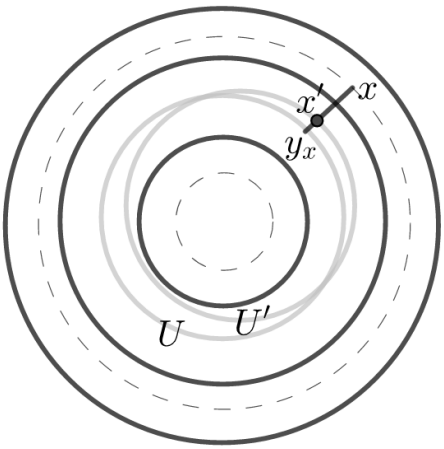
This concludes Step 4.

To summarize, note that Step 4 implies $g_0(x) = g_1(x), \forall x \in S$. In another words, suppose g_0 and g_1 are both optimal solutions, then $g_0\Big|_S = g_1\Big|_S$, and $L(g_0) = L(g_1)$. On the other hand, since $I(g)$ depends on g only through these (function value on the clusters and Lipschitz constant), any function that agrees with g_0 in these two respects are optimal.

2.7.10 Proof of Corollary 2.4

Let $U' = \{g' = 1/2\}$ be the level set at $1/2$ of another solution g', we argue that U' must coincide with U within the prescribed region.

Step 1. By Theorem 2.6, L is unique.



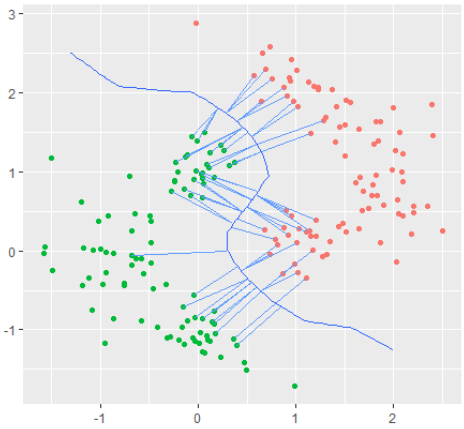


Figure 2.7: Left: illustration of proof of corollary in the disk and annulus case; dashed lines denotes boundaries of A. Right: data illustration of "sweeping normals" and inferred region of uniqueness.

Step 2. For any point $x \in A \cap S$, $g^*(x)$ has the form $g^*(x) = 1/2 - Ld(x, U)$ (on $A \cap S_1$) or $g^*(x) = 1/2 + Ld(x, U)$ (on $A \cap S_2$). Suppose U' intersects with any line segment from x outside U, and suppose x' is one such intersection, then $d(x, U') \leq d(x, x') < d(x, U)$, which implies (if $x \in A \cap S_1$) $g'(x) = 1/2 - Ld(x, U') > 1/2 - Ld(x, U) = g^*(x)$, a contradiction to Theorem 2.6 which states that g^* and g' should have the same function value on the clusters.

Step 3. Let U_N be the collection of points on U which are shared end points of a pair of line segments drawn respectively from the two clusters. Any element in U_N is an intermediate point on a path p (formed by the union of the two line segments) between S_1 and S_2. By Step 2, U' cannot intersect p outside U; on the other hand, $U' \cap p$ cannot be empty by intermediate value theorem, so it must be that U' intersects with p exactly on U_N, i.e., U' coincides with U on p. It follows that function values of g' and g^* are the same everywhere on the pair of line segments. This concludes the proof.

2.7.11 Proof of Lemma 2.1

We show that the objective function I is linear in L when $A \supset S_1 \cup S_2$. This implies that for any fixed $U^{(0)}$, suppose $A \supset S_1 \cup S_2$, then either increasing or decreasing the Lipschitz constant of g_0 alone will give a better solution.

When $A \supset S_1 \cup S_2$, the expression of I simplifies to

$$
\begin{aligned}
I_1 &= \int_{S_1} (\frac{1}{2} - Ld(x, U^{(0)}))dP + \int_{S_2} (\frac{1}{2} - Ld(x, U^{(0)}))dP, \\
I_2 &= \lambda_2 L, \\
E[g_0] &= \int_{S_1} (\frac{1}{2} - Ld(x, U^{(0)}))dP + \int_{S_2} (\frac{1}{2} + Ld(x, U^{(0)}))dP \\
&= \frac{1}{2} + L(\int_{S_2} d(x, U^{(0)})dP - \int_{S_1} d(x, U^{(0)})dP), \\
I_3 &= \max\{E[g_0], 1 - E[g_0]\} = \frac{1}{2} + \\
&\begin{cases}
L(\int_{S_2} d(x, U^{(0)})dP - \int_{S_1} d(x, U^{(0)})dP), & \int_{S_2} d(x, U^{(0)})dP > \int_{S_1} d(x, U^{(0)})dP \\
0, & \int_{S_2} d(x, U^{(0)})dP = \int_{S_1} d(x, U^{(0)})dP \\
-L(\int_{S_2} d(x, U^{(0)})dP - \int_{S_1} d(x, U^{(0)})dP), & \int_{S_2} d(x, U^{(0)})dP < \int_{S_1} d(x, U^{(0)})dP.
\end{cases}
\end{aligned}
$$

For $U^{(0)}$ fixed, the relation between $\int_{S_2} d(x, U^{(0)})dP$ and $\int_{S_1} d(x, U^{(0)})dP$ is fixed, so when $A^\circ \supset S_1 \cup S_2$ (need A° instead of A in order to take derivative, this restriction will be removed a few paragraphs later), we obtain the derivative of I_1 with respect to L:

$$
I_1'(L) = -\int_{S_1} d(x, U^{(0)})dP - \int_{S_2} d(x, U^{(0)})dP, \tag{2.19}
$$

the derivative of I_3 with respect to L:

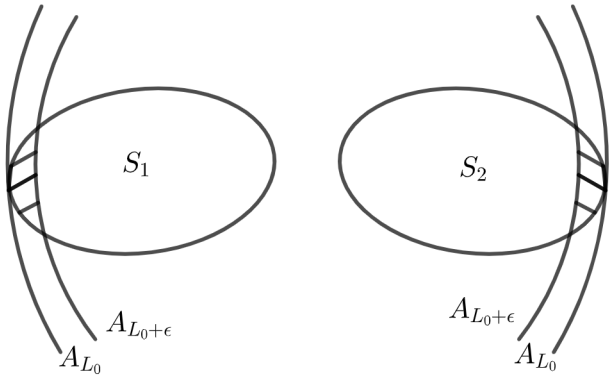
$$I_3'(L) = \begin{cases} \lambda_3(\int_{S_2} d(x, U^{(0)})dP - \int_{S_1} d(x, U^{(0)})dP), & \int_{S_2} d(x, U^{(0)})dP > \int_{S_1} d(x, U^{(0)})dP \\ 0, & \int_{S_2} d(x, U^{(0)})dP = \int_{S_1} d(x, U^{(0)})dP \\ \lambda_3(\int_{S_1} d(x, U^{(0)})dP - \int_{S_2} d(x, U^{(0)})dP), & \int_{S_2} d(x, U^{(0)})dP < \int_{S_1} d(x, U^{(0)})dP, \end{cases} \tag{2.20}$$

the derivative of I with respect to L:

$$I_L' = \lambda_2 + \begin{cases} (\lambda_3 - 1)\int_{S_2} d(x, U^{(0)})dP - (\lambda_3 + 1)\int_{S_1} d(x, U^{(0)})dP, \\ \qquad \text{when } \int_{S_2} d(x, U^{(0)})dP > \int_{S_1} d(x, U^{(0)})dP; \\ -\int_{S_1} d(x, U^{(0)})dP - \int_{S_2} d(x, U^{(0)})dP, \\ \qquad \text{when } \int_{S_2} d(x, U^{(0)})dP = \int_{S_1} d(x, U^{(0)})dP; \\ (\lambda_3 - 1)\int_{S_1} d(x, U^{(0)})dP - (\lambda_3 + 1)\int_{S_2} d(x, U^{(0)})dP, \\ \qquad \text{when } \int_{S_2} d(x, U^{(0)})dP < \int_{S_1} d(x, U^{(0)})dP. \end{cases}$$

In any case, this derivative does not depend on L, so we have, depending on the value of λ_2 and the relation between $\int_{S_2} d(x, U^{(0)})dP$ and $\int_{S_1} d(x, U^{(0)})dP$, either $I_L' > 0$ or $I_L' < 0$ or $I_L' = 0$. This shows linearity with respect to L when $A^\circ \supset S_1 \cup S_2$. Note that when $I_L' = 0$, as we decrease L all the way to 0 the I value is not changed, and eventually we arrive at the constant function $g \equiv 1/2$, which cannot be optimal (since the optimal g should satisfy $g|_{S_1} < 1/2, g|_{S_2} > 1/2$).

The analysis is not yet complete because when ∂A just touches the boundary $S_1 \cup S_2$, further increasing L will let ∂A intersect the interior of $S_1 \cup S_2$, so the right derivative may have a different expression. We analyze this case below.



When $\frac{1}{2L_0} = \sup_{x \in S_1 \cup S_2} d(x, U^{(0)})$, that is, when ∂A just touches the boundary $S_1 \cup S_2$, further increasing L will let ∂A intersect the interior of $S_1 \cup S_2$. For such L_0, with other things being fixed, view I as a function of L and look at $\lim_{\epsilon \to 0^+} \frac{I(L_0+\epsilon)-I(L_0)}{d\epsilon}$ (note that in this case, $\lim_{\epsilon \to 0^-} \frac{I(L_0+\epsilon)-I(L_0)}{d\epsilon}$ is still equal to the expression for I'_L given above). Denote $S = S_1 \cup S_2, A_{L_0} := A = B_{\frac{1}{2L_0}}(U^{(0)}), A_{L_0+\epsilon} = B_{\frac{1}{2(L_0+\epsilon)}}(U^{(0)})$ be the shrunk ball when increasing L by ϵ, and $g_{L_0+\epsilon}$ be the corresponding g function. Note that for any g function such that

$$g(x) = \begin{cases} \max\{\frac{1}{2} - Ld(x, U), 0\}, & x \in S_1 \\ \min\{\frac{1}{2} + Ld(x, U), 1\}, & x \in S_2 \end{cases},$$

we have

$$g \wedge (1-g)(x) = \max\{\frac{1}{2} - Ld(x, U), 0\}, x \in S,$$

and

$$\begin{aligned} I_1 &= E[g \wedge (1-g)] \\ &= \int \max\{\frac{1}{2} - Ld(x, U), 0\} dP \\ &= \int_S \max\{\frac{1}{2} - Ld(x, U), 0\} dP. \end{aligned}$$

Therefore (whether $A_{L_0+\epsilon}$ just touches the interior of S_1, or S_2, or both),

$$\begin{aligned} I_1(L_0+\epsilon) - I_1(L_0) &= \int_S \max\{\frac{1}{2} - (L_0+\epsilon)d(x,U^{(0)}), 0\}dP - \int_S (\frac{1}{2} - L_0 d(x,U^{(0)}))dP \\ &= \int_{S\cap A_{L_0+\epsilon}} (\frac{1}{2} - (L_0+\epsilon))d(x,U^{(0)})dP - \int_S (\frac{1}{2} - L_0 d(x,U^{(0)}))dP \\ &= -\int_{S\cap A_{L_0+\epsilon}} \epsilon \cdot d(x,U^{(0)})dP - \int_{S\cap A^C_{L_0+\epsilon}} (\frac{1}{2} - L_0 d(x,U^{(0)}))dP. \end{aligned}$$

The second term is $o(\epsilon)$: for any $x \in A^C_{L_0+\epsilon}$, $d(x,U^{(0)}) \geq \frac{1}{2(L_0+\epsilon)}$, so $\frac{1}{2} - L_0 d(x,U^{(0)}) \leq \frac{1}{2} - L_0 \cdot \frac{1}{2(L_0+\epsilon)} = \frac{\epsilon}{L_0+\epsilon}$, and

$$\int_{S\cap A^C_{L_0+\epsilon}} (\frac{1}{2} - L_0 d(x,U^{(0)}))dP \leq \frac{\epsilon}{L_0+\epsilon} P(S\cap A^C_{L_0+\epsilon}),$$

$$\frac{\int_{S\cap A^C_{L_0+\epsilon}} (\frac{1}{2} - L_0 d(x,U^{(0)}))dP}{\epsilon} \leq \frac{1}{L_0+\epsilon} P(S\cap A^C_{L_0+\epsilon}) \overset{\epsilon\to 0}{\longrightarrow} 0 \text{ (by absolute continuity of } P).$$

Since $\int_{S\cap A_{L_0+\epsilon}} d(x,U^{(0)})dP \overset{\epsilon\to 0}{\longrightarrow} \int_S d(x,U^{(0)})dP$, we have

$$\frac{I_1(L_0+\epsilon) - I_1(L_0)}{\epsilon} \overset{\epsilon\to 0}{\longrightarrow} -\int_S d(x,U^{(0)})dP.$$

This agrees with the derivative of I_1 with respect to L in (2.19). Next we analyze the I_3 term,

note that

$$
\begin{aligned}
E[g_{L_0+\epsilon}I_{\{x\in S_1\}}] - E[g_0 I_{\{x\in S_1\}}] &= \int_{S_1} \max\{\frac{1}{2} - (L_0+\epsilon)d(x,U^{(0)}), 0\}dP \\
&\quad - \int_{S_1} (\frac{1}{2} - L_0 d(x,U^{(0)}))dP \\
&= -\int_{S_1\cap A_{L_0+\epsilon}} \epsilon \cdot d(x,U^{(0)})dP \\
&\quad - \int_{S_1\cap A^C_{L_0+\epsilon}} (\frac{1}{2} - L_0 d(x,U^{(0)}))dP, \\
E[g_{L_0+\epsilon}I_{\{x\in S_2\}}] - E[g_0 I_{\{x\in S_2\}}] &= \int_{S_2} \min\{\frac{1}{2} + (L_0+\epsilon)d(x,U^{(0)}), 1\}dP \\
&\quad - \int_{S_2} (\frac{1}{2} + L_0 d(x,U^{(0)}))dP \\
&= \int_{S_2\cap A_{L_0+\epsilon}} (\frac{1}{2} + (L_0+\epsilon)d(x,U^{(0)}))dP + \int_{S_2\cap A^C_{L_0+\epsilon}} 1\, dP \\
&\quad - \int_{S_2\cap A_{L_0+\epsilon}} (\frac{1}{2} + L_0 d(x,U^{(0)}))dP \\
&\quad - \int_{S_2\cap A^C_{L_0+\epsilon}} (\frac{1}{2} + L_0 d(x,U^{(0)}))dP \\
&= \int_{S_2\cap A_{L_0+\epsilon}} \epsilon \cdot d(x,U^{(0)})dP + \int_{S_2\cap A^C_{L_0+\epsilon}} (\frac{1}{2} - L_0 d(x,U^{(0)}))dP.
\end{aligned}
$$

From the analysis of I_1 term, we know both $\int_{S_1\cap A^C_{L_0+\epsilon}} (\frac{1}{2} - L_0 d(x,U^{(0)}))dP$ and $\int_{S_2\cap A^C_{L_0+\epsilon}} (\frac{1}{2} - L_0 d(x,U^{(0)}))dP$ are $o(\epsilon)$, thus

$$
\frac{E[g_{L_0+\epsilon}] - E[g_0]}{\epsilon} \xrightarrow{\epsilon\to 0^+} \int_{S_2} d(x,U^{(0)})dP - \int_{S_1} d(x,U^{(0)})dP.
$$

When either $E[g_0] < 1/2$ or $E[g_0] > 1/2$, we have $E[g_{L_0+\epsilon}] < 1/2$ or $E[g_{L_0+\epsilon}] > 1/2$ accordingly when ϵ small enough. When $E[g_0] = 1/2$, note that the g_0 we are considering here still

satisfies $A \supset S_1 \cup S_2$, so

$$E[g_0] = \frac{1}{2} + L(\int_{S_2} d(x, U^{(0)})dP - \int_{S_1} d(x, U^{(0)})dP),$$

and

$$E[g_0] = \frac{1}{2} \iff \int_{S_2} d(x, U^{(0)})dP = \int_{S_1} d(x, U^{(0)})dP,$$

in which case

$$\frac{E[g_{L_0+\epsilon}] - E[g_0]}{\epsilon} \overset{\epsilon \to 0^+}{\longrightarrow} 0,$$

and

$$\Big|\frac{I_3(L_0+\epsilon) - I_3(L_0)}{\epsilon}\Big| \leq \lambda_3 \Big|\frac{E[g_{L_0+\epsilon}] - E[g_0]}{\epsilon}\Big| \overset{\epsilon \to 0^+}{\longrightarrow} 0.$$

Similarly,

$$E[g_0] > \frac{1}{2} \iff \int_{S_2} d(x, U^{(0)})dP > \int_{S_1} d(x, U^{(0)})dP,$$

$$E[g_0] < \frac{1}{2} \iff \int_{S_2} d(x, U^{(0)})dP < \int_{S_1} d(x, U^{(0)})dP,$$

we obtain

$$\frac{I_3(L_0+\epsilon) - I_3(L_0)}{\epsilon} \overset{\epsilon \to 0^+}{\longrightarrow}$$

$$\begin{cases} \lambda_3(\int_{S_2} d(x, U^{(0)})dP - \int_{S_1} d(x, U^{(0)})dP), & \int_{S_2} d(x, U^{(0)})dP > \int_{S_1} d(x, U^{(0)})dP \\ 0, & \int_{S_2} d(x, U^{(0)})dP = \int_{S_1} d(x, U^{(0)})dP \\ \lambda_3(\int_{S_1} d(x, U^{(0)})dP - \int_{S_2} d(x, U^{(0)})dP), & \int_{S_2} d(x, U^{(0)})dP < \int_{S_1} d(x, U^{(0)})dP \end{cases},$$

which also agrees with the derivative of I_3 with respect to L in (2.20). Therefore,

$$\lim_{\epsilon\to 0^+}\frac{I(L_0+\epsilon)-I(L_0)}{\epsilon}=\lambda_2+\lim_{\epsilon\to 0^+}\frac{I_1(L_0+\epsilon)-I_1(L_0)}{\epsilon}+\lim_{\epsilon\to 0^+}\frac{I_3(L_0+\epsilon)-I_3(L_0)}{\epsilon}$$

has exactly the same form as before. This means the right derivative agrees with the left, so the limit exists and from linearity with respect to L, we can conclude when $A \supset S_1 \cup S_2$, the corresponding g function can't be optimal.

2.7.12 Proof of Lemma 2.2

First, consider the following easier problem:

Lemma 2.3. *Let Ω be a nonempty open bounded subset of $\mathbb{R}^n$. Suppose that the function $g : \bar{\Omega} \to \mathbb{R}$ is Lipschitz, equal to zero on the boundary of Ω, and satisfies $\nabla g = 0$ a.e. in Ω. Then g is identically zero.*

Proof of Lemma 2.3. Since g is equal to 0 on the boundary of Ω, one may extend g to all of $\mathbb{R}^n$ where $g|_{\Omega^C} = 0$. Let $E = \{x : \nabla g(x) \neq 0 \text{ or } g \text{ is not differentiable at } x\}, m$ denotes n dimensional Lebesgue measure. By Rademacher's theorem, $m(E) = 0$. Let I_E be indicator function of E, let $(x_1, \cdots, x_n)$ denotes the n coordinates and $y = (x_2, \cdots, x_n)$, we may write

$$m(E) = \int_{\mathbb{R}^n} I_E dm = \int_{\mathbb{R}^{n-1}} dy \int_{\mathbb{R}} I_E(\cdot, y) dx_1.$$

By Fubini's theorem [A.7], $m_1(E_y) = \int_{\mathbb{R}} I_E(\cdot, y)dx_1 = 0, a.e.\ y \in \mathbb{R}^{n-1}$, where m_1 is Lebesgue measure in one dimension, and E_y is the one dimensional slice of E by fixing the last $(n-1)$ coordinates to be y. This says for almost every line in a fixed direction (here in the direction of

the first coordinate), the set of points in E has 0 measure in 1-d. Taking any such line, which must intersect either the boundary of Ω or Ω^C, so that there exists some point x on the line where $g(x) = 0$. For any other point z on the same line, by the fundamental theorem of calculus in 1-d, let $v = (1, 0, \cdots, 0), t_z = z_1 - x_1$,

$$g(z) - g(x) = \int_0^{t_z} \frac{dg}{dt}(x + tv)dt = \int_0^{t_z} \nabla g(x + tv) \cdot v dt = 0,$$

since ∇g is almost everywhere 0 on the line. Thus $g(z) = g(x) = 0$. This shows $g \equiv 0$ on the line $\{x \in \mathbb{R}^n : (x_2, \cdots, x_n) = y\}$, for any y such that $m_1(E_y) = 0$. Since $m_1(E_y) = 0$ holds for almost every $y \in \mathbb{R}^{n-1}$, we get $g \equiv 0$ a.e. on $\mathbb{R}^n$. By continuity of g, $g \equiv 0$ on all of $\mathbb{R}^n$. □

Remark (Remark on the use of Fubini's theorem). Fubini's theorem reduces the n dimensional measure 0 set E to one dimensional measure 0 slices of E, so fundamental theorem of calculus can be applied in that one dimension. In higher dimension, it is unclear whether a Lipschitz function is always absolutely continuous so that versions of fundamental theorem of calculus or Lebesgue decomposition theorem holds.

Remark. Pondering on the above proof, similar technique can be used to deal with situations where the boundary condition is changed to some other conditions. Here is a simple version in $\mathbb{R}^2$ that also works using the Fubini arguments:

Let Ω be a rectangle. If g is Lipschitz and $\nabla g = 0$ a.e. on Ω, $g = 0$ on the intersection of a vertical line with the rectangle (a line segment that "separates" the rectangle), then $g \equiv 0$ on Ω.

Now we are ready to prove Lemma 2.2 using similar idea.

Main proof of Lemma 2.2:

The result holds if Ω is a hyperrectangle $(a_1, b_1) \times \cdots \times (a_n, b_n)$: it suffices to take an open rectangle A inside D. For simplicity, we can choose A such that its edges are all aligned with Ω, that is, $A = (a'_1, b'_1) \times \cdots \times (a'_n, b'_n)$. Let $A_1 = (a_1, b_1) \times (a'_2, b'_2) \cdots \times (a'_n, b'_n)$ be the expansion of A in the first coordinate. Using $g|_A = 0$ and applying the Fubini argument to A_1, we get $g|_{A_1} = 0$. Similarly, by expanding the second coordinate in A_1, and so on until we arrive at Ω after n-steps, yields $g|_\Omega = 0$.

Back to the general case. For any point $x \in D, y \in \Omega \backslash D$, there is a continuous path $p(t), t \in [0, 1]$ from x to y such that $p(0) = x, p(1) = y$. We start from an open rectangle in D surrounding x where $g = 0$, then build a chain of rectangles to extend this property to point y. For every $t \in [0, 1]$ there is an open hyperrectangular neighborhood R_t of the point $p(t)$ that lies in Ω. Since $p[0, 1]$ is closed and bounded, the collection of all these rectangles, which covers the path, has a finite subcover $\{R_{t_k}\}, k = 0, 1, \cdots, N$. Assume w.l.o.g. that $t_0 < \cdots < t_N$, in addition, we may require that $t_0 = 0, t_N = 1$. Note that this subcover forms a chain: all successive rectangles (R_{t_i} and $R_{t_{i+1}}$) intersect pairwise and $R_{t_i} \cap R_{t_{i+1}}$ is open. Take R_{t_1} for example, since R_0 can be chosen such that $g|_{R_0} = 0$, $R_0 \cap R_{t_1}$ is an open set in R_{t_1} where $g = 0$. By the rectangle result in the last paragraph, $g|_{R_{t_1}} = 0$. Do this sequentially along the chain, we eventually get $g|_{R_{t_N}} = 0$. Since $y \in R_{t_N}, g(y) = 0$.

2.8 Appendix: difference between discrete and continuous measure

This section shows that not all results using Lebesgue densities and positive Lebesgue density can extend to discrete P_X. It also draws distinctions between ideal-clustering results and those based on data.

We do a simple 1-d analysis below to show that $\tilde{g}$ (2.4) can be optimal for discrete P_X, but is never optimal for continuous P_X. This also agrees with numerical findings (Figure 2.8).

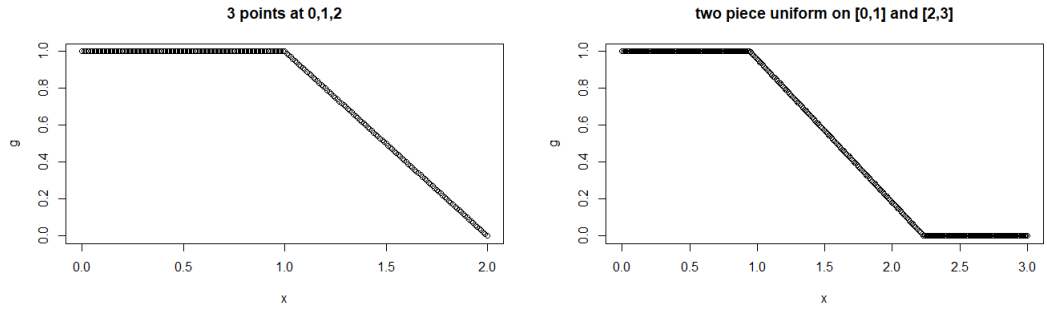


Figure 2.8: Optimal solution under a discrete P_X (left), and a continuous P_X (right). Parameters: (left) $p(0) = 0.2, p(1) = 0.3, p(2) = 0.5, \lambda_2 = 0.1, \lambda_3 = 0.5$; (right) $\lambda_2 = \exp(-3), \lambda_3 = 0.5$.

Continuous case

Suppose P_X is absolutely continuous with density $f(x)$ supported on two disjoint intervals S_1, S_2. For simplicity, assume further that $f(x)$ is symmetric about x_0. Then $\tilde{g}$ is not optimal.

Proof. Rotate $\tilde{g}$ around x_0, so that the Lipschitz constant decreases from L_0 to $L_0 - \epsilon$, for some small $\epsilon > 0$. By symmetry, both the rotated function and $\tilde{g}$ has equal proportions, so the difference in $I(g)$ is

$$\begin{aligned} I(g_{\text{rotate}}) - I(\tilde{g}) &= 2 \int_{x_0 - \frac{1}{2(L_0-\epsilon)}}^{x_0 - \frac{1}{2L_0}} [(L_0 - \epsilon)(x - x_0) + \frac{1}{2}] f(x) dx + \lambda_2 (L_0 - \epsilon - L_0) \\ &\leq [(L_0 - \epsilon) \cdot -\frac{1}{2L_0} + \frac{1}{2}] \cdot P([x_0 - \frac{1}{2(L_0-\epsilon)}, x_0 - \frac{1}{2L_0}]) - \lambda_2 \epsilon \\ &= O(\epsilon^2) - \lambda_2 \epsilon, \text{ (the first term is roughly } \frac{\epsilon}{2L_0} \cdot f(x_0 - \frac{1}{2L_0}) \cdot \frac{\epsilon}{L_0(L_0-\epsilon)}) \end{aligned}$$

which is negative as $\epsilon \to 0^+$, so rotating $\tilde{g}$ locally alone would decrease $I(g)$. Therefore $\tilde{g}$ is not

optimal. □

Discrete case

Suppose P_X has positive probability mass on the boundary of S_1, S_2, so $P([x_0 - \frac{1}{2(L_0-\epsilon)}, x_0 - \frac{1}{2L_0}]) \xrightarrow{\epsilon \to 0} P(\{a\}) > 0$, where $a = x_0 - \frac{1}{2L_0}$ is a boundary point. Still consider a symmetric distribution, now

$$I(g_{\text{rotate}}) - I(\tilde{g}) = 2P(\{a\}) \cdot \frac{\epsilon}{2L_0} - \lambda_2 \epsilon,$$

which is positive as long as $\lambda_2 < \frac{P(\{a\})}{L_0}$.

Remark. This does not say $\tilde{g}$ is optimal yet, only that $I(g_{rotate}) > I(\tilde{g})$ when λ_2 is small enough. Nevertheless, the situation for $\tilde{g}$ being optimal is much better than in the continuous case, and numerical results suggest that this is often the case.

Continuous case, general dimension

We generalize the argument in the above 1-d analysis for continuous case to the general sharp cluster model (C1) with P symmetric about an hyperplane H.

Let $\tilde{g}$ be any function such that $\tilde{g}\Big|_{S_1} = 0, \tilde{g}\Big|_{S_2} = 1, L(\tilde{g}) = L_0$. By Theorem 2.1, we can further require $\tilde{g}$ to satisfy

$$\tilde{g}(x) = \max\{\frac{1}{2} - L_0\, d(x, H), 0\},\ \forall x \in H_1 \supset S_1; \tilde{g}(x) = \min\{\frac{1}{2} + L_0\, d(x, H), 0\},\ \forall x \in H_2 \supset S_2, \tag{2.21}$$

with $d(S_1, H) = d(S_2, H) = \frac{1}{2L_0}$, where H_1, H_2 are the two half spaces separated by H. Similar

to the 1-d case, we "relax" the Lipschitz constant of $\tilde{g}$ by ϵ:

$$\tilde{g}_\epsilon(x) = \max\{\frac{1}{2} - (L_0 - \epsilon)d(x, H), 0\},\ \forall x \in H_1 \supset S_1;$$

$$\tilde{g}_\epsilon(x) = \min\{\frac{1}{2} + (L_0 - \epsilon)d(x, H), 0\},\ \forall x \in H_2 \supset S_2.$$

It suffices to show that the increase in I_1 when relaxing the Lipschitz constant by ϵ is of order $O(\epsilon^p)$ for some $p > 1$. We have

$$\begin{aligned} I_1(\tilde{g}_\epsilon) - I_1(\tilde{g}) &= \int_{x\in S,\ \frac{1}{2L_0}\le d(x,H)\le \frac{1}{2(L_0-\epsilon)}} \epsilon \cdot d(x, H)dP \\ &= \epsilon \int_{x\in S,\ \frac{1}{2L_0}\le d(x,H)\le \frac{1}{2(L_0-\epsilon)}} d(x, H)dP, \end{aligned}$$

where $d(x, H) \overset{\epsilon\to 0}{\longrightarrow} \frac{1}{2L_0}$ in the domain of integration $\{x \in S : \frac{1}{2L_0} \le d(x, H) \le \frac{1}{2(L_0-\epsilon)}\}$. Therefore,

$$\limsup_{\epsilon\to 0^+} \epsilon^{-1}(I(\tilde{g}_\epsilon) - I(\tilde{g})) \le \frac{1}{2L_0} \limsup_{\epsilon\to 0^+} P(\{x \in S : \frac{1}{2L_0} \le d(x, H) \le \frac{1}{2(L_0 - \epsilon)}\}),$$

where the probability mass on the R.H.S is of order $O(\epsilon^k)$ if P has k-dimensional density (in another word, P is absolutely continuous on $S = S_1 \cup S_2$ where S_1, S_2 are k-dimensional in $\mathbb{R}^d$). Thus we obtain

$$I_1(\tilde{g}_\epsilon) - I_1(\tilde{g}) = O(\epsilon^{1+k}),$$

which establishes that $\tilde{g}$ as in (2.21) is not optimal. This also means that any function g such that $g\Big|_{S_1} = 0, g\Big|_{S_2} = 1, L(g) = L_0$ is not optimal, because they all have the same I value as $I(\tilde{g})$.

Comment on the use of connectedness of S_k in the consistency/bipartition result

Proofs of Theorem 2.2-2.4 rely on the fact that if $1/2 - g^*$ changes sign on a single cluster, in particular, if $g^*(x_1) = 0, g^*(x_2) = 1$ for some $x_1, x_2 \in S_k$, then by Lipschitz continuity of g^* and intermediate value theorem (and that S_k is connected), we have $g^*(x) = 1/2$ for some $x \in S_k$. This implies $I_1(g^*)$ is large, and so such g^* cannot be optimal, after comparison with some other functions such as $\tilde{g}$. It is not the case for discrete P_X. For example, if S_1 contains only two support points x, y, and $g^*(x) = 0, g^*(y) = 1$, then $I_1(g^*) = 0$.

2.9 Appendix: determine optimal surface U

The results here are incomplete, so it is put in appendix.

Now that some special cases are dealt with in section 2.4, we would really like to step towards a general method to characterize the surfaces or curves U for an optimal g-function. One idea is to find useful one-dimensional perturbations leading to necessary conditions for optimality. Here we consider the simplest perturbation: *rigid motions*.

Let S_1, S_2 denote the (well-separated) clusters in d dimensions, and suppose U is a $(d-1)$-dimensional oriented manifold separating them.

Translation:

Consider for any unit vector $v \in \mathbb{R}^d$,

$$U_v(\epsilon) \equiv \{u + \epsilon \cdot v : \quad u \in U\},$$

and look at the integrals

$$H_k(U, L, f_X) := \int_{S_k} \max\{0, \min\{1, \frac{1}{2} + (-1)^k L \cdot d(x, U)\}\}\, f_X(x)\, dx,$$

when U is replaced by $U_v(\epsilon)$. This replacement corresponds to a small rigid motion of U by translation. Because the rigid motion is a distance-preserving map on $\mathbb{R}^d$, its effect can be handled by a change of variables in the integrals. When the cluster S_2 is everywhere more than ϵ distant from U,

$$\begin{aligned} H_k(U_v(\epsilon), L, f_X) &= \int_{S_k} \max\{0, \min\{1, \frac{1}{2} + (-1)^k L \cdot d(x, U + \epsilon \cdot v)\}\} f_X(x) dx \\ &= \int_{S_k} \max\{0, \min\{1, \frac{1}{2} + (-1)^k L \cdot d(y + \epsilon \cdot v, U + \epsilon \cdot v)\}\} f_X(y + \epsilon \cdot v) dy \\ &= H_k(U, L, f_X(\cdot + \epsilon v)), \end{aligned}$$

because $d(y + \epsilon \cdot v,\, U + \epsilon \cdot v) = d(y, U)$. Therefore as $\epsilon \to 0$

$$\frac{1}{\epsilon} \left(H_k(U_v(\epsilon), L, f_X) \;-\; H_k(U, L, f_X)\right) \;\to$$

$$\int_{S_k} v' \nabla\, f_X(y)\, \max\{0, \min\{1, \frac{1}{2} + (-1)^k L \cdot d(y,\, U)\}\}\, dy,$$

and this leads to an equation combining these limits to express (a necessary condition for) the optimality of U. When $k = 1$,

$$\frac{dH_1(U_v(\epsilon), L, f_X)}{d\epsilon}\Big|_{\epsilon=0} = \int_{S_k} v' \nabla\, f_X(y)\, \max\{0, \frac{1}{2} - L \cdot d(y,\, U)\}\, dy.$$

Note that the I_1 term is now expressed as $I_1(U_v(\epsilon), L, f_X) = H_1(U_v(\epsilon), L, f_X)+H_2(U_v(\epsilon), L, f_X)$, so

$$\begin{aligned}\frac{dI_1}{d\epsilon}\Big|_{\epsilon=0} &= \frac{dH_1}{d\epsilon} + \frac{dH_2}{d\epsilon} \\ &= \sum_{k=1,2} \int_{S_k} \max\{\frac{1}{2} - Ld(y, U), 0\} v \cdot \nabla f_X(y) dy.\end{aligned}$$

In order to illustrate the method clearly, from now on, we will set $\frac{dI_1}{d\epsilon}\Big|_{\epsilon=0} = 0$ to serve as a first order optimality condition for U, although it should really be $0 \in \frac{\partial I}{\partial \epsilon}\Big|_{\epsilon=0}$. In fact, from the one dimensional analysis in the examples from 2.2, we have

$$\frac{\partial I}{\partial \epsilon}\Big|_{\epsilon=0} = \frac{\partial (I_1 + I_3)}{\partial \epsilon}\Big|_{\epsilon=0} = \begin{cases} (1+\lambda_3)H_1'(0) + (1-\lambda_3)H_2'(0), & E[g] > \frac{1}{2} \\ (1-\lambda_3)H_1'(0) + (1+\lambda_3)H_2'(0), & E[g] < \frac{1}{2} \\ [\min, \max], & E[g] = \frac{1}{2} \end{cases},$$

where $E[g] = \int_{S_1} g dP + \int_{S_2} g dP = \int_{S_1} g dP + (\int_{S_2} 1 dP - \int_{S_2} (1-g) dP) = H_1 + P(S_2) - H_2$. Thus all the necessary information is H_1, H_2 and their derivatives at $\epsilon = 0$. For simplicity of illustration, we will not include the I_3 part, which complicates the discussion in this section.

Assume that $f_X(x) \in C_c^1$, denote $f_1(x) = f_X(x) I_{\{x \in S_1\}}$, so f_1 has compact support S_1, we

can proceed with integration by part [A.2]. Let $f_U(y) := \max\{\frac{1}{2} - Ld(y,U), 0\}$,

$$\begin{aligned}
\frac{dH_1}{d\epsilon}\Big|_{\epsilon=0} &= \int_{S_1} \langle f_U(y)v, \nabla f_X(y)\rangle dy \\
&= \int_{\mathbb{R}^d} \langle f_U(y)v, \nabla f_1(y)\rangle dy \\
&= -\int_{\mathbb{R}^d} \left(\nabla \cdot (f_U(y)v)\right) f_1(y) dy \\
&= -\int_{\mathbb{R}^d} \langle \nabla f_U(y), v\rangle f_1(y) dy \quad (v \text{ is a fixed vector}) \\
&= L\int_{y:\ d(y,U)\le \frac{1}{2L}} \langle \nabla d(y,U), v\rangle f_1(y) dy \\
&= L\int_{y\in S_1:\ d(y,U)\le \frac{1}{2L}} \langle \nabla d(y,U), v\rangle f_X(y) dy.
\end{aligned}$$

Rotation:

Let $U_g = \{g(u), u \in U\}$, where $g(u) = R(u-u_0)+u_0$ $(R^TR = 1, det(R) = 1)$ is a rotation around u_0. Let

$$\begin{aligned}
H(U_g, L, f_X) &:= \int_{S_k} \max\{\frac{1}{2} - Ld(x, g(U)), 0\} f_X(x) dx \\
&\overset{x=g(y)}{=} \int_{S_k} \max\{\frac{1}{2} - Ld(g(y), g(U)), 0\} f_X(g(y)) det(Dg) dy \\
&= \int_{S_k} \max\{\frac{1}{2} - Ld(y, U), 0\} f_X(g(y)) \cdot 1 dy \\
&= \int_{S_k} f_U(y) f_X(g(y)) dy.
\end{aligned}$$

For simplicity, consider rotation in $\mathbb{R}^2$. Denote $R_\theta = \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix}$, we have

$$H_\theta = \int_{S_k} f_U(y) f_X(R_\theta(y-u_0)+u_0)dy,$$
$$H'_\theta = \int_{S_k} f_U(y) \frac{df_X(R_\theta(y-u_0)+u_0)}{d\theta} dy,$$

where

$$\begin{aligned}
\frac{df_X(R_\theta(y-u_0)+u_0)}{d\theta}\Big|_{\theta=0} &= R'_\theta(y-u_0)\cdot\nabla f_X(R_\theta(y-u_0)+u_0)\Big|_{\theta=0} \\
&= \langle \begin{pmatrix} -\sin\theta & \cos\theta \\ -\cos\theta & -\sin\theta \end{pmatrix} (y-u_0), \nabla f_X(R_\theta(y-u_0)+u_0)\rangle\Big|_{\theta=0} \\
&= \langle \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} (y-u_0), \nabla f_X(y)\rangle.
\end{aligned}$$

So far we have obtained two types of balance equations for optimal U:

Balance equation from translation with v:

$$\sum_{k=1,2} \int_{S_k} \langle f_U(y)v, \nabla f_X(y)\rangle dy = 0, \quad \forall v;$$

Balance equation from rotation around u_0 in $\mathbb{R}^2$:

$$\sum_{k=1,2} \int_{S_k} f_U(y) \langle \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} (y-u_0), \nabla f_X(y)\rangle dy = 0, \quad \forall u_0.$$

Reduce the number of equations:

It suffices to look at $v = (1,0),(0,1)$, which spans all possible balance equations from translation. Let $u_0 = (0,0)$, i.e. rotation around the origin, we have $\begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} y = (y_2, -y_1)$. Since for any other rotation $u_0 = (a,b)$, $\begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} (y - u_0) = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} (y_1 - a, y_2 - b) = (y_2 - b, a - y_1) = (y_2, -y_1) + (-b, a)$, so the balance equation from rotation around (a,b) can be reproduced by summing up the balance equation from rotation around $(0,0)$ and translation with $(-b,a)$. Therefore there are essentially only 3 equations above, for $v = (1,0),(0,1)$, and $u_0 = (0,0)$.

Apply integration by part on the 3 equations (we have seen how this is done in the translation case, see remark below for the rotation case) to obtain

$$\sum_{k=1,2} \int_{S_k} \langle (y_2, -y_1), \nabla f_U(y) \rangle f_X(y) dy = 0,$$

$$\sum_{k=1,2} \int_{S_k} \langle (1,0), \nabla f_U(y) \rangle f_X(y) dy = 0,$$

$$\sum_{k=1,2} \int_{S_k} \langle (0,1), \nabla f_U(y) \rangle f_X(y) dy = 0;$$

or equivalently,

$$\sum_{k=1,2} \int_{y \in S_k, d(y,U) \leq \frac{1}{2L}} \langle (y_2, -y_1), \nabla d(y,U) \rangle f_X(y) dy = 0, \quad (1)$$

$$\sum_{k=1,2} \int_{y \in S_k, d(y,U) \leq \frac{1}{2L}} \langle (1,0), \nabla d(y,U) \rangle f_X(y) dy = 0, \quad (2)$$

$$\sum_{k=1,2} \int_{y \in S_k, d(y,U) \leq \frac{1}{2L}} \langle (0,1), \nabla d(y,U) \rangle f_X(y) dy = 0. \quad (3)$$

Remark (Integration by part details for rotation around $(0,0)$).

$$\begin{aligned}
\int_{S_k} \langle f_U(y)(y_2, -y_1), \nabla f_X(y) \rangle dy &= -\int_{S_k} \nabla \cdot \left(f_U(y)(y_2, -y_1) \right) f_X(y) dy \\
&= -\int_{S_k} \left(\frac{df_U(y) y_2}{dy_1} - \frac{df_U(y) y_1}{dy_2} \right) f_X(y) dy \\
&= -\int_{S_k} \langle (y_2, -y_1), \nabla f_U(y) \rangle f_X(y) dy \\
&= -L \int_{y \in S_k, d(y,U) \leq \frac{1}{2L}} \langle (y_2, -y_1), \nabla d(y,U) \rangle f_X(y) dy.
\end{aligned}$$

Remark. These equations illustrate how one can use one dimensional perturbations to find necessary conditions. Given f_X and for fixed L, the three balance equations (1), (2), (3) are sufficient to find an optimal hyperplane U in $\mathbb{R}^2$. However, the scope of the analysis is still limited to find a complete set of equations for a general U surface. We propose that further tools from geometric measure theory (GMT, [37, 52]) should be utilized to answer this question. This future direction will be described in Chapter 5. Results in this chapter are then regarded as preliminary constructions and preparations before applying GMT.

Chapter 3: Computational aspects

Chapter 1 and 2 focus on theoretical aspects of the problem. In this chapter, various computational questions are discussed when implementing the method in practice. This includes the development of a main algorithm, illustration of solutions to the ideal and data problem using the algorithm, and handling questions and models that are not covered by available theory. Some additional conceptual questions are involved.

Organization of the chapter: The main algorithm, Algorithm 1, is based on the idea of alternating minimization. In section 3.4, we explore remedies for the scalability issue using subsampling ideas. In section 3.3, we distinguish different settings for classification/clustering in a population or a data-based problem, which the reader should keep in mind in any simulation experiment. Conceptual issues in tuning parameter selection are discussed in section 3.8, by clarifying the notion of cross-validation and cross-stability. Other elements of the chapter, including Monte Carlo studies, are chosen selectively for mathematical and statistical insights – such as effect of dimension and shape (section 3.6), relation between clustering and classification (section 3.9), and statistical inference in terms of variability of the decision boundary U (section 3.5). A real data example is given in section 3.10.

3.1 Main algorithm

Implementation of the main algorithm is based on Theorem 1.5 from Chapter 1. Theorem 1.5 shows that with data, the variational problem (1.15) is equivalent to a finite dimensional (n dimensional) optimization problem (1.16). The optimal g in the original problem (1.15) is obtained by Lipschitz extension of a function whose values at data points are the optimal a in (1.16), in particular, (1.18) ensures that the range of the extended function is in $[0, 1]$. However, (1.16) is still non-convex, because the function $x \wedge (1 - x)$ appearing in the first term is concave in $[0, 1]$. Fortunately, the other two terms can be turned into linear programming form:

$$\min_{a_1,\cdots,a_n\in[0,1]} \left\{ \frac{1}{n}\sum_{i=1}^{n} \min\{a_i, 1-a_i\} + \lambda_2 \max_{d(x_i,x_j)\neq 0} \frac{a_i - a_j}{d(x_i, x_j)} + \lambda_3 \max\{\bar{a}, 1-\bar{a}\} \right\}$$

is equivalent to

$$\text{minimize } \frac{1}{n}\sum_{i=1}^{n} \min\{a_i, 1-a_i\} + \lambda_2\rho + \lambda_3 r \tag{3.1}$$

$$\text{subject to } \frac{a_i - a_j}{d(x_i, x_j)} \leq \rho, i \neq j$$

$$\bar{a} \leq r, 1 - \bar{a} \leq r$$

$$0 \leq a_i \leq 1, i = 1, 2, ..., n.$$

This first step turns the second and third term in (1.16) into linear programming form.

We further rewrites (3.1) into a mixed-integer program by introducing an auxiliary vector variable z. Let z be a vector variable, where each $z_i, i = 1, \cdots, n$ is binary. The following

program about (a, z) is equivalent to (3.1):

$$\text{minimize } \frac{1}{n}\sum_{i=1}^{n}\left\{I_{[z_i=0]}a_i + I_{[z_i=1]}(1-a_i)\right\} + \lambda_2\rho + \lambda_3 r \tag{3.2}$$

$$\text{subject to } \frac{a_i - a_j}{d(x_i, x_j)} \leq \rho, i \neq j;\ \bar{a} \leq r,\ 1-\bar{a} \leq r;\ 0 \leq a_i \leq 1, i = 1, 2, ..., n;$$

$$z_i \in \{0, 1\}, \quad i = 1, \cdots, n.$$

We propose to use alternating minimization to solve (3.2), leading to Algorithm 1. In the context of clustering and classification, it can be understood as combining Lloyd's algorithm for K-means clustering [34] with Luxburg's linear program for Lipschitz classifier [65], see section 3.11.1.

Alternating minimization

Alternating minimization is a very general principle that has been used to solve a variety of nonconvex optimization problems, especially optimization problems arising from unsupervised learning, see e.g. [25]. Often, these problems are not jointly convex in (x, y) (for two vector variables x and y), but are convex or have closed-form solutions whenever x or y is fixed. Each alternating step is sometimes reduced to solving a known or well-resolved optimization problem.

A generic alternating minimization algorithm minimizes an objective function $f : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ in the following way: start with an initialization (x^0, y^0), then for $t = 0, 1, 2, \cdots$, do

$$x^{t+1} \leftarrow \arg\min_{x \in \mathcal{X}} f(x, y^t), \quad y^{t+1} \leftarrow \arg\min_{y \in \mathcal{Y}} f(x^{t+1}, y), \tag{3.3}$$

until convergence.

Remark. Lloyd's algorithm for K-means clustering is one example of the alternating minimization principle, which, in terms of the K-means formulation (1.9), can be seen as alternating minimization between cluster assignments $C_1, \cdots, C_K$ and cluster centers $c_1, \cdots, c_K$.

Basic property of alternating minimization

The following monotonicity lemma always holds for an alternating minimization algorithm.

Lemma 3.1 (monotonicity of alternating minimization)**.** *Suppose the current solution* (x^t, y^t) *is updated according to (3.3), then*

$$f(x^{t+1}, y^{t+1}) \leq f(x^t, y^t),$$

and equality holds iff (x^t, y^t) *is a bistable point:* $x^t \in \arg\min_{x \in \mathcal{X}} f(x, y^t), y^t \in \arg\min_{y \in \mathcal{Y}} f(x^t, y)$.

Remark. In other words, Lemma 3.1 ensures strict improvement of f value as long as the current point is not bistable. This is a routine result, see e.g. [25]. Monotonicity implies convergence, at least to a local optimal point.

Algorithm 1

The following iterative procedure is proposed to solve (3.2):

1. Randomly pick a starting point z^0, and start multiple times to avoid locally optimal solutions. Or, use another clustering algorithm to initialize, e.g., spectral clustering.

2. Fix z, so that (3.2) becomes a linear program. The solution is the update for a.

3. Fix a, update z. It suffices to

$$\text{minimize} \quad \left\{ I_{[z_i=0]} a_i + I_{[z_i=1]} (1 - a_i) \right\}$$

for each i with a_i fixed. Therefore $z_i = 0$ if $a_i < 1/2$; $z_i = 1$ if $a_i > 1/2$; when $a_i = 1/2$, z_i can be either 0 or 1.

4. Repeat 2&3 until the decrease in the objective value is smaller than some ϵ, or until z and a do not change any more.

The above steps are summarized into Algorithm 1.

Algorithm 1 Alternating Minimization

1. Initialize $z \in \mathbb{R}^n$.
   while not converged, do
2. For fixed z, update $a \in \mathbb{R}^n$ as the solution of the linear program (3.2).
3. For fixed a, update z as $z_i = I_{\{a_i \geq 1/2\}}, i = 1, \cdots, n$.
   end while
4. return a, z. The clustering function g_n is expressed by a as

$$g_n(x) = \frac{1}{2} \min_{i=1,\cdots,n} \{a_i + Ld(x, x_i)\} + \frac{1}{2} \max_{i=1,\cdots,n} \{a_i - Ld(x, x_i)\}, \tag{3.4}$$

   where $L = \max_{d(x_i,x_j) \neq 0} \frac{a_i - a_j}{d(x_i,x_j)}$.

Remark (1). (3.2) can be further formulated into a bilinear program:

$$\begin{aligned}
\text{minimize } & \frac{1}{n}\sum_{i=1}^{n}\{a_i(1-z_i)+(1-a_i)z_i\}+\lambda_2\rho+\lambda_3 r \\
\text{subject to } & \frac{a_i-a_j}{d(x_i,x_j)}\le\rho, i\neq j;\ \bar{a}\le r,\ 1-\bar{a}\le r;\ 0\le a_i\le 1, i=1,2,...,n; \\
& 0\le z_i\le 1, i=1,2,...,n.
\end{aligned}$$

This is because $a_i(1-z_i)+(1-a_i)z_i$ is linear in z_i, thus the optimal z_i's should be either 0 or 1.

Remark (2). Why choose Algorithm 1 instead of other alternatives, such as solvers for mixed-integer quadratic programming or bilinear programming? A reason is that alternating minimization is a general principle that may also generalize to the other formulations of the problem. LP is used here because of the Lipschitz formulation, while different algorithms may be involved if other nonparametric or parametric formulations are considered.

Computational complexity

Two important complexities (that are not to be confused) in Algorithm 1 are:

- z-iteration: number of alternating steps in alternating minimization. This number is usually small in most synthetic and real datasets we have run (Table 3.2), but the general performance deserves more comprehensive empirical study.

- Iterations within LP for fixed z: the complexity of a linear program done via simplex algorithm is often small in practice, despite its exponential complexity in the worst case. This can be partly explained by the theory of linear programming via random inputs [56], which leads to polynomial-time performance in expectation. Empirically, the complexity

of this particular LP is about $O(n^3)$ using a general purpose LP solver (Table 3.3), this is similar to the average case bound [56]. Note that we can only give a rough sense of the complexity because even for a fixed dataset, we have a different LP for every different z.

Remark. A maybe related linear programming problem is the optimal transportation problem, which typically has $O(n^2)$ variables and $O(n)$ constraints. The optimal transportation problem has a long history, so various algorithmic ideas developed therein (see [40]) may be borrowed to better solve our particular problem. We do not pursue this direction in the thesis.

computing distance matrix	$O(n^2)$ for Euclidean distance, see remark
number of z-iterations	see "theoretical issue of alternating minimization"
linear program for each fixed z	$\approx O(n^3)$ *
evaluating function value at given point	$O(n)$ (searching for minimum among n points)
memory complexity	$O(n^2)$ by sparse matrix representation of constraints

Table 3.1: computational complexity in different parts of algorithm; *: the rate $O(n^3)$ is an average case bound for linear programming ([56]), which better agrees with empirical performance (compared to a worst case bound), see Table 3.3.

The linear programming formulation of the Lipschitz term enforces constraints for every pair of points, therefore storing the full constraint matrix (the A matrix in a canonical form $Ax \leq b$) has memory complexity $O(n^3)$ ($O(n^2)$ constraints of $O(n)$ variables), which can be too large to handle for a linear programming software even before the computational bottleneck. Fortunately, the number of nonzero entries of this matrix is only $O(n^2)$, so a sparse matrix routine can resolve this issue, such as the "dense constraint" option in R package "lpSolve".

Tables 3.2 and 3.3 provide a brief empirical overview of the possible effect of shape, degree of separation, noise level and sample size on (1) actual number of iterations and (2) actual running time of each LP solver. The results will be further explained in the next section 3.2.

Distribution		number of z-iterations		
		average	worst case	SD
two piece uniform	$\epsilon = 1$	2.73	12	1.19
	$\epsilon = 0.5$	2.99	9	1.27
	$\epsilon = 0.1$	3.87	12	1.83
	$\epsilon = 0$ (uniform on $[0, 2]$)	4.13	15	2.04
	$\epsilon = 1, \delta = 0.1$	3.45	12	1.31
	$\epsilon = 1, \delta = 0.3$	4.15	14	2.27
	$\epsilon = 1, \delta = 0.5$	4.28	17	2.50
disk and annulus	$\epsilon = 1$	7.73	27	3.96
	$\epsilon = 0.5$	7.36	20	3.50
	$\epsilon = 0.1$	6.8	18	2.79
	$\epsilon = 0$ (uniform on $B_{[0,2]}$)	6.22	15	2.34
	$\epsilon = 1, \delta = 0.1$	8.07	25	4.23
	$\epsilon = 1, \delta = 0.3$	7.59	17	3.25
	$\epsilon = 1, \delta = 0.5$	7.03	19	3.38

Table 3.2: average/worst case/standard deviation of the number of z-iterations over 100 runs for different problems, $n = 200$;
$\epsilon = d(S_1, S_2)$; δ: noise probability, equal to 0 by default.
Two piece uniform data: 1-d uniform distribution on $[0, 1] \cup [1 + \epsilon, 2 + \epsilon]$ when $\delta = 0$; when $\delta > 0$, observe noise (uniform distribution on $[1, 1 + \epsilon]$) with probability δ.
Disk and annulus data: 2-d uniform distribution on $B_{[0,1]} \cup B_{[1+\epsilon,2+\epsilon]}$ when $\delta = 0$; when $\delta > 0$, observe noise (uniform distribution on $B_{[1,1+\epsilon]}$) with probability δ.
For two piece uniform data, the number of z-iterations increases when the clusters become closer or when the noise level is increased. The reverse phenomenon is seen for the disk and annulus data.

Remark in higher dimension

In principle, dimension may affect number of iterations and actual running time of LP as well. However, numerical experience suggests that the additional cost in these two respects is not as important as the growth of cost for increasing sample size. These observations are further

sample size	running time of LP (sec)		
	average	worst case	SD
50	0.02	0.05	0.01
100	0.13	2.12	0.14
150	0.90	39.39	3.23
200	4.08	242.72	19.09
250	21.46	1311	99.34

Table 3.3: average/worst case/standard deviation of running time of an LP solver for different sample sizes in 100 Monte Carlo samples, under random initialization of Algorithm 1. Since a linear program is run for each z-iteration, the average is taken over 100×"average number of z-iteration" (as displayed in Table 3.2 for $n = 200$) linear programming runs. Data: $Unif([0, 1] \cup [2, 3])$.

distance	average number of iterations	average running time of LP (sec)
Euclidean distance	5.5	2.4
Wasserstein distance	4.2	2.5

Table 3.4: Average number of iterations and average running time of LP for a random unlabeled sample of size 200 from digit 0 and digit 1 in the MNIST dataset. The input dimension here is 256, but the running time does not increase much from the 1-d and 2-d datasets in Table 3.2 and 3.3.

supported by real data examples, see Table 3.4. Therefore, for Algorithm 1, higher dimension may be more of theoretical concern rather than computational (the additional computational cost is mainly in computing a distance matrix, which for Euclidean distance is $O(n^2 d)$). Whether the algorithm finds the global optimum (Table 3.5) is a notable issue because local optima can be more abundant in higher dimension.

Though it is not possible to exhaust all possible problem characteristics, some selected examples are presented in section 3.6, where three problem characteristics are considered: dimension, degree of separation and shape of the clusters.

Theoretical issues of alternating minimization

The monotonicity Lemma 3.1 is not able to explain the empirically small number of z-iterations (see Table 3.2) of Algorithm 1. It also does not address the question that to what extent a global optimum can be found by Algorithm 1 on top of a local optimum. For example, table 3.5 shows that simple random initialization can lead to a large number of local optima, which may prevent the algorithm from finding the global optimum. This will be further discussed in the future direction section in Chapter 5.

Initialization

By default we will run the Algorithm 1 under 10 random initializations, and the final solution is one that yields the smallest objective value in the end. Alternatively, the output of another clustering algorithm, such as spectral clustering, can be used initially. Such choices often lead to more rapid convergence of Algorithm 1. This practice can be also seen as a postprocessing step upon other clustering algorithms – since one shortcoming of many machine learning algorithms is that the output is not smooth: either clustering is done only on the data points or the memberships are categorical.

Initialization within subsampling

Later in Algorithm 2, one method uses output from the previous subsample as the initial for running algorithm on the next subsample. This is helpful, as can be seen in Table 3.9, the number of iterations is substantially decreased.

3.2 Illustration of the method

In this section we present a pure illustration of the method on simple synthetic datasets. More complicated examples and real datasets will be considered in later sections.

Two piece uniform

The data is generated from a uniform distribution on $[0,1] \cup [2,3]$, sample size $n = 200$. Let D be the distance matrix computed from data. Implementation of Algorithm 1 is coded in R function "Lclust". The only necessary input is the distance matrix. Default tuning parameters are $\lambda_2 = 0.1 * mean(D), \lambda_3 = 0.5$, where $mean(D)$ denotes sample average of entries of the distance matrix. This choice ensures invariance of clustering under uniform scaling of the variables.

output = Lclust(D=D, one_d = TRUE)

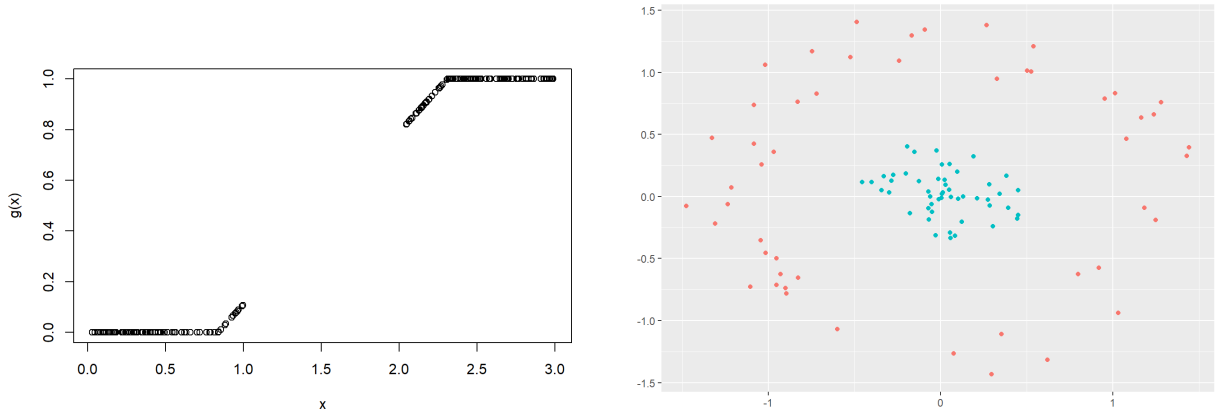


Figure 3.1: Left: fitted clustering function values at data points; data: $Unif[0,1] \cup [2,3]$, $n = 200$. Right: clustering membership (z values) at data points by color; data: $Unif(B_{[0,1]} \cup B_{[2,3]})$, $n = 200$

The function "Lclust" will return a list of outputs. The main output of the algorithm contains:

a: length n vector, function values of g at data points;

z: length n binary vector, cluster membership;

L: Lipschitz constant of the fitted solution,

where "a" and "L" determines the clustering function g_n by (3.4), based on which the prediction at any future point can be computed. Other output includes the estimated proportion of the clusters, number of iterations, optimal value of the objective function, and a classification error which is value of the first term of the objective function under the optimal g. Below is the output list from the two piece uniform data:

a: 0 0 1 0.972 0 ...

z: 0 0 1 1 0 1 1 1 1 0 ...

L: 0.679

proportion: 0.5

num_of_iterations: 2

objective: 0.35

classification error: 0.0188

Remark. If the original data is supplied rather than a distance matrix, then the function will first compute a Euclidean distance matrix by default, and the output list will contain the original data as well. This is useful later to develop subsampling functionality on top of the basic function.

In function "predict.Lclust", "a" and "L" will be extracted from an Lclust object and they are used to construct the fitted g function. For example, we can do prediction at a new data point:

predict.Lclust(output, newdata = 1.5)

0.6943509

We can also do prediction for a data matrix (n by p):

newdata = matrix(seq(0, 3, 0.1), ncol = 1)

predict.Lclust(output, newdata = newdata)

The choice of uniform distribution is not significant, see Figure 3.10 for some empirical results on other choice of distributions.

Disk and annulus

The data is generated from a uniform distribution on the union of disk $B_{1/2}(0)$ and annulus $B_{[1,3/2]}$, where $B_{[a,b]} = \overline{B_b(0) \backslash B_a(0)}$ denotes an annulus, sample size $n = 200$.

output = Lclust(x=x, two_d = TRUE)

L: 1.176801

proportion: 0.5

num_of_iterations: 7

objective: 0.4036826

classification error: 0.02245751

Table 3.5 lists several local optima for this problem and their relative frequency. For $d = 3$ (ball and annulus), random initialization fails to find the global optimum in 100 runs among a large number of local optima (so a table is not displayed), possibly due to the annulus structure is less clear in higher dimension for a small sample. This can be overcome by using the output of another clustering algorithm, such as spectral clustering to initialize Algorithm 1.

objective value	percentage
0.47193	8%
0.40687*	20%
0.45013	15%
0.5**	14%
0.47362	14%
other	29%

Table 3.5: Objective values and percentage of local optima by running Algorithm 1 using random initialization for 100 times. Data: disk and annulus in 2d, $n = 200$. *: global optimum; **: flat solution; other: other local optima that appear less than 3 times in 100 runs. If spectral clustering is used initially instead, then the global optimum can always be found for this example.

3.3 Three simulation settings

In this section we distinguish between three simulation settings described in Table 3.6: a population problem, a classification problem and a clustering problem. They are equally important scenarios, serving different purposes in the simulations, see Table 3.7. The three settings are dealt with by related but slightly different algorithms, contrasted in Table 3.8.

Table 3.6: Three settings

1	population clustering/classification	know P_X, derive/compute ideal clustering function
2	supervised-classification	$\{X_i\}_{i=1}^n$ i.i.d drawn from P_X, Y_i observed.
3	unsupervised-clustering	$\{X_i\}_{i=1}^n$ i.i.d drawn from P_X, Y_i not observed.

The three settings are ubiquitous in clustering and classification. One can consider the three settings simultaneously for other ideal population models (such as the Gaussian mixture model) along with methods and algorithms developed under that model.

Setting 1 (population clustering/classification): Suppose P_X is known, and is generated from C1. The goal is to find either a mathematical or numerical solution to the variational prob-

Table 3.7: purpose served for three settings

setting 1	understand the method – what is the clustering function (as the optimal solution of the variational problem) when the distribution P_X is known
setting 2& 3	show that there is a unifying framework for clustering and classification with Lipschitz decision functions; also as a way to assess performance of setting 3 under ideal conditions (illustrated in section 3.9)
setting 3 + tuning parameter selection	relevant to real data applications of clustering

Table 3.8: methods for three settings

setting 1	sample directly from the model, then solve by LP (3.5).
setting 2	LP, similar to [65].
setting 3	alternating LP (Algorithm 1). Works well (in a few iterations) empirically when clusters are well-separated.

lem (1.5). In this chapter we desire a numerical solution. First, sample points $x_1, \cdots, x_n$ from the model. For these sampled points, we know which are from S_1, and which are from S_2. Then we can use theoretical results developed in Chapter 2 to make reduction. By Theorem 2.4, under suitable choice of λ's, we can assume w.l.o.g that the optimal solution g satisfies $g|_{S_1} < 1/2, g|_{S_2} > 1/2$. The problem is then reduced to a convex problem, as explained in section 2.4, the only difference being that here we are dealing with a discrete version. In fact, it can be written as the linear program:

$$\text{minimize } \frac{1}{n}\left[\sum_{x_i\in S_1} a_i + \sum_{x_i\in S_2}(1-a_i)\right] + \lambda_2\rho + \lambda_3 r \tag{3.5}$$

$$\text{subject to } \frac{a_i - a_j}{d(x_i,x_j)} \le \rho, i \ne j;\ \bar{a} \le r, 1-\bar{a} \le r;$$

$$0 \le a_i \le 1/2, x_i \in S_1;\ 1/2 \le a_i \le 1, x_i \in S_2.$$

Remark. Suppose the optimal solution can be proved to lie in a certain subset of the feasible region, then we can reduce the search space. In this case, we have convexity in the reduced space.

Setting 2 (Classification): A natural classification algorithm arises as a byproduct from Algorithm 1. In Algorithm 1, the vector z can be regarded as an indicator of cluster membership. In classification, where we observe (X_i, Y_i) pairs, we can replace the z_i's in step (2) by Y_i's. This gives the optimization problem in the classification context:

$$\text{minimize } \frac{1}{n}\sum_{i=1}^{n}\left\{I_{[Y_i=0]}a_i + I_{[Y_i=1]}(1-a_i)\right\} + \lambda_2\rho + \lambda_3 r \tag{3.6}$$

$$\text{subject to } \frac{a_i - a_j}{d(x_i, x_j)} \leq \rho, d(x_i, x_j) \neq 0;\ \bar{a} \leq r, 1-\bar{a} \leq r;\ 0 \leq a_i \leq 1, i = 1, 2, ..., n.$$

This is a linear program, equivalent to running step 2 of Algorithm 1 once with $z = Y$. The output classification function is still given by the Lipschitz extension (3.4). In other words, (3.6) is solved by running Algorithm 1 with initial labels Y and with exactly one iteration. This allows us to write the optimization program for clustering and classification together.

Remark. In classification, we can set $\lambda_3 = 0$ (constraint on proportion is not necessary), in which case the method will have only one tuning parameter λ_2, and can be seen as a variant of Luxburg's linear program (3.13) for Lipschitz classifier, with a different loss function and a different range for the classifier.

Setting 3 (Clustering): This is implemented by Algorithm 1, where the output is a Lipschitz clustering function. Contrasted with classification (3.6), it can be viewed as searching for

labelings that minimize the (empirical) Lipschitz regularized risk.

3.4 Scaling up the algorithm

For very large n, the original program becomes infeasible because it involves solving a large linear program, with $O(n)$ variables and $O(n^2)$ constraints. See also Table 3.1 and several remarks that follows. In order to make the algorithm scalable, we explore several ideas using subsampling ([42]). Generally speaking, each time the algorithm is run on a random subsample of the original data with sample size m much smaller than n. The process is carried on B subsamples either repeatedly (so is purely parallelizable) or iteratively (for better subsamples or better estimate), then a final solution is proposed by aggregating the B solutions.

Other than dealing with computational bottleneck, subsampling can also answer important questions about our ideal and data problems. For the ideal problem, it allows us to get a better numerical solution of the variational problem by sampling more points from a model. For the data problem, it comes with inferential advantages. In section 3.5, we use it to construct a confidence band for the decision boundary U. Consistency of such a subsampling scheme is addressed in section 3.11.4.

Subsampling and aggregation

In classical empirical bootstrap, each bootstrap sample of size n is drawn from the empirical distribution P_n, which amounts to sampling n points with replacement from the original data $X_1, \cdots, X_n$. In our case, m-out-of-n bootstrap is used instead. This is necessary because of computational constraints: for large n, running the original program on full data, more precisely,

the large linear program (with n variables and $O(n^2)$ constraints) in Algorithm 1, is infeasible. Using subsampling, we need only to solve a linear program with m variables each time, where m can be a data-adaptive choice such as $\sqrt{n}$. Such choice has much supportive theory in the subsampling literature [41] ($m \to \infty, m/n \to 0$). In practice, m can be set to a small number (e.g., $m = 100$) for which the program has efficient and reliable computational performance on the given machine.

Alignment in clustering

For a clustering problem, the aligning step is necessary when aggregating, because by symmetry (and here when $K = 2$), if g is a solution, then $1 - g$ is also a solution, and an algorithmic procedure alone cannot tell the two solutions apart. Therefore, a simple average of B solutions $g_1, \cdots, g_B$ would be meaningless if they are not "aligned". This is detailed in step 2 of Algorithm 2.

Algorithm 2

In short, the subsampling version of the algorithm runs Algorithm 1 on B subsamples of size m, then aligns the B solutions to get an aggregated solution ("aggregating" can be implemented in ways other than simple averaging, see, e.g., step 3 of Algorithm 2).

Algorithm 2 subsample + aggregate

1. Take B random subsamples of X of size m, denoted by $X^{(1)}, \cdots, X^{(B)}$.

   Method = parallel:

   Run Algorithm 1 "in parallel" on the B subsamples, $g^{(1)}, \cdots, g^{(B)}$ denote the B solutions.

   Method = sequential:

   For $i = 1, \cdots, B-1$, use the prediction of ith clustering function on $(i+1)$th subsample, $g^{(i)}(X^{(i+1)})$, as the initial z value when running Algorithm 1 on the $(i+1)$th subsample.

2. Let $v = (v_1, \cdots, v_B)$ be an alignment indicator vector, where $v_1 = 1$, and for $i = 2, \cdots, B$, $v_i = I\{d(g^{(1)}, g^{(i)}) < d(g^{(1)}, 1 - g^{(i)})\}$ for some distance metric between g's, such as the L_1 distance (3.8) on the merged sample $X^{(1)} \cup X^{(i)}$.

3. Let $\tilde{g}^{(i)}(x) = \{v_i g^{(i)}(x) + (1 - v_i)(1 - g^{(i)}(x))\}$ denote the ith aligned solution. The aggregated clustering function is

   Method = average:

$$g(x) = \frac{1}{B} \sum_{i=1}^{B} \tilde{g}^{(i)}(x);$$

   Method = pointwise median:

$$g(x) = \text{sample median}\{\tilde{g}^{(1)}(x), \cdots, \tilde{g}^{(B)}(x)\}.$$

Choice of m and B

A good choice of m and B should balance statistical and computational performance. Suppose the LP run on original sample with size n has complexity $O(n^3)$, then the complexity of step 1 (running Algorithm 1 on all B subsamples of size m) is $O(m^3 B)\times$ average number of z iterations, the complexity of step 2 (alignment) is $O(mB)$, evaluating the function value at future point takes $O(mB)$ steps. Memory complexity (mainly to store the linear programming constraint matrix) is $O(m^2)$. Suppose $m = \sqrt{n}, B = \sqrt{n}$, then the total computational cost is reduced to $O(n^2)$. On the other hand, a reasonably large m and B ensures good statistical performance.

number of z-iterations	1	2	3	4	5	7
frequency	0	14	65	14	5	2

method = parallel

1	2	3	4	5	6
18	60	19	2	0	1

method = sequential

Table 3.9: comparing number of iterations in a single run of step 1 of Algorithm 2 using the parallel and the sequential approach, B = 100. The sequential approach has fewer number of iterations, many converge in a single iteration, illustrating the usefulness of initialization from the output of earlier subsamples.

Repeated observations in subsample

A typical random subsampling procedure involves sampling with replacement, therefore there is a positive probability that some points in the subsample are repeated observations. For certain clustering methods such as hierarchical clustering, this will change the combinatorics and cause potential issues. The situation is much better here, except for a technical modification of the linear program: in (3.1) or (3.2), the constraints $\frac{a_i - a_j}{d(x_i, x_j)} \leq \rho$ should be changed to $a_i - a_j \leq \rho \cdot d(x_i, x_j)$ (so the pair of constraints for i, j enforces $a_i = a_j$ whenever $d(x_i, x_j) = 0$). Thus, the repeated observations play a part in the first and third term in (3.1) and (3.2), but does not contribute to the Lipschitz term.

Retain necessary condition of solution

The averaged solution in the aggregation stage of Algorithm 2 violates optimality condition (Theorem 1.1), see left plot in Figure 3.2. This is expected: when proving uniqueness of solution (section 2.5), we have already used the fact that the average of two different solutions (both satisfy N.C.1) does not satisfy the optimality condition anymore (Figure 2.4 illustrates this observation in dimension one). Besides loss of theoretical property, such issue also brings up practical questions: if a solution solved by the subsampling version of the algorithm does not behave the same way

as the original version, then it hinders a unified interpretation of the solution under the method.

For this reason, pointwise median of the functions (method = pointwise median in Algorithm 2) is proposed in place of the average. The right plot in Figure 3.2 shows that N.C.1 is roughly satisfied.

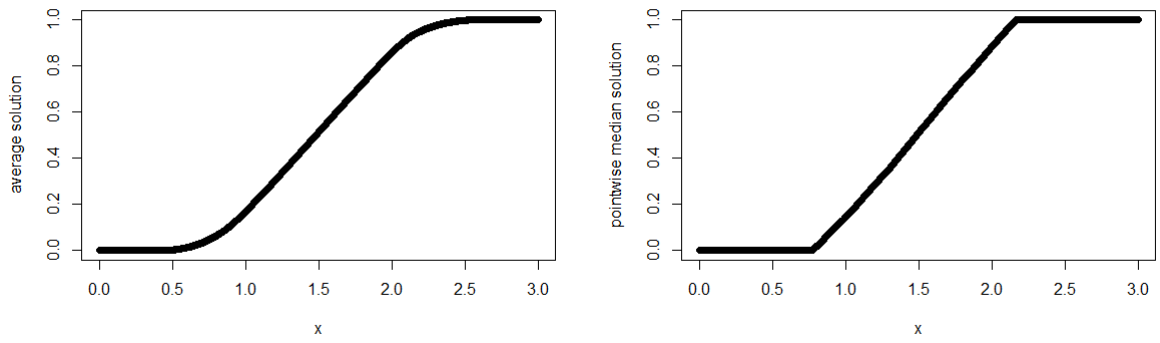


Figure 3.2: Left: method = average; Right: method = pointwise median. P_X: two piece uniform on $[0, 1] \cup [2, 3], n = 1000, m = 50, B = 100$.

Subsample discriminating points

For clustering and classification, an alternative strategy to scale up the algorithm is to run the algorithm on a set of discriminating points rather than on full data, under the assumption that a small amount of discriminating points near the decision boundary determines the clustering result, and deleting points far-away from the boundary does not change problem structure much.

Such a procedure can proceed as follows: based on an initial solution $g^{(0)}$ or any later solution $g^{(i)}$, estimate an uncertainty level of each point in the current clustering, then resample from the original data with sampling probability proportional to the uncertainty level. By repeating this resampling procedure for a number of times, more important points close to the decision boundary are gathered in the subsample.

For example, suppose we use $\min\{g(x), 1 - g(x)\}$ as an uncertainty level, then for many points whose g values are 0 or 1 (optimality conditions suggest that such points are abundant) indicating that they are perfectly-clustered, they will have 0 probability to appear in the new sample. The size of each new sample can be set to be a constant for which computational cost is reasonable. No aggregation stage is needed here. See Figure 3.3 for an illustration of the idea.

Figure 3.3: Illustration of subsampling discriminating points. Left: two-spheres data, $n = 2000$. Middle: discriminating points estimated from a preliminary run on a random subsample of size 200. Right: resample 200 points with sampling probability proportional to the estimated uncertainty level $p(x_i) \propto \min\{g_n(x_i), 1 - g_n(x_i)\}$. The algorithm is run on this new subsample in the next iteration.

3.5 Confidence band for decision boundary U

A natural mathematical object of interest in this method is the decision boundary U, whose importance can be seen from Theorem 2.1: the optimal solution is almost uniquely determined by U and the Lipschitz constant (up to a sign at each point). Therefore it will be nice to quantify the variability of the data-based estimate of U, which gives much information about the variability of the solution. More precisely, the goal is to find a set C_α based on data such that $P(C_\alpha \supset U) = 1 - \alpha$, for some confidence level α. The concept of confidence band differs from a pointwise confidence interval for g^* in the sense that it asks for uniform coverage of the target set U, rather

than pointwise coverage of the function value at a particular point.

Since U is the level set of the optimal solution g at $1/2$, the problem of finding a confidence band for the decision boundary is related to the general problem of constructing a confidence region for the level set of a function estimate [38]. Such a method was originally developed in pursuit of a confidence region for the level set of a kernel density estimator. It was later used in the density-estimate-based clustering literature, where clusters are formed by connected components of the level set of density estimates.

The method

We elaborate the general methodology proposed in [38] in a simplified manner, at the cost of losing some mathematical rigor. Then we show how this leads to a practical confidence band for U in our problem. The remark at the end explains what are the additional technical details one needs to take care to treat this more rigorously. The method is based on bootstrap and has the advantage that it does not require particular geometric assumptions on U.

Suppose U is the set of zeros of a function h, i.e., $U := \{x : h(x) = 0\}$ (it may be helpful to mentally think of U as a $(d-1)$-manifold in the preceding discussion). Based on data $X_1, \cdots, X_n$ and an estimate $\hat{h}_n$ of h, the goal is to construct a confidence band for U. In our case we require that both h and $\hat{h}_n$ are Lipschitz. We can proceed by relating this question to the statistic $Z_n = \sup_{x \in U} |\hat{h}_n(x) - h(x)|$, whose distribution may be hard to derive explicitly but can be approximated using (Efron's nonparametric) bootstrap: let $X_1^*, \cdots X_n^*$ be a bootstrap sample drawn with replacement from $\{X_1, \cdots, X_n\}$, and $\hat{h}_n^*(\cdot) := \hat{h}_n(\cdot\,; X_1^*, \cdots X_n^*)$ is the estimate of h using the bootstrap sample. Then $Z_n^* := \sup_{x:\ \hat{h}_n(x)=0} |\hat{h}_n^*(x) - \hat{h}_n(x)|$ is an estimate of Z_n. Let F

denote CDF of Z_n, and F^* the corresponding CDF of Z_n^*.

The set $C_\alpha = \{x : |\hat{h}_n(x)| \leq F^{-1}(1-\alpha)\}$ is a $(1-\alpha)$ confidence band for U, from the following derivation

$$
\begin{aligned}
& P(Z_n \leq F^{-1}(1-\alpha)) = 1-\alpha \\
\iff & P(\sup_{x\in U} |\hat{h}_n(x) - h(x)| \leq F^{-1}(1-\alpha)) = 1-\alpha \\
\overset{(1)}{\iff} & P(\{x : |\hat{h}_n(x)| \leq F^{-1}(1-\alpha)\} \supset \{x : h(x) = 0\}) = 1-\alpha \\
\iff & P(C_\alpha \supset U) = 1-\alpha,
\end{aligned}
$$

where (1) is because the two events are equivalent: $\sup_{x\in U} |\hat{h}_n(x) - h(x)| \leq F^{-1}(1-\alpha)$ implies for any $x \in U = \{x : h(x) = 0\}$, $|\hat{h}_n(x) - 0| \leq F^{-1}(1-\alpha)$, so $x \in C_\alpha$, therefore $C_\alpha \supset U$; conversely, suppose $C_\alpha \supset U$, by definition of C_α, for any $x \in U \subset C_\alpha$, $|\hat{h}_n(x)| \leq F^{-1}(1-\alpha)$, this implies $\sup_{x\in U} |\hat{h}_n(x) - h(x)| \leq F^{-1}(1-\alpha)$.

Replacing F with F^* in C_α gives a practical approximate confidence band. The justification of the bootstrap approximation here depends on the smoothness of the random function $\hat{h}_n$ and appropriate (functional) central limit theorem holding for $\hat{h}_n - h$.

Now back to our problem, since U is the $\frac{1}{2}$-level set of g_0 (where g_0 is the solution of a variational problem). Write $h = g_0 - \frac{1}{2}$ (so that $U = \{h = 0\}$) and use the confidence set C_α described above. Let $\hat{g}_n$ denote the data-based estimate of g_0 and F be the CDF of $\sup_{x\in U} |\hat{g}_n(x) - g_0(x)|$, we arrive at the following $(1-\alpha)$ confidence band for U

$$C_\alpha = \{x : |\hat{g}_n(x) - \frac{1}{2}| \leq F^{-1}(1-\alpha)\}.$$

By optimality condition of the variational problem (assume it holds also for the data-based $\hat{g}_n$):

$$\hat{g}_n(x) = \begin{cases} \max\{\frac{1}{2} - \hat{L}_n d(x, \hat{U}_n), 0\}, & x \in \hat{U}_{n,1} \\ \frac{1}{2}, & x \in \hat{U}_n \\ \min\{\frac{1}{2} + \hat{L}_n d(x, \hat{U}_n), 1\}, & x \in \hat{U}_{n,2} \end{cases},$$

where $\hat{U}_n = \{x : \hat{g}_n(x) = 1/2\}, \hat{U}_{n,1} = \{x : \hat{g}_n(x) < 1/2\}, \hat{U}_{n,2} = \{x : \hat{g}_n(x) > 1/2\}; \hat{L}_n = L(\hat{g}_n)$, plugging into C_α:

$$C_\alpha = \{x : d(x, \hat{U}_n) \leq \frac{F^{-1}(1-\alpha)}{\hat{L}_n}\}. \tag{3.7}$$

The derived confidence band thus has the form

$$\{x : d(x, \hat{U}_n) \leq c\}.$$

The confidence band (3.7) can be made practical by approximating F by a suitable bootstrapped version

$$\sup_{x \in \tilde{U}_n} |\hat{g}_n^*(x) - \hat{g}_n(x)|,$$

where $\tilde{U}_n$ is some approximating set of $\hat{U}_n$ (such as a finite set of points whose $\hat{g}_n$ values are close to $1/2$). This is possible because all the g functions involved are Lipschitz continuous.

Remark (technical details). A more rigorous treatment of this subject would at least involve relaxing the level sets in Z_n and Z_n^* over which the supremum is taken to a small tube around them with an adaptive choice of bandwidth, e.g., consider $\{x : |g_n(x) - 1/2| \leq a_n\}$ instead

of $U_n = \{x : g_n(x) = 1/2\}$, where $a_n \xrightarrow{n} 0$, followed by a careful analysis of the coverage probability under suitable convergence rate of a_n, plus the approximation error of the bootstrap approximation Z_n^* to Z_n, see Lemma 2.1 in [38]. Theorem 3.1 in [38] did a case analysis on level sets of kernel density estimator using bootstrap approximations.

Remark (compactness). The confidence region for the entire surface U can be very large, because there is a lot of variability in the tails of U where data are sparse (for noiseless models, the tails are not even uniquely defined). Therefore in practice, we only consider confidence band for the restriction of U within certain compact region D. It is then expected that as we narrow down the region of interest, the resulting confidence band for $U \cap D$ will become narrower.

Remark (implementation). It requires a good amount of data to have a uniform confidence band that is not too wide. On the other hand, computational problems can arise when sample size is large, as explained in section 3.4. In this case, g_n and each bootstrapped estimate g_b^* are computed from algorithm 2 using B subsamples of size m.

Remark (nominal coverage probability). It will be desirable to do a numerical experiment or simulation to verify that the confidence band has the approximate nominal coverage probabilities. However, such an experiment requires many (100 or more) such confidence band implementations. At the moment, computing one confidence band such as in Figure 3.4 already takes about 3 hours.

Example

Figure 3.4 illustrates the confidence band method through a Gaussian mixture example. However, these apparently plausible numerical results given here go beyond what can so far be

proven rigorously, or verified computationally by nominal coverage probability.

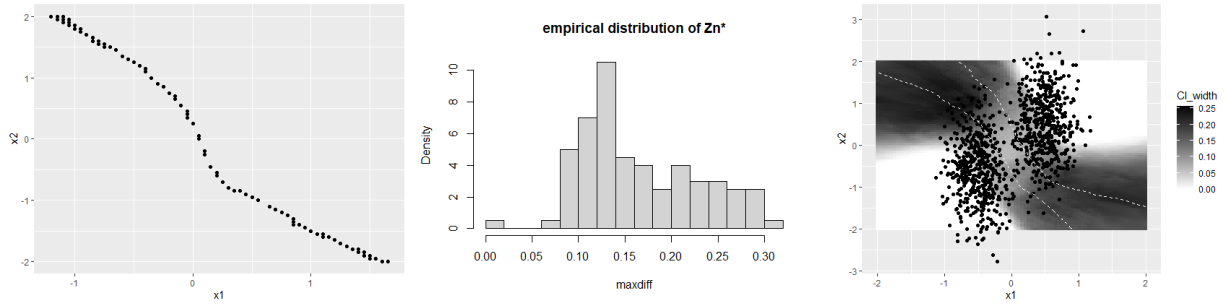


Figure 3.4: Left: $\tilde{U}_n = \{x \in \{x_1, \cdots, x_n\} : |g_n(x) - 1/2| \leq 0.01\}$, as an approximation set to $U_n = \{x : g_n(x) = 1/2\}$. Middle: histogram for the empirical distribution of the statistic $Z_n^* = \sup_{x \in \tilde{U}_n} |g_n(x) - g_b^*(x)|$. Right: a 90% confidence band for U based on (3.7), within the region $[-2, 2] \times [-2, 2]$, supplemented by widths of pointwise confidence intervals for g^* (greyscale values). Data: mixture of two gaussians with mean vector $\mu_1 = -\mu_2 = (0.5, 0.5)$ and common covariance matrix $\begin{pmatrix} 0.05 & 0 \\ 0 & 0.5 \end{pmatrix}$, $n = 1000, m = 50, B = 100, B' = 100$. Here g_n and each g_b^* ($b = 1, \cdots, B'$) are computed from algorithm 2 using B subsamples of size m, and by the pointwise median method.
Implications for clustering (last figure): dark grey regions on the tail within confidence band: pointwise CIs are wide and clustering is confused; dark grey regions on the tail outside confidence band: pointwise CIs are wide but clustering is clear; white regions at bottomleft and topright: at least 90% of the empirical bootstrap solutions has value exactly 0 (or 1) at these points (so that CI reduces to a single point, at either 0 or 1).

3.6 Monte Carlo study 1: effect of dimension, shape and degree of separation

In this section we conduct a Monte Carlo study on the effect of dimension, shape and degree of separation on computation (number of iterations, run time of linear program) and quality of solution (smoothness, correctness). Table 3.10 lays out the empirical behavior for selected models. The interpretation of the columns should take account of how the dimension grows. The disk annulus data are generated by uniform distribution on $B_{[0,3]} \backslash B_{[1,2]}$ (notation defined in section 3.2), i.e., with fixed radius and growing dimension. The d-dimensional two-spheres data are generated in two ways (both uniform distributions): (1) two d-dimensional unit spheres centered

at 1_d and -1_d; (2) distance between the two d-dimensional unit spheres is fixed to be $\sqrt{2}$. It is worth mentioning that these spherical shell clusters are nonconvex. In the first two-spheres case, it is clear that clustering is easier in higher dimension, as can be seen from smaller Lipschitz constant and less classification error, while the same phenomenon is not quite clear for the second case and for the disk and annulus data. The differences in computation time for higher dimensions is not large.

In the last two columns, the two quantities $I_{n,1}$ and R_n (see definition in (2.6) and (2.7)) are both related to classification/clustering errors, but should be carefully distinguished, see section 2.2.3. They are the same (which is in several rows of Table 3.10) when bipartition (C3) holds for g_n. In general, $I_{n,1}$ is smaller than R_n.

Notions of margin

There are two types of notions of margin: population-based and data-based ([59]). A population-based notion of margin is a function of the joint distribution of (X, Y) which quantifies the degree of separation between clusters. Natural candidates are distance between compact sets for well-separated compact clusters, and Mahalanobis distance for Gaussian mixture model.

Historically, the data-based notion of margin came from the support vector machine literature – geometrically it is the distance (or twice the distance) from any "support vector" to the optimal hyperplane, and from functional analysis point of view is the norm of the linear decision functional [50]. This notion is generalized to the metric space setting by [65], where the corresponding quantity is the reciprocal of the Lipschitz constant. We refer to the many discussions in [65] about this generalization, especially the functional analysis point of view.

Distribution	dimension	z-iterations	run time	L	$I_{n,1}(g_n)$	$R_n(g_n)$
Disk and annulus	2	5	3.22	0.57	0.03	0.03
	3	6	3.42	0.58	**0.03**	**0.05**
	5	3	2.74	0.55	**0.07**	**0.15**
	10	5	2.12	0.49	**0.02**	**0.22**
Two spheres (fixed distance along each coordinate)	2	3	21.56	0.57	0.06	0.06
	3	5	2.17	0.41	0.02	0.02
	5	4	2.53	0.28	0.009	0.009
	10	3	2.67	0.18	0.003	0.003
	20	2	2.45	0.12	0.002	0.002
Two spheres (fixed distance between centers)	2	3	21.56	0.57	0.06	0.06
	3	3	3.69	0.51	0.04	0.04
	5	6	3.63	0.47	0.02	0.02
	10	1	3.01	1.12	**0.04**	**0.46**
	20	1	2.98	0.96	**0.02**	**0.41**

Table 3.10: dimension effect in some ideal models. Sample size $n = 200$, solution is computed under 10 random initializations. The iteration and run time columns show relatively little effect on computing time in higher dimension (even under random initialization). The estimated Lipschitz constant indicates how smoothness of the clustering change with dimension.
For an estimated clustering function g_n, the second last column is the "classification error" given by value of the first term in the objective function $I_{n,1}(g_n)$; the last column is the empirical clustering risk of g_n evaluated from true labeling. The latter (R_n) can only be available in a Monte carlo study when the truth is known for validation, and is not accessible to real data. $I_{n,1}$ and R_n are equal when g_n satisfies bipartite condition C3, which can be confirmed in a Monte Carlo study. The bold-faced numbers indicate cases when $I_{n,1}$ underestimates the true clustering error, sometimes severely due to reasons such as convergence to local optimum.

There is an asymptotic agreement of the two notions for well-separated clusters for our case: convergence of $1/L_n$ (data-based margin) to $d(S_1, S_2)$ (population margin) by Corollary 2.5, when tuning parameters are selected appropriately.

Table 3.11 illustrates that $1/L_n$ can be a data-based indicator of "degree of separation" only weakly depending on dimension and shape. But it is well-defined even when clusters are

not well-separated and in which case "population margin" may not have a clear definition.

Distribution	dimension	$1/L_n$	$d(S_1, S_2)$
Disk and annulus $d(S_1, S_2) = \sqrt{d}$	2	1.75	1.41
	3	1.72	1.73
	5	1.82	2.24
	10	2.04	3.16
Two separated spheres $d(S_1, S_2) = 2\sqrt{d} - 2$	2	1.75	0.83
	3	2.44	1.46
	5	3.57	2.47
	10	5.56	4.32
	20	8.33	6.94

Table 3.11: $1/L_n$ as a notion of margin. This table is computed from the examples in Table 3.10, with $n = 200$ for data randomly sampled from each P_X model.

Table 3.11 is a bit crude in the sense that L_n can depend on other quantities. Table 3.16 in the appendix shows that, in particular, it can depend on the distribution within the clusters.

3.7 Monte Carlo study 2: 6-component Gaussians, visualization and diagnostics

We provide a dataset distributed according to a mixture of 6 bivariate-Gaussian components (top figure in Figure 3.5, $n = 200$) as an additional example where nonlinear shape is present. The 6 Gaussian components are created to resemble "two crescents" (the left 3 Gaussians as one cluster and the right 3 Gaussians as another). Figure 3.5 presents several locally optimal solutions discovered by Algorithm 1 based on the dataset. This example is further used to illustrate several visualization and diagnostic tools.

We provide several tools for visualization and diagnostics after the solution is fitted. These

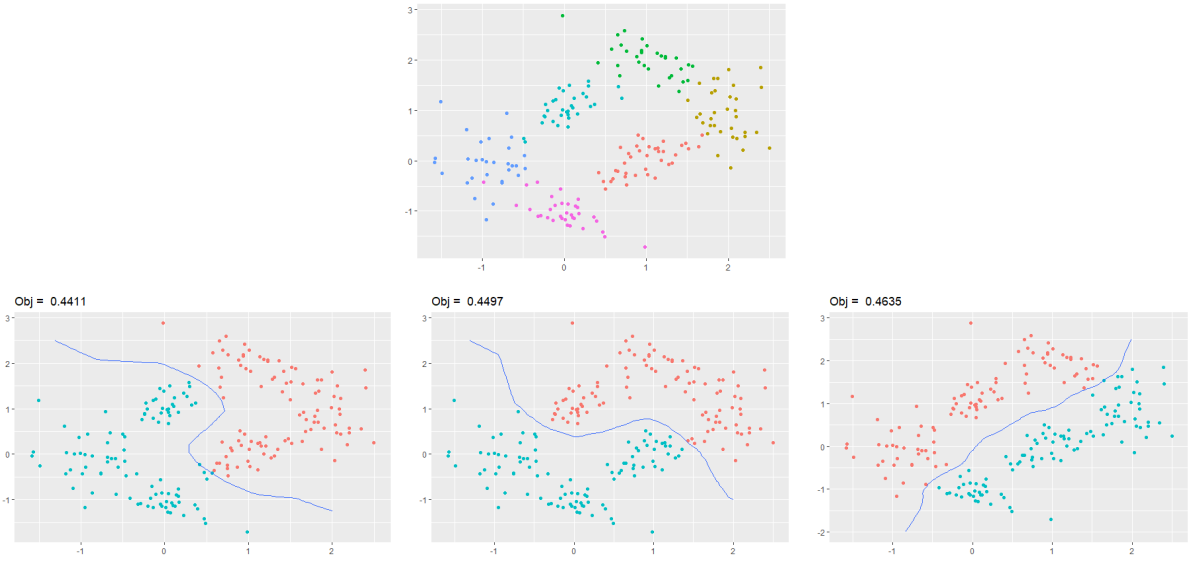


Figure 3.5: The 6-component mixed Gaussian data (top) and several locally optimal solutions found by Algorithm 1 (from different random starting points). Objective values of these solutions are displayed in the title. The number of z-iterations are $4, 4, 7$ respectively. These local optima each correspond to a certain (but different) decomposition of the 6 Gaussian components. Even though the algorithm is intended for $K = 2$, the similar objective values for these apparently different locally optimal clusterings suggest multiple clusters.

tools are suggested by theory in previous chapters, but would implicitly assume the U or other level sets have sufficient regularity to work well.

Level sets and gradient vector field

The piecewise linear solution in 1-d is easy to imagine, but is much less easy to visualize in higher dimension. In view of N.C.1, a multiple-path plot (Figure 3.6) is used to visualize the approximate gradient directions, where each path roughly shows how the clustering function increases from 0 to 1 (or decreases from 1 to 0).

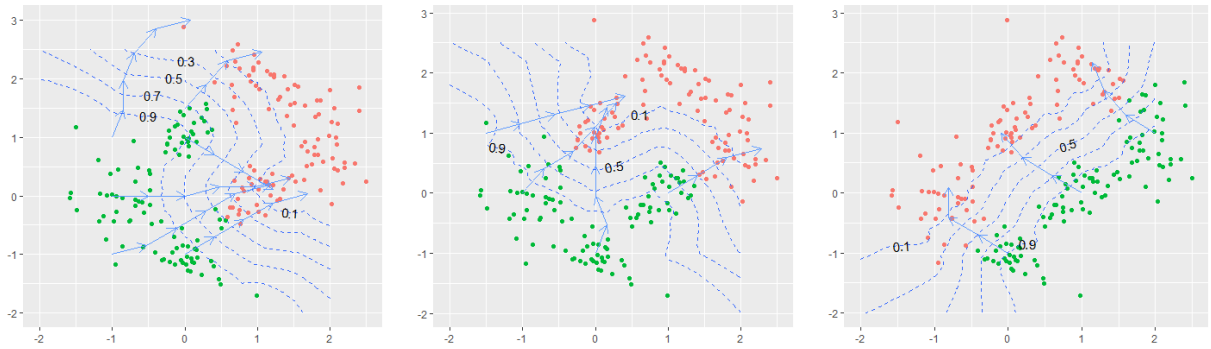


Figure 3.6: Visualization in 2-dimension. The arrows indicate approximate local gradient directions. Dashed curves are level sets plotted as contours of the g function. The three plots correspond to the three solutions in Figure 3.5.
The level sets and gradient-paths are drawn separately – level sets are created from contour plots of g_n, and each arrow is a continuation of the previous arrow on the same path in the following way: for a prescribed step size, determine the steepest descent/ascent direction from a list of angles. Theoretically, the level sets and the gradient directions should be orthogonal, as step sizes go to zero, and if the steepest direction could be computed exactly.

Check necessary condition for optimality empirically

We can check empirically whether the solution computed from Algorithm 1 satisfies N.C.2, by comparing function value with the prediction by N.C.2 based on an approximate set U. Specifically, let L_n be the estimated Lipschitz constant of a solution g_n computed from Algorithm 1, and $\tilde{U}$ be a finite set of points as an approximation to $U_n = \{g_n = 1/2\}$. Suppose N.C.2 is satisfied for g_n, then for any point x we should have either $g_n(x) \approx \max\{d(x, \tilde{U}), 0\}$ or $g_n(x) \approx \min\{d(x, \tilde{U}), 1\}$. For convenience, we check this on data points, and compute the difference between actual function values and the prediction given by N.C.2 (Table 3.12). In fact, there is no reason not to check at some other points.

type of difference	median	mean	max
solution 1	0	0.0018	0.0182
solution 2	0	0.0015	0.0093
solution 3	0	0.0030	0.0592

Table 3.12: Each row presents median/average/maximum difference (among $n = 200$ data points) between function values of a candidate solution with the prediction by N.C.2, using an approximate U set and the estimated L. The three solutions correspond to the ones postulated in Figure 3.5 for the 6-component Gaussian data. Here $\tilde{U} := \{x : |g_n(x) - 1/2| \leq 0.01, x$ is on the grid $[-2, 2] \times [-2, 2.5]$ with cell length $0.05\}$. All three solutions agree closely with N.C.2, while solution 1 and solution 2 appear better.

Margin plot

The variational problem (1.5) is mathematically well-defined for any P_X. This, however, means we could still get the same kind of solution as in Figure 3.1 for a 1-d uniform distribution on a compact interval. Therefore, it is important to develop further diagnostic tools to answer the question: does the solution suggest there is a strong/weak clustering structure for the underlying distribution?

Figure 3.7 offers a natural way to visualize empirical probability measure according to distance from the decision boundary U. The x-axis represents $d(\cdot, U)/L$, the distance to U normalized by the Lipschitz constant, and y-axis represents $P_n(\{x : d(x, U)/L \leq t\})$. It suffices to plot normalized distance within $[0, 1/2)$, outside which function values will be either 0 or 1 by N.C.2 (where points are perfectly clustered). A positive feature of such a "margin" plot is that it is always a 1-d plot, regardless of the data dimension. Ideally or as sample size tends to infinity, for uniform distribution the plot will be exactly a slope. For well-separated clusters, the y-coordinate will be 0 for small distances.

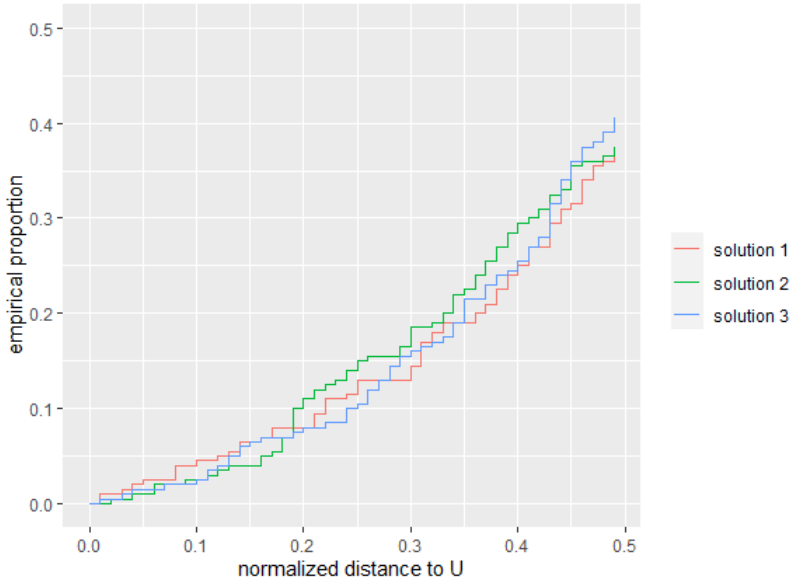


Figure 3.7: Margin plots of the three solutions in Figure 3.5 for 6-Gaussians data. "Normalized" means the distances are divided by the Lipschitz constant, so that the plot is independent of scale. Ideally, a curve that stays close to 0 at the beginning and goes up rapidly when x-coordinate approaches $1/2$ would indicate a strong clustering structure.

3.8 Tuning parameter selection

In previous sections we adopted preliminary choice of tuning parameters: $\lambda_2 = 0.1 * mean(D), \lambda_3 = 0.5$, where $mean(D)$ denotes average of the entries of distance matrix D. This choice makes the procedure invariant under uniform scaling of variables, but is at best an empirical choice. In this section we first discuss the conceptual problem of tuning parameter selection related to cross validation and stability ideas. Much of the discussion is generic for clustering regardless of the method being used (some literature review is supplied at the end). Then we propose a practical procedure to select tuning parameters in our case.

Cross validation and cross stability

This section discusses the conceptual question of tuning parameter selection. We illustrate through the notion of cross validation (for classification), cross stability (for clustering), and another intermediate notion, with the goal of clarifying their differences and relations.

For simplicity and clarity of illustration, consider randomly splitting a dataset X (size n) into two subsets $X_{(1)}$ and $X_{(2)}$, where (1) and (2) denote two subsets of $\{1, \cdots, n\}$. Let $Y_{(1)}, Y_{(2)}$ denote vectors of true classes associated with $X_{(1)}, X_{(2)}$. The two subsets are not necessarily equal-sized, and more generally, they can be pairs of random subsamples whose union is only part of the original dataset. For a given tuning parameter set λ, let $\hat{g}^{(1)}$ and $\hat{g}^{(2)}$ denote the clustering functions computed from $X_{(1)}, X_{(2)}$ respectively by Algorithm 1.

1. Cross-stability of clustering for a given λ: When $Y_{(1)}, Y_{(2)}$ are not observed, we may evaluate the stability or sensitivity of clustering by comparing the two clustering functions estimated from the two data subsets. The comparison is based on either some function norm (we use L_1 below) of the difference between g-functions, or a binary-based metric applied to the vector of cluster labels. We refer to [68] for related definitions.

L_1 **cross-stability**: let

$$d(\hat{g}^{(1)}, \hat{g}^{(2)}) := E|\hat{g}^{(1)}(X) - \hat{g}^{(2)}(X)|, \tag{3.8}$$

with the empirical version $\frac{1}{n}\sum_{i=1}^{n} |\hat{g}^{(1)}(X_i) - \hat{g}^{(2)}(X_i)|$. The L_1 cross-stability measure for the

clustering, after alignment of the classes, is defined as

$$Instab(\lambda) := \min\{d(\hat{g}^{(1)}, \hat{g}^{(2)}), d(\hat{g}^{(1)}, 1 - \hat{g}^{(2)})\}, \tag{3.9}$$

named "instability" [5] because larger difference between $\hat{g}^{(1)}$ and $\hat{g}^{(2)}$ indicates less stability.

Binary-based cross-stability: a binary clustering rule f can be obtained from a continuous clustering function g by $f(x) = I\{g(x) > 1/2\}$ (or some other thresholding), then define

$$d(\hat{f}^{(1)}, \hat{f}^{(2)}) := E[I\{\hat{f}^{(1)}(X) = \hat{f}^{(2)}(X)\}], \tag{3.10}$$

with the empirical version $\frac{1}{n}\sum_{i=1}^{n} I\{\hat{f}^{(1)}(X_i) = \hat{f}^{(2)}(X_i)\}$. The corresponding cross-stability measure is

$$Instab(\lambda) := \min\{d(\hat{f}^{(1)}, \hat{f}^{(2)}), d(\hat{f}^{(1)}, 1 - \hat{f}^{(2)})\}.$$

2. Cross-validation error of clustering for a given λ: This is an intermediate stability measure between cross-stability and cross-validation. Suppose we fit a clustering rule $\hat{f}^{(1)}$ using $X_{(1)}$, and validate by $(X_{(2)}, Y_{(2)})$; fit a clustering rule $\hat{f}^{(2)}$ using $X_{(2)}$, and validate by $(X_{(1)}, Y_{(1)})$ (i.e., have access to $Y_{(1)}, Y_{(2)}$ for testing but not training), then a binary-based cross-validation error is

$$\frac{1}{n}\sum_{i=1}^{n}\left[\sum_{i\in(2)} I\{\hat{f}^{(1)}(X_i) = Y_i\} + \sum_{i\in(1)} I\{\hat{f}^{(2)}(X_i) = Y_i\}\right]. \tag{3.11}$$

It is important to note that the $\hat{f}$ here depends on X only and not on Y. Expression (3.12) looks the same, but in it $\hat{f}$ depends on both X and Y.

This measure offers a theoretical way to justify clustering. It is only available in a Monte

Carlo study. The main purpose of introducing this measure is to provide a bridge between the purely unsupervised cross-stability (above) and the purely supervised cross-validation error (below).

Remark. This setting is not to be confused with a "semi-supervised" setting where the training set is partially labeled. Here the training set is still unlabeled.

3. Cross-validation error of classification for a given λ: In the classification setting, $Y_{(1)}, Y_{(2)}$ are observed. In our case, the classification function $g^{(1)}, g^{(2)}$ are obtained from (3.6) followed by (3.4), and the binary classification rule $f^{(1)}, f^{(2)}$ can be obtained from $g^{(1)}, g^{(2)}$ in the same way as above. The cross validation error is

$$\frac{1}{n}\sum_{i=1}^{n}\left[\sum_{i\in(2)} I\{\hat{f}^{(1)}(X_i) = Y_i\} + \sum_{i\in(1)} I\{\hat{f}^{(2)}(X_i) = Y_i\}\right]. \tag{3.12}$$

Although (3.11) and (3.12) have the same expression, the classifier $\hat{f}^{(1)}$ in (3.12) is a function of $(X_{(1)}, Y_{(1)})$, while in (3.11) the clustering function $\hat{f}^{(1)}$ is a function of $X_{(1)}$ only.

The Monte Carlo study in section 3.9 implements these measures, with a comparison in Table 3.14.

Discussion about cross validation and clustering stability literature

Early work about theory of cross-validation type procedures includes [53]. A paper by Jun Shao [51] studies cross-validation in the context of linear model selection. Much later work surrounds using cross validation to select tuning parameters in nonparametric classification and regression, see [70, 71], and [21].

While the literature for supervised problems is mathematically rigorous, the same cannot

be said for unsupervised problems such as clustering. Although there are analogous definitions of cross-validation error [68], and other practically useful metrics such as rand index [47], there seems no way internally to provably tell the quality of clustering by a stability approach alone. Some attempts in the context of clustering, often under the name "clustering stability", include [5, 66] in the machine learning literature, and [68] in the statistics literature, among others, but the extent of these results are limited. In some sense, the limit theorem in [5] is not so useful, because any clustering with a unique population solution has stability going to zero in the large sample limit. The issue in [68] is that it relies on an extra definition, that the true clustering is the one that maximizes stability.

Overall, cluster stability is an important conceptual question that cries out for further investigation, although perhaps the question of consistency should be addressed first before any procedure to select the tuning parameter is proposed. That is, these stability questions should only be rendered within a set of consistent clustering rules under the model, and talking about tuning parameter selection alone without appropriate model restriction can be a void question. As an example, consider a "constant clustering" that gives out the same clustering regardless of data. Such a procedure is stable, which already questions the validity of many stability measures, but a constant solution will not be statistically consistent in the first place.

A stability-based selection procedure

Practically, one could find λ values that minimize the stability measure (3.9). To deal with multiple tuning parameters, we minimize "coordinate-wise", in Algorithm 3.

Some justification for this stability-based selection is given in the next Monte Carlo section

Algorithm 3 Tuning Parameter Selection

1. Initialize or use preliminary choice of $(\lambda_2^0, \lambda_3^0)$;

   For $k = 0, 1, \cdots$, and let s_2^k, s_3^k be finite sets (may or may not depend on k) where λ_2, λ_3 are chosen from.

2. Fix λ_3^k, $\lambda_2^{k+1} \leftarrow \min_{\lambda_2 \in s_2^k} Instab(\lambda_2, \lambda_3^k)$, where $Instab$ is the cross stability measure defined in (3.9);

3. Fix λ_2^{k+1}, $\lambda_3^{k+1} \leftarrow \min_{\lambda_3 \in s_3^k} Instab(\lambda_2^{k+1}, \lambda_3)$;

   Repeat step 2 and 3, stop when $Instab(\lambda_2^{k+1}, \lambda_3^{k+1}) \approx Instab(\lambda_2^k, \lambda_3^k)$.

3.9, where the value of this stability statistic is shown to be close to the cross validation error in the supervised case under the same choice of λ.

Checking theoretical conditions for tuning parameters

In Chapter 2, theoretical conditions for tuning parameters are given under which consistency and bipartition holds. We give a basic example to check these conditions in practice: the uniform distribution on $[0, 1] \cup [2, 3]$, with tuning parameters $\lambda_2 = 0.01, \lambda_3 = 0.04$. In this case the quantities involved in the conditions can be exactly computed.

In order to check for bipartition (C3), recall that the parameter c involved in the condition depends on the following quantity about P_X:

$$C_{S_k}(L, a, b) = \inf\{\int_{S_k} f dP : f(x_1) = a, f(x_2) = b, x_1, x_2 \in S_k, L(f) \leq L, 0 \leq f \leq 1\}.$$

Let c_k be the number such that $C_{S_k}(L_0 + C\alpha_0, \frac{c_k}{\pi_k}, \frac{1}{2}) = c_k, k = 1, 2$, then $c := \min\{c_1, c_2\}$.

In the given uniform case, $\alpha_0 = \pi_1 = \pi_2 = 1/2, L_0 = 1$. By symmetry, $C_{S_1}(L, a, b) =$

$C_{S_2}(L, a, b)$ and $c_1 = c_2$. Now suppose $0 \leq a < b \leq \frac{1}{2}, L > 1$, then

$$C_{S_1}(L, a, b) = \int_0^{b/L} (b - Lx)dx = \frac{b^2}{2L},$$

so we obtain in this case $c = C_{S_1}(L_0 + C\alpha_0, \frac{c}{\pi_1}, \frac{1}{2}) \equiv \frac{1/4}{2+C}$. Let $C = 4$, then $c = \frac{1}{24}$.

Check C2 (condition for consistency): $2L_0/(1 - \alpha_0) = 2/0.5 = 4 = \frac{\lambda_3}{\lambda_2} = C$.

Check C3 (bipartite condition): $\frac{L_0}{1-c-\alpha_0} < \frac{L_0}{1-\alpha_0} = 2 < \frac{\lambda_3}{\lambda_2} = C$, and $\lambda_2 L_0 + \lambda_3 \alpha_0 = 0.01 + 0.02 = 0.03 < c$.

Therefore, both conditions are satisfied.

3.9 Monte Carlo study 3: classification and clustering in two-component Gaussian mixture model

In this section we consider a two-component Gaussian mixture in 2-d, with class proportion $\pi = (0.4, 0.6)$, mean vectors $\mu_1 = (-1, 1), \mu_2 = (1, -1)$, and common covariance matrix $\Sigma^{1/2} = \begin{pmatrix} 1.5 & 0 \\ 0 & 2 \end{pmatrix}, n = 80$. One goal of this section is to illustrate the two notions: cross-stability for clustering and cross-validation for classification introduced in the last section.

Classification and clustering

For fixed λ, we compute clustering (without labels) and classification (with labels) based on the same (many) training set/testing set split, compare the misclustering error (an average cross-stability error) and misclassification error (an average cross-validation error), also with the monte carlo truth for probability of "correct" clustering.

Part 1

The average confusion matrix from 100 Monte Carlo simulations for clustering and classification (use corresponding algorithm under setting 2 and setting 3 in Table 3.8) is given in Table 3.13, using $\lambda = (0.5, e^{-1}, 0.5)$. The label "1" and "2" denotes true labels for the two gaussians centered at μ_1 and μ_2. To make sure clusters are aligned, one should either choose the same initial z each time or align each simulation (find the best confusion matrix, i.e., closest to diagonal, among all permutations of columns).

	0	1
1	0.3432	0.0486
2	0.1562	0.4519

	0	1
1	0.3564	0.0355
2	0.1481	0.4600

Table 3.13: Left: average confusion matrix for clustering; Right: average confusion matrix for classification.

In Table 3.13, the average confusion matrix for clustering is close to the one for classification, indicating that in this normal mixture example, the clustering approach works reasonably when labels are unobserved in the sense that it achieves similar performance as the supervised setting.

Part 2

In Part 1, clustering error and classification error are evaluated using the true labels in the monte carlo sample. Since these monte carlo estimates of the clustering/classification errors are not available for a real dataset, moreover, in the clustering setting the labels are not observed, we need some internal measures (based on available data) to evaluate these errors – in particular, cross validation error (for classification) and cross stability (for clustering) as an "estimate" of

the true error. Definition of these measures are given in section 3.8.

In Table 3.14, the same λ value is used as part 1 and the four measures are compared based on 50 equal-sized random splits of the data. The four measures are comparable, indicating that cross-stability for clustering (in the unsupervised setting) is closely resembles the cross-validation error for classification.

	L_1 cross-stability	binary cross-stability	intermediate	cv error
average	0.2798	0.3105	0.2995	0.2680
SD	0.0979	0.1260	0.0788	0.0623

Table 3.14: Four stability/validation measures (described in section 3.8) computed from 50 equal-sized random splits of a two-component Gaussian mixture data with sample size 80, using the same tuning parameter λ: (1) L_1 cross-stability for clustering; (2) binary cross-stability for clustering; (3) cross-validation error under the intermediate setting; (4) the familiar "K-fold" cross-validation error for classification with $K = 2$.
There is a natural order among the last three: the purely supervised cross-validation error is the smallest, then the intermediate one, the purely unsupervised binary cross-stability is largest. Interestingly, the same order also holds for standard deviation.
The four measures in the table are comparable, indicating that cross-stability for clustering (in the unsupervised setting) closely resembles the cross-validation error for classification.

Remark. The order among the standard deviations for the last three measures may suggest a theorem to prove under certain conditions.

3.10 Real data analysis: Boston housing data

In this section, we apply our method to the Boston housing dataset. This dataset was used as an illustrative example in [23], and is available in R package "MASS". The sample size is $n = 506$, with $p = 13$ variables. We use the same variable transformations following [23]: $x_{(1,3,5,6,8,9,10,14)} = log(x_{(1,3,5,6,8,9,10,14)}), x_7 = x_7^{2.5}, x_{11} = exp(0.4x_{11}), x_{13} = \sqrt{x_{13}}$. After the transformation, all variables are then scaled. The river variable (binary) is deleted from the study.

In [23], hierarchical clustering with Wald's method is presented, and interpretation of the result shows evidence for two clusters: towns with high living quality and towns with low living quality.

We run Algorithm 1 on full data using default tuning. Four initialization procedures are compared: (1) 10 random initializations; (2) hierarchical clustering (Wald's method); (3) K-means clustering; (4) spectral clustering. From Figure 3.8 and 3.9, except for the random initialization, all other three initialization methods give sensible but slightly different clusterings indicating high and low living quality, each as a local optimum of the objective function found by Algorithm 1.

Ordering among solutions

By looking at several variable pairs in Figure 3.8 and 3.9, the clusterings given by the three initializations (hierarchical, K-means and spectral) can be seen as successively adding more towns to the high living-quality group (more red dots replacing the black ones in the figure), leading to more unbalanced clusters. A further investigation shows that this seemingly clear order in the plot comes from comparison between the labelings (z values), while such order is not clearly present when values of the clustering function (g values) are compared instead.

Trade-offs among solutions

In Table 3.15, the classification problem found by spectral initialization has the lowest classification error and is most smooth (smallest Lipschitz constant) among the methods presented, but the unbalanced estimated proportions and an extreme grouping of several variables such as "rad", "tax" and "pratio" in Figure 3.8 and 3.9 may need further investigation. The second and

third row under the hierarchical and K-means column show a typical trade-off between Lipschitz constant (smoothness) and classification error: the solution from K-means is smoother, but has a larger classification error.

initialization method	random (10)	hclust	K-means	spectral
objective value	0.490	0.440* (0.444)	0.449 (0.462)	0.509 (0.509)
classification error	0.094	0.065 (0.068)	0.086 (0.078)	0.045* (0.045)
Lipschitz constant	0.308	0.265 (0.264)	0.238 (0.229)	0.237* (0.237)
proportion	0.50	0.50 (0.50)	0.50 (0.45)	0.30* (0.30)
number of iterations	10	3	6	1

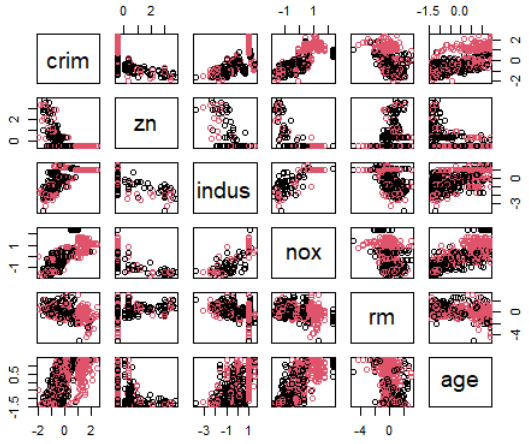
Table 3.15: Clusterings found by different initializations. Several output quantities of interest are displayed. Data: Boston housing. Initializing by hierarchical clustering gives the smallest objective value among the methods presented here. * indicates the smallest number within each row. The values inside brackets are "initial input values" from the output of another clustering algorithm, by passing only one iteration to turn a discrete clustering into a Lipschitz continuous one. Not all quantities are improved from the initial, as our objective function tries to find a certain balance among them.
It is important to note that, as mentioned in section 3.6 and Table 3.10, the "classification error" here refers to $I_{n,1}(g_n)$, and since there is no "true label" for this dataset, the "true classification error" $R_n(g_n)$ is not possible to get.

3.11 Appendix

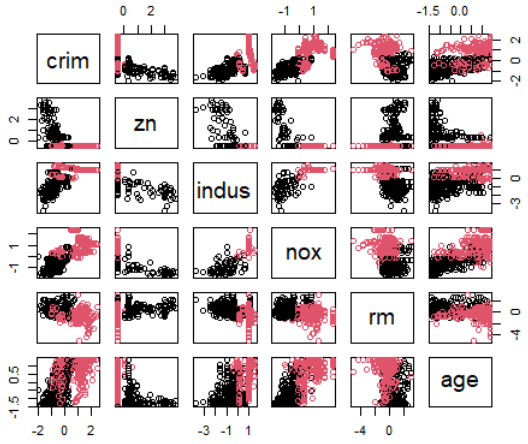
3.11.1 Lloyd's algorithm and Luxburg's linear program

Lloyd's algorithm for K-means clustering

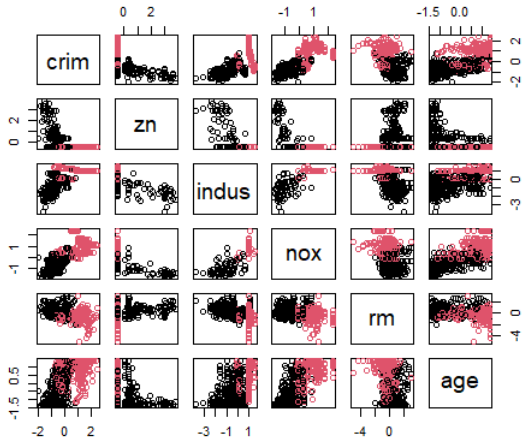
We briefly recall Lloyd's algorithm [34] because the alternating minimization steps therein bear much high-level resemblance to the our main algorithm. Since exact minimization of K-means (1.9) requires combinatorial optimization, in practice often a heuristic algorithm is implemented. The most common one is Lloyd's algorithm, which proceeds by iterative refinement of



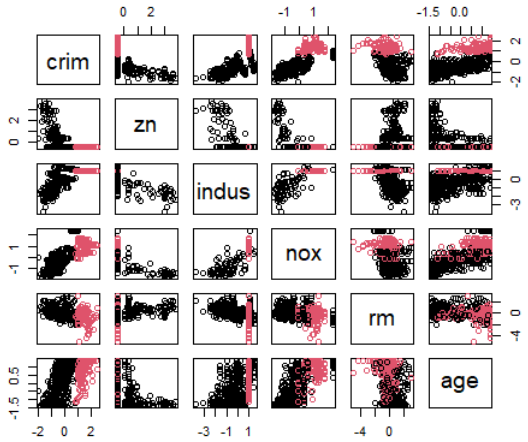
(a) initial = random



(b) initial = hclust



(c) initial = kmeans



(d) initial = spectral

Figure 3.8: Pairwise clustering plot using different initialization algorithms. Data: Boston housing. Variable 1-6.

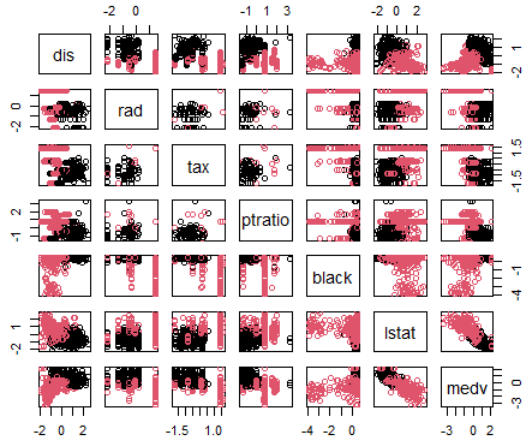
the cluster centers and cluster assignments:

Step 1 Pick K data points randomly as initial cluster centers $c_1^0, \cdots, c_k^0$. Then for $t = 0, 1, 2, \cdots$:

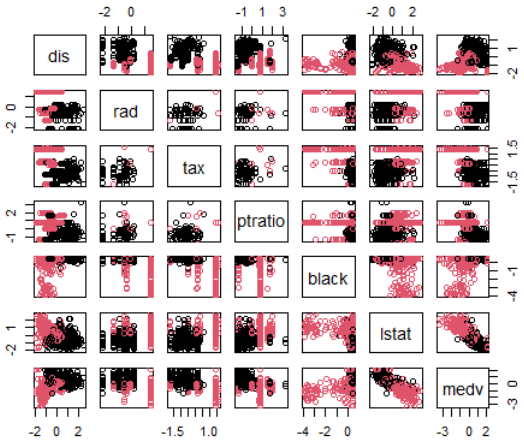
Step 2 Given the cluster centers $c_1^t, ..., c_k^t$, update the membership of X_i's: $X_i \in C_j^{t+1}$ iff $c_j^t = \arg\min_{c \in \{c_1^t, ..., c_k^t\}} ||X_i - c||^2$.

Step 3 Given the membership information, i.e. $\{i : X_i \in C_j^{t+1}\}$ where C_j^{t+1} denotes jth cluster,

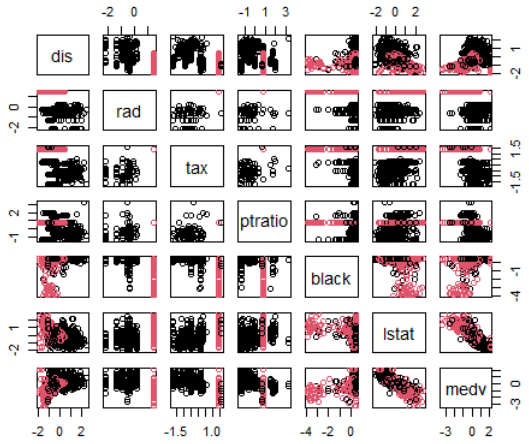
(a) initial = random



(b) initial = hclust



(c) initial = kmeans



(d) initial = spectral

Figure 3.9: Pairwise clustering plot continued: variable 7-13.

update the cluster centers by $c_j^{t+1} = \dfrac{\sum\limits_{X_i \in C_j^{t+1}} X_i}{|C_j^{t+1}|}, j = 1, \cdots, k.$

Step 4 Repeat steps 2& 3 until certain convergence criteria is met: when decrease in the objective value (1.9) is smaller than some tolerance, or when cluster assignments no longer change.

Now we explain the high-level resemblance of these steps to our main algorithm (Algorithm 1). In terms of the mixed-integer formulation (3.2), the integer variable z resembles memberships

of data points, while the continuous variable a resembles values of the clustering function. Each iteration of Algorithm 1 can be understood in this way: when the membership of data points are fixed, the problem becomes a classification problem, in our case it is essentially a Lipschitz classifier known in the literature (introduced next), which yields a fitted classification/clustering function (corresponding to step 3 above); when the classification/clustering function is fixed, data points are then relabeled (corresponding to step 2 above).

Luxburg's linear program for Lipschitz classifier

A Lipschitz classifier [65] minimizes the classification error while controlling the Lipschitz constant of the decision function:

$$\inf_{f \in Lip(\mathcal{X})} l(y_i f(x_i)) + \lambda L(f)$$

where $y_i \in \{-1, 1\}$, and l is the hinge loss for classification: $l(y_i f(x_i)) = \max\{0, 1 - y_i f(x_i)\}$.

[65] fits the Lipschitz classifier by solving

$$\min_{a_1, \cdots, a_n} \sum_{i=1}^{n} l(y_i a_i) + \lambda \max_{i,j} \frac{a_i - a_j}{d(x_i, x_j)}, \tag{3.13}$$

which can be written into a linear program:

$$\min_{a_1,\cdots,a_n} \sum_{i=1}^{n} \xi_i + \lambda\rho$$

$$\text{subject to } \xi_i \geq 0,$$

$$y_i a_i \geq 1 - \xi_i,$$

$$\rho \geq \frac{a_i - a_j}{d(x_i, x_j)}.$$

Remark. In [65], the range for the classifier f is all real numbers, while the range of the clustering function g we adopt is in $[0, 1]$.

3.11.2 Additional simulations on distributional effects

Figure 3.10 shows the effect of distribution on the clustering result over a family of beta distributions, replacing the uniform density in the two-piece uniform example. Figure 3.11 and Table 3.16 shows the effect of radial distribution on $1/L_n$ in the disk and annulus data.

radial distribution	$1/L_n$
beta(5,1)	1.22
beta(1,1)	1.75
beta(1,3)	2.27

Table 3.16: Distribution effect on margin, under the three disk and annulus data described in Figure 3.11, using different beta distributions in the radial direction. The "margin" $1/L_n$ becomes larger when the distribution on the clusters concentrate away from each other, which matches intuition: degree of separation is higher from left to right in Figure 3.11.

3.11.3 Subsampling inference for computationally infeasible problems

Consider estimating a statistical functional $\theta(P)$ by an estimator $\hat{\theta}_n$ (often formulated as solution to an optimization problem), where $\hat{\theta}_n$ is hard to compute for very large n (either infeasible or the complexity grows too fast with n). Using classical empirical bootstrap for statistical inference, such as constructing CI for $\theta(P)$, requires computing $\hat{\theta}_n$ repeatedly for resamples of size n, which becomes even more computationally prohibitive. On the other hand, m-out-of-n bootstrap is often able to produce equally valid inference under the condition that $m \to \infty, m/n \to 0$ [41]. Such subsampling procedure can involve more postprocessing steps than empirical bootstrap, but doing so are worthwhile as much more computational cost is saved by computing $\hat{\theta}_m$ instead of $\hat{\theta}_n$. This may turn an infeasible problem to a feasible one, or reduce the complexity significantly. For example, when the complexity of the optimization problem is $O(n^3)$, then taking $m = \sqrt{n}$ reduces each run to $O(n^{1.5})$, and if $B = \sqrt{n}$ subsamples are taken then the total cost is $O(n^2)$, which is a solid reduction in computation. Computing was not one of the main motives to use subsampling in [41, 42] at that time, instead these approaches were advocated so as to weaken the assumptions to apply bootstrap methods for valid inference, or as a remedy in cases when empirical bootstrap fails.

3.11.4 Consistency of aggregated solution

This section studies consistency of a solution computed from the subsample-aggregate approach proposed in Algorithm 2, which may be of independent interest. The main technical tool is U-statistics.

Suppose the original dataset has size n while we utilize B repeated subsamples with size

m to estimate g_n, the original solution at sample size n, because of computational constraints. We study consistency of such an estimate when both m and B grow with n, but possibly slowly. In summary, consistency follows from two parts:

1. the aggregated solution is close to the "average behavior" at size m;

2. the "average behavior" at size m is consistent as long as m grows with n.

The first part will be dealt with under a specific subsampling scheme introduced below (see later remark for the difference between this scheme and the usual m-out-of-n bootstrap). The second part is straightforward.

For any $m < n$, conditioning on data $X_1, \cdots, X_n$, let $g_{m,1}, g_{m,2}, \cdots, g_{m,\binom{n}{m}}$ denote any (fixed) ordered solutions from all $\binom{n}{m}$ data subsets of size m. Consider the following subsampling and aggregating scheme: Let $g_1^*, \cdots, g_B^*$ be i.i.d and for any $b = 1, \cdots, B, g_b^* = g_{m,i}$ with probability $\frac{1}{\binom{n}{m}}, i = 1, \cdots, \binom{n}{m}$. Define the aggregated solution

$$\hat{g}_{ag} := \frac{1}{B} \sum_{b=1}^{B} g_b^*.$$

Under this scheme, the randomness in this average is split into two parts: one involving the symmetric dependence structure among all subsamples of size m, which will be dealt with by a U-statistics result; the other is about an average of B independent random variables, conditioning on the fixed order of $\binom{n}{m}$ solutions.

In the preceding theorem, we assume the alignment problem in clustering can be dealt with separately. One could choose a base point in the data space (e.g., a point far-away from all data points) and require that all g functions have the same value (0 for example) at this point.

Theorem 3.1. *Suppose the variational problem (1.5) has solution g^*, and consider any point x*

at which the data problem is consistent: $g_n(x) \xrightarrow{p} g^*(x)$. *When* $B \nearrow \infty, m \nearrow \infty, m = o(n)$, *the subsample-aggregate scheme above is also consistent at this point:* $\hat{g}_{ag}(x) \xrightarrow{p} g^*(x)$.

Remark (1). Section 2.6 summarizes various forms of uniqueness/consistency results around the original population/data problem, indicating different extents to which consistency property holds. The focus here is to show what are the additional analysis needed to justify a subsampling procedure. With suitable choice of m, B, subsampling can be applied where consistency and uniqueness is justified or assumed for the original data and population problem, and where the major concern is computation.

Remark (2). There is some difference between the subsampling studied here and what is usually implemented in an m out of n bootstrap with replacement. Here although some of B subsamples may be repeated, there are no repeated observations within each subsample. This difference can be handled by the difference between U-statistic and V-statistic discussed in the next section.

Proof. Let $U_n(x) := \frac{1}{\binom{n}{m}} \sum_{i=1}^{\binom{n}{m}} g_{m,i}(x)$. Then $U_n(x)$ is a U-statistic with degree m. By an exponential inequality for U-statistic [A.8] which dates back to Hoeffding, and the fact that all g values lies in $[0, 1]$, we have, unconditionally,

$$P(U_n(x) - E[U_n(x)] \geq t) \leq \exp(-2\lfloor n/m \rfloor t^2), \tag{3.14}$$

where $E[U_n(x)] = E[g_{m,i}(x)]$.

Note that conditioning on $X_1, \cdots, X_n$, under the subsampling scheme we have

$$E[g_b^*(x)|X_1, \cdots, X_n] = U_n(x),$$

and by Hoeffing inequality for i.i.d sum,

$$P(\frac{1}{B}\sum_{b=1}^{B} g_b^*(x) - U_n(x) \geq t | X_1, \cdots, X_n) \leq \exp(-2Bt^2).$$

We also have, unconditionally,

$$\begin{aligned} P(\frac{1}{B}\sum_{b=1}^{B} g_b^*(x) - U_n(x) \geq t) &= E[E[I(\frac{1}{B}\sum_{b=1}^{B} g_b^*(x) - U_n(x) \geq t) | X_1, \cdots, X_n]] \\ &= E[P(\frac{1}{B}\sum_{b=1}^{B} g_b^*(x) - U_n(x) \geq t | X_1, \cdots, X_n)] \\ &\leq \exp(-2Bt^2). \end{aligned} \tag{3.15}$$

Combining (3.14) and (3.15),

$$\begin{aligned} P(\frac{1}{B}\sum_{b=1}^{B} g_b^*(x) - E[U_n(x)] \geq t) &= P(\frac{1}{B}\sum_{b=1}^{B} g_b^*(x) - U_n(x) + U_n(x) - E[U_n(x)] \geq t) \\ &\leq P(\frac{1}{B}\sum_{b=1}^{B} g_b^*(x) - U_n(x) \geq t/2) \\ &\quad + P(U_n(x) - E[U_n(x)] \geq t/2) \end{aligned}$$

(using $P(X+Y \geq t) \leq P(\{X \geq t/2\} \cup \{Y \geq t/2\}) \leq P(X \geq t/2) + P(Y \geq t/2)$)

$$\leq \exp(-Bt^2/2) + \exp(-\lfloor n/m \rfloor t^2/2).$$

This shows when $B \nearrow \infty, m = o(n)$, the value of the aggregated solution at any point x, $\hat{g}_{ag}(x)$ is close to $E[U_n(x)] = E[g_{m,i}(x)]$, the average behavior at sample size m. This finishes the first part.

For the remaining part, note that whenever the original solution (at size n) is consistent,

then as long as m grows with n (which can grow slowly), the average behavior at size m will also be consistent. Specifically, suppose $g_n(x) \xrightarrow{p} g^*(x)$, then by [A.12], $E[U_n(x)] = E[g_{m,i}(x)] \xrightarrow{n\to\infty} g^*(x)$. □

Further remarks on Theorem 3.1:

Remark (1). From this analysis, m, B can even grow as slowly as $\log(n)$ to be consistent. However, the actual performance at any given sample size needs to be checked by Monte Carlo.

Remark (2). The aggregation method studied here is pointwise average. An averaged solution, however, no longer satisfies the optimality conditions. Another method implemented is pointwise median (see step 3 of Algorithm 2 and Figure 3.2), which seems to yield finite sample solutions that roughly satisfy optimality conditions. The proof could be a bit different for the median approach.

Remark (3). Different tools will be needed to prove consistency for m fixed, or under an adaptive scheme other than pure random subsampling, such as described previously in "subsample discriminating points".

3.11.5 U-statistics and V-statistics

Definition: for any "kernel" h, a permutation symmetric function of its arguments,

U-statistic: $U_n = \frac{1}{\binom{n}{m}} \sum_c h(X_{i_1}, \cdots, X_{i_m})$, where c denotes all combinations of m distinct elements $\{i_1, \cdots, i_m\}$ from $\{1, \cdots, n\}$.

V-statistic: $V_n = \frac{1}{n^m} \sum_{i_1=1}^{n} \cdots \sum_{i_m=1}^{n} h(X_{i_1}, \cdots, X_{i_m})$.

It can be seen from the two definitions that the difference is in sampling with or without replacement. Suppose we adopt the aggregated solution from the classical m-out-of-n bootstrap

(sampling with replacement), then the U-statistic $U_n(x)$ appeared in section 3.11.4 will be replaced by the corresponding V-statistic with the same kernel.

A result from Serfling ([49], section 5.7.3) says the asymptotic behavior of U-statistic and V-statistic are very similar–in particular, they share the same central limit theorem. However, one needs to take caution when applying this result, because Serfling's proof argument treats m as a fixed number – while in our case m grows with n.

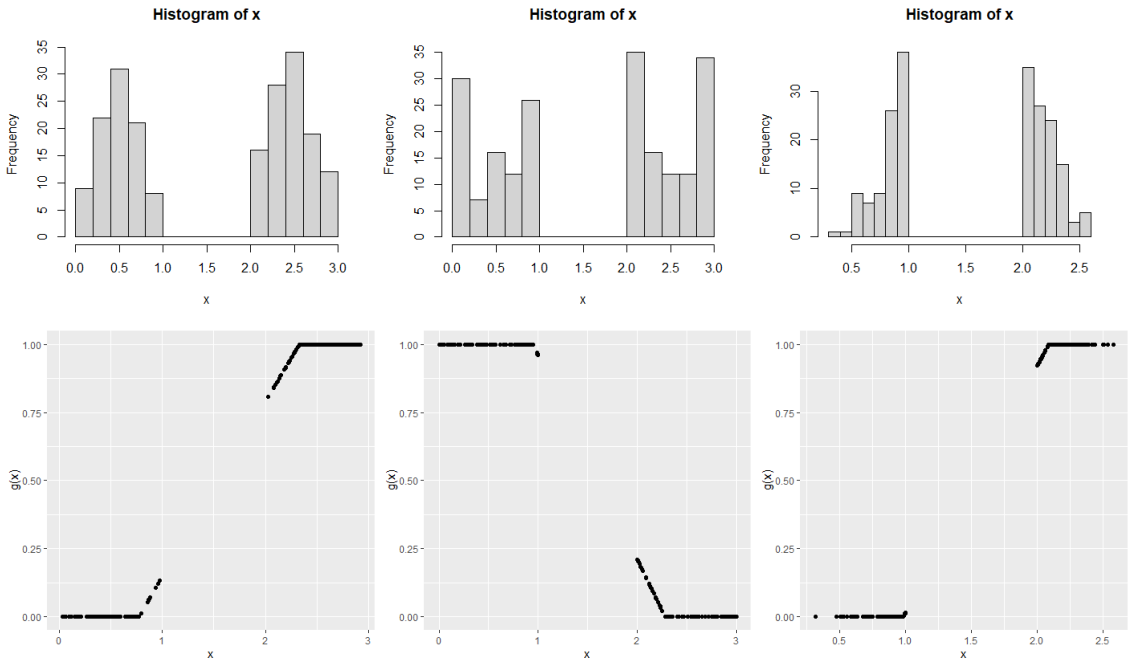
3.11.6 Other subsampling schemes

Subsample the constraints

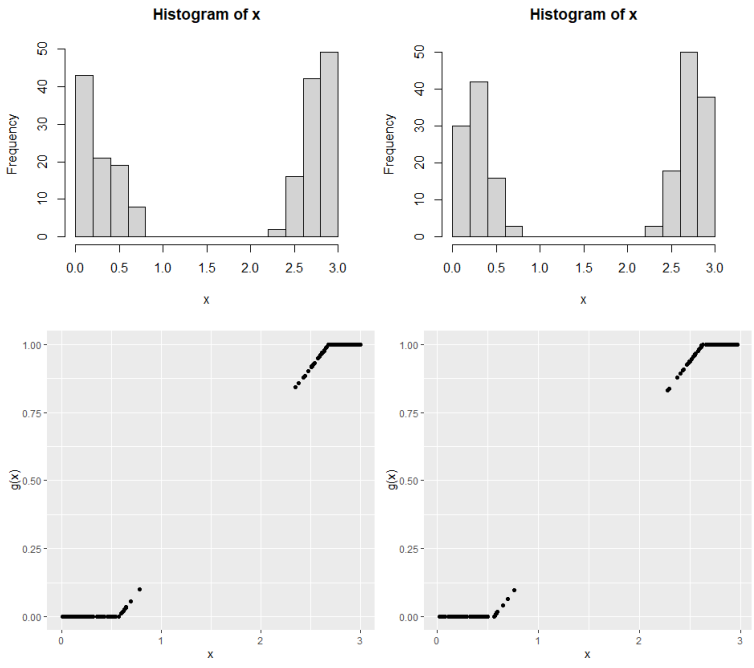
The computational bottleneck of Algorithm 1 depends a lot on the large constraint matrix in the linear program, in terms of both running time and storage. A natural thought is to subsample the constraint matrix. However, it appears that subsampling rows of constraints uniformly can lead to very unstable performance. Further knowledge regarding the potential active constraints is needed to prevent them from being deleted, while getting rid of redundant constraints.

Subsample the variables

Another different idea is to subsample the variables when the number of variables is large, and consider subsampling in both directions (both the data and the variables) when both the sample size and the dimension are large. This direction is out of scope here.

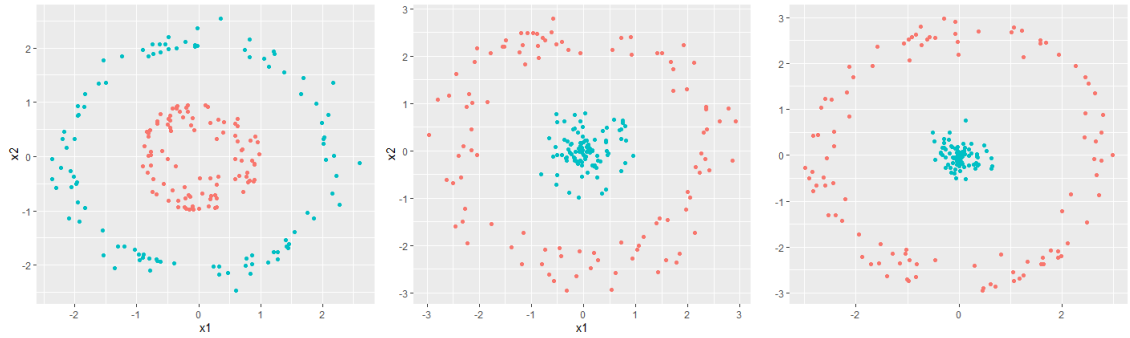


(a) beta(2,2) (b) beta(0.5,0.5) (c) beta(5,1)



(d) beta(1,3) (e) beta(2,5)

Figure 3.10: Fitted values on the data points from beta distributed clusters with different parameters. Actual distribution on each cluster does not show much effect on the form of solution, only the ”turning points” are changed.



(a) beta(5,1) (b) beta(1,1) (c) beta(1,3)

Figure 3.11: Disk and annulus data with different radial distributions, generated by the product of two independent random variables γ and β, where γ denotes uniform distribution on S^1, and β denotes the radial distribution specified by a beta family: β follows $beta(a, b)$ or $(3 - beta(a, b))$ each with probability $1/2$, for some $a > 0, b > 0$. The case $beta(1, 1)$ corresponds to the uniform case used in Table 3.11. Each subfigure (a) (b) (c) has the same support (on the same disk and annulus), but the distribution concentrates differently.

Chapter 4: Generalizations

In this chapter, we initiate three major extensions of (1.5): general penalty function, multiclass, and model with noise – while still under Lipschitz regularization. Section 4.4 discusses possible variants of the formulation other than Lipschitz function spaces for the decision functions g.

Many results for $K > 2$ do not involve new proof ideas, and only require finding the right analogue statements, some can even be repeated verbatim. The emphasis on $K = 2$ in previous chapters is not a limitation: while multiple clusters can be an important practical question, the case $K = 2$ captures the essence of the theory (also more clear for theoretical understanding). The extension from 2 to K is more of an engineering step. We also refer back to section 3.7 for an approach to still use the 2-cluster formulation to explore multiple clusters from further knowledge of multiple local optima.

4.1 General penalty function

In this section we show that N.C.2 holds for general penalty function ρ when $K = 2$. The proof of Theorem 2.1 relies on I_3 only through the comparison Lemma 1.1, and only in the last step. Therefore it suffices to show that such a comparison lemma holds for general penalty functions, with modified requirement on λ_3.

Let $\rho(\cdot)$ be a nonnegative function on $[0,1]$, and we change $I_3 = \lambda_3 \max\{E[g], 1-E[g]\}$ in (1.5) used in previous chapters to $I_3 = \lambda_3 \cdot \rho(E[g])$, with other terms fixed.

Suppose ρ is Lipschitz with constant C:

$$|\rho(x) - \rho(y)| \le C|x-y|,$$

then for any two functions g, g', we have $I_3(g') - I_3(g) \le \lambda_3 \cdot C|E[g'] - E[g]|$. From the proof of Lemma 1.1, this inequality would lead us to the same conclusion in the lemma when $\lambda_3 < 1/C$. Examples covered by this case include $\rho(x) = (1-x)^2 + x^2$ or $\max\{(1-x)^2, x^2\}$.

Corollary 4.1. *The statements in Theorem 2.1 hold for general penalty function ρ (replacing $\max\{E[g], 1-E[g]\}$) satisfying $|\rho(x) - \rho(y)| \le C|x-y|$ for some constant C, for any $x, y \in [0,1]$, when the condition $\lambda_3 < 1$ is replaced by $\lambda_3 < 1/C$, with other things fixed.*

Remark. For penalty functions such as $\rho(x) = \frac{1}{x} + \frac{1}{1-x}$, which are not Lipschitz on the whole interval $[0,1]$, one could first use a competitor (such as a nonzero constant function) to obtain a positive lower bound on $E[g^*]$ for any optimal solution g^* (i.e., excluding the possibility of extreme proportions). Suppose the lower bound is δ, then ρ is Lipschitz within $[\delta, 1-\delta]$, and the same argument above can be applied within this interval. In other words, the solution is unchanged if $\rho(\cdot)$ is replaced by $\min\{\frac{1}{x} + \frac{1}{1-x}, C\}$ for some constant C, which reduces to the case in Corollary 4.1.

4.2 Some multiclass theory

A Pollard-type consistency theorem for $K > 2$ was already provided in Theorem 1.4. Here we consider an analogue of Theorem 2.3 about model-based consistency for $K > 2$, with general penalty function ρ. First we define an extension of the sharp cluster model C1 in Chapter 2.

K sharp clusters

Consider the generating distribution P to be K sharp clusters $S_1, \cdots, S_k \subset \mathbb{R}^d$ with $P(S_k) = \pi_k > 0, k = 1, \cdots, K$, where "sharp" means:

(1) density exists for P and is lower bounded away from 0 on $\cup_k S_k$, constant 0 on $(\cup_k S_k)^C$;

(C4)

(2) S_k's are compact, connected, disjoint.

Denote $L_0 = \max_{i \neq j} \frac{1}{d(S_i, S_j)}, \pi_0 = (\pi_1, \cdots, \pi_K)$. Let $\mathcal{G}_K := \{\{g_1, ..., g_K\} : \sum_{k=1}^{K} g_k = 1, g_k \geq 0, k = 1, \cdots, K\}$ denote the collection of sets of K clustering functions.

For $k = 1, \cdots, K$, let $\tilde{g}_k$ be any function such that $\tilde{g}_k|_{S_k} = 1, \tilde{g}_k|_{S_j} = 0$ for any $j \neq k$, and $\tilde{g}_k|_{(\cup_{j=1}^K S_j)^C}$ is a Lipschitz extension of $\tilde{g}_k|_{\cup_{j=1}^K S_j}$, so that $L(\tilde{g}_k) = \max_{j \neq k} \frac{1}{d(S_j, S_k)}$. Let $\tilde{g} = \{\tilde{g}_1, \cdots, \tilde{g}_K\}$. Define the clustering risk (as in section 2.2.3) of g under K sharp clusters as

$$R(g) := d_H(g, \tilde{g}),$$

where $d_H(g^1, g^2) = \max_i \min_j ||g_i^1 - g_j^2||_{L_1(P)}$ for any $g^1, g^2 \in \mathcal{G}_K$, as in Theorem 1.4.

Now we state some technical conditions on ρ needed in Theorem 4.1 below. Let $\rho(x_1, \cdots, x_K)$

be a nonnegative function on the probability simplex $\{(x_1, \cdots, x_K) : \sum_k x_k = 1, x_k \geq 0\}$, satisfying the following conditions:

1. ρ is continuous, symmetric in its arguments, i.e., $\rho(x_1, \cdots, x_K) = \rho(x'_1, \cdots, x'_K)$ for any $x = (x_1, \cdots, x_K)$ and any permutation $x' = (x'_1, \cdots, x'_K)$ of x.

2. $\rho(x_1, \cdots, x_K) < \infty$, for any $x_1 \neq 0, \cdots, x_K \neq 0$.

3. $\rho(0, x_2, \cdots, x_K) > \rho(y_1, \cdots, y_K)$, for any $x_2, \cdots, x_K$, and $y_1 \neq 0, \cdots, y_K \neq 0$.

Examples of ρ include $\sum_{k=1}^{K} \frac{1}{x_k}, \max_k \frac{1}{x_k}, \max_k (1 - x_k), \max_k (1 - x_k)^2$. In practice, this should be chosen for computational tractability.

Remark. For the first two examples, $\rho(0, x_2, \cdots, x_K) = \infty$; for the latter two, $\rho(0, x_2, \cdots, x_K) < \infty$ (e.g., for $\max_k (1 - x_k)$, $\rho(0, x_2, \cdots, x_K) = 1$). The effect of this difference is discussed in the last part of the proof of Theorem 4.1.

Consistency under sharp cluster model with $K > 2$

With all notations and conditions stated above, we give a generalization of Theorem 2.3. Under model C4,

Theorem 4.1. *Let the generalized data-based objective function be*

$$I_n(g) = P_n[\min_k \{1 - g_k(X)\}] + \lambda_{2,n} \max_k L(g_k) + \lambda_{3,n} \rho(P_n[g_1], \cdots, P_n[g_k]),$$

where ρ satisfies the conditions stated above. Denote any minimizer

$$g_n(\cdot, \lambda_{2,n}, \lambda_{3,n}) \in \arg\min_{g \in \mathcal{G}_K} I_n(g).$$

Suppose $\lambda_{2,n}$ *and* $\lambda_{3,n}$ *are chosen such that* $\frac{L_0}{\min_{x_2,\cdots,x_k} \rho(0,x_2,\cdots,x_k)-\rho(\pi_0)}\lambda_{2,n} < \lambda_{3,n} \leq C\lambda_{2,n}, \lambda_{3,n} \to 0$, C *is any constant, then*

$$\lim_n R(g_n(\cdot, \lambda_{2,n}, \lambda_{3,n})) = 0 \text{ a.s.}.$$

In the case when $\rho(0, x_2, \cdots, x_k) = \infty$, *it is sufficient that* $C_1 \leq \frac{\lambda_{3,n}}{\lambda_{2,n}} \leq C_2$ *for some constants* $0 < C_1 \leq C_2$ *(which does not depend on* L_0 *and* π_0*).*

The proof is in section 4.5.1.

4.3 Sharp cluster model with noise

In this section we consider the sharp cluster with noise model

$$P_X = (1-\delta)P_s + \delta P_\epsilon, \tag{C5}$$

where P_s is a sharp cluster model with $K = 2$ (C1), where $P_s(S_1) = \pi_1, P_s(S_2) = \pi_2$, and P_ϵ is some noise model in the ambient space, δ is the probability of observing noise. We prove a 1-d uniqueness result in such noise case, as an extension to Theorem 2.5.

First, we show that a "bipartite" theorem (Theorem 2.4) is still attainable, with some extra condition involving the noise probability δ.

Perfect separation under noise

Under model C5,

Theorem 4.2. *Let* $\alpha_0 = \max\{\pi_1, \pi_2\}$. *When* $\delta < \frac{(1-\alpha_0)^2}{8\alpha_0}$, *there exists a range for* (λ_2, λ_3) *depending on* δ *under which any optimal solution* g^* *for (1.5) satisfies that* $\frac{1}{2} - g^*$ *has different*

signs on S_1, S_2.

Remark. The analysis is similar to the noiseless case (Theorem 2.4). Here there is an extra condition on δ, requiring the noise probability to be small.

The proof is in section 4.5.2.

Uniqueness under noise in 1-d for fixed L

Under model C5,

Theorem 4.3. *Let the density of* P_X *in* S_1, S_2 *be bounded below by* ϵ_s*, and the noise density (density outside* S*) be bounded above by* ϵ_{noise}*. Suppose* $\epsilon_s/\epsilon_{noise} > 4/(1-\lambda_3)$*, and* $\delta, \lambda_2, \lambda_3$ *satisfy the condition in Theorem 4.2. Then (1.5) has unique solution in* $\mathcal{G} = \{g : g|_{S_1} < 1/2, g|_{S_2} > 1/2\}$.

The proof is in section 4.5.3.

Remark (Proof strategy). For $i = 0, 1$, let

$$g_i(x) := \begin{cases} 0, & x \in (-\infty, x_i - \frac{1}{2L}]; \\ L(x - x_i) + \frac{1}{2}, & x \in (x_i - \frac{1}{2L}, x_i + \frac{1}{2L}); \\ 1, & x \in [x_i + \frac{1}{2L}, \infty). \end{cases}$$

Then define

$$g_t^\circ(x) := \begin{cases} 0, & x \in (-\infty, x_t - \frac{1}{2L}]; \\ L(x - x_t) + \frac{1}{2}, & x \in (x_t - \frac{1}{2L}, x_t + \frac{1}{2L}); \\ 1, & x \in [x_t + \frac{1}{2L}, \infty) \end{cases}$$

where $x_t = tx_0 + (1-t)x_1$.

The main proof strategy for uniqueness in the noise model is to replace the role of g_t in Theorem 2.5 by g_t°. Note that g_t° is not a convex combination of g_0, g_1. In the proof, we will sometimes use the function $g_t = tg_0 + (1-t)g_1$ as an intermediate function to compare. By Theorem 2.1, when $\lambda_3 < 1$, we have $I(g_t^\circ) < I(g_t)$. Therefore, even though $I(g_t) < tI(g_0) + (1-t)I(g_1)$ no longer holds because of the presence of noise (specifically, the restricted convexity argument no longer holds), it is still possible that $I(g_t^\circ) < tI(g_0) + (1-t)I(g_1)$, which would enable us to replace the role of g_t in Theorem 2.5 by g_t°. This will be the main idea in the noise case: show that $I(g_t^\circ) \leq tI(g_0) + (1-t)I(g_1)$ under some condition on the density. Uniqueness then follows: if g_0, g_1 are both assumed to be optimal solutions, then we can find a better solution g_t° unless $g_0 = g_1$.

Remark on other possible extensions

Other possible extensions may span different combinations of general penalty function, multiclass and noise model on top of the three demonstrated extensions in this chapter. We do not pursue all these extensions, but give an example of how to think about any of them, say, "uniqueness for multiple K under general penalty function in general dimension". In Chapter 2, uniqueness was a final product built upon a series of previous results. First one needs an analogue of N.C.2 such as Corollary 4.1, and a consistency theorem such as Theorem 4.1 followed by an extension of the bipartite condition C3 (this condition in Chapter 2, which comes out of the proof of a consistency theorem, is needed in the statement of the uniqueness theorem). Then one finds the analogue uniqueness statement of Theorem 2.6 (note that any uniqueness result for K-clusters should be indifferent to permutations). Technical efforts are still needed to carefully check that

each step of proof of Theorem 2.6 can be generalized seamlessly. In particular, whether key referring results such as N.C.2 and Lemma 1.1 can be utilized in a similar manner in proof as before, and whether technical lemmas such as Lemma 2.1 are still valid in a general sense.

4.4 Some variants of the formulation

This thesis has been focused on the development of the Lipschitz formulation (1.5). Other formulations under the general criterion (1.1) might be possible. For example, one that has been studied in the imaging literature is the total variation norm ([9, 20, 55]). This assumes the clustering function g lives in the function space BV (bounded variation).

Total variation formulation

We analyze a total variation variant of (1.5) in one-dimension:

$$\underset{g:\mathbb{R}^d\to[0,1],\ g\in BV}{\text{minimize}} \ E[g\wedge(1-g)]+\lambda_2\int|g'|dP+\lambda_3\max\{E[g],1-E[g]\}.$$

For any candidate solution g, consider any point x_0 such that $0<g(x_0)<1/2$, and a local variation $\tilde{g}$ of g around the neighborhood of x_0, $\tilde{g}=g-\eta$, where η is differentiable, nonnegative, and equal to zero outside the neighborhood. Suppose $\lambda_3<1$, then

$$\begin{aligned} I(g)-I(\tilde{g}) &= E[\eta]+\lambda_2\Big(\int|g'|dP-\int|(g-\eta)'|dP\pm\lambda_3\int\eta dP\Big) \\ &\geq \int\eta dP-\lambda_2\int|\eta'|dP\pm\lambda_3\int\eta dP \\ &\geq (1-\lambda_3)\int\eta dP-\lambda_2\int|\eta'|dP. \end{aligned}$$

Suppose we let $\eta = \epsilon\, e^{\frac{1-\lambda_3}{\lambda_2}x}$ (which is a solution to $(1-\lambda_3)\eta = \lambda_2|\eta'|$), and choose ϵ small enough such that $g - \eta > 0$ in a neighborhood of x_0. Then $I(g) - I(\tilde{g}) \geq 0$, so $\tilde{g}$ is variation that improves g. This offers a necessary condition for optimality: let g^* be any optimal solution, then for any point x such that $g^*(x) < 1/2$, it must be that $g^*(x) = 0$; similarly, for any point such that $g^*(x) > 1/2$, it must be that $g^*(x) = 1$. Therefore, the total variation formulation starts with a continuous formulation, but always yields 0-1 valued solutions.

The above can be contrasted with the necessary condition in the Lipschitz formulation (Theorem 1.1), the difference being that the Lipschitz solution has an additional "transition" region.

Parametric formulation: the logistic case

We give an example of a parametric formulation of (1.5) that may be more tractable:

$$\text{minimize} \quad E[g \wedge (1-g)] + \lambda_2 L(g) + \lambda_3 \max\{E[g], 1 - E[g]\}$$

subject to the parametric constraint that $g(x) = \frac{1}{1+e^{\langle \beta, x\rangle}}$, where $\beta \in \mathbb{R}^d$.

Since logistic functions are differentiable, $L(g) = \max_x ||\nabla g(x)||_2 = \max_x |\frac{e^{\langle \beta, x\rangle}}{(1+e^{\langle \beta, x\rangle})^2}| \cdot ||\beta||_2 = \max_{y>0} |\frac{y}{(1+y)^2}| \cdot ||\beta||_2 = \frac{||\beta||_2}{4}$, where the Lipschitz constant is achieved when $y = e^{\langle \beta, x\rangle} = 1$, i.e. when $\langle \beta, x\rangle = 0$ (so is achieved along the hyperplane which is the level set of the logistic function at $1/2$).

Consequence for theory and computing

The previous two examples shows that a different formulation for (1.5) may lead to different solution properties. Difference can also be expected in computing. A proposal here is to still use alternating minimization as a general principle, while within each "z-iteration step" (see Chapter 3) computation is to be done using method specific to the formulation: linear programming for Lipschitz, existing algorithms for total variation, and logistic regression for the parametric case.

It is also possible that alternative forms motivated from PDE or optimal control theory may offer better theoretical properties or are more convenient to optimize.

4.5 Proofs of chapter 4

4.5.1 Proof of Theorem 4.1

$$\begin{aligned} I_n(\tilde{g}) &= \lambda_{2,n} \max_k L(\tilde{g}_k) + \lambda_{3,n}\rho(P_n[\tilde{g}_1], \cdots, P_n[\tilde{g}_K]) \\ &= \lambda_{2,n} L_0 + \lambda_{3,n}\rho(P_n[I_{S_1}], \cdots, P_n[I_{S_K}]). \end{aligned}$$

Since $P_n[I_{S_k}] \xrightarrow{a.s.} \pi_k, k = 1, \cdots, K$, and ρ is continuous, we have $\rho(P_n[I_{S_1}], \cdots, P_n[I_{S_K}]) \xrightarrow{a.s.} \rho(\pi_1, \cdots, \pi_K) < \infty$. Therefore when $\lambda_{2,n}, \lambda_{3,n} \to 0, I_n(\tilde{g}) \to 0$.

To simplify the notation, denote $g_n(\cdot, \lambda_{2,n}, \lambda_{3,n})$ by g_n, $g_n = (g_{n,1}, \cdots, g_{n,K})$. Since $g_n \in \arg\min_g I_n(g)$, we always have $I_n(g_n) \leq I_n(\tilde{g})$.

The proof is divided into 3 parts:

1. There exists a data point x_n in S_k such that $\min_k\{1-g_{n,k}\}(x_n) < \frac{I_n(\tilde{g})}{P_n[I_{S_k}]}$, and $\lim_n P_n[\min_k\{1-g_{n,k}\}] = \lim_n P[\min_k\{1-g_{n,k}\}] = 0$.

2. For any $k = 1, \cdots, K$, $\arg\min_j(1 - g_{n,j}(x))$ is constant on S_k, when n large enough, a.s.

3. For any $k = 1, \cdots, K$, there is exactly one $k^* \in \{1, \cdots, K\}$ such that $\arg\min_j(1 - g_{n,j}(x))|S_{k^*} \equiv k$, when n large enough, a.s.

By 1,2,3, we can assume w.l.o.g that $\arg\min_j(1 - g_{n,j}(x))|S_k \equiv k$, when n large enough, a.s.. Therefore

$$
\begin{aligned}
R(g_n) &= \max_i \min_j P[|\tilde{g}_i - g_{n,j}|] \\
&= \max_i \min_j \{P[(1 - g_{n,j})I_{S_i}] + P[g_{n,j}I_{S_i^C}]\} \\
&\leq \max_i \{P[(1 - g_{n,i})I_{S_i}] + \sum_{j \neq i} P[(1 - g_{n,j})I_{S_j}]\} \\
&= \sum_{k=1}^{K} P[(1 - g_{n,k})I_{S_k}] \\
&= P[\min_k(1 - g_{n,k})] \ \ (\text{use } \arg\min_j(1 - g_{n,j}(x))|S_k \equiv k) \\
&\longrightarrow 0, a.s.,
\end{aligned}
$$

proving the claim.

1. Suppose for every $X_i \in S_k$, $\min_k\{1 - g_{n,k}\}(X_i) \geq \frac{I_n(\tilde{g})}{P_n[I_{S_k}]}$, then

$I_n(g_n) \geq P_n[\min_k\{1-g_{n,k}\}I_{S_k}] > \frac{1}{n}\sum_{i:X_i\in S_k}\frac{I_n(\tilde{g})}{P_n[I_{S_k}]} = I_n(\tilde{g})$, a contradiction as g_n is a minimizer of I_n.

Since $I_n(\tilde{g}) \to 0$, and $P_n[\min_k\{1-g_{n,k}\}] \leq I_n(g_n) \leq I_n(\tilde{g})$, we get $P_n[\min_k\{1-g_{n,k}\}] \to 0$.

The Lipschitz constant of $g_{n,k}$ is uniformly bounded almost surely. In fact, when $\lambda_{3,n} \leq C\lambda_{2,n}$, we have

$$\begin{aligned}\lambda_{2,n}L(g_{n,k}) \leq I_n(g_n) \leq I_n(\tilde{g}) &= \lambda_{2,n}L_0 + \lambda_{3,n}\rho(P_n[I_{S_1}],\cdots,P_n[I_{S_K}]) \\ &\leq \lambda_{2,n}(L_0 + C\rho(P_n[I_{S_1}],\cdots,P_n[I_{S_K}]))\end{aligned}$$

$$L(g_{n,k}) \leq L_0 + C\rho(P_n[I_{S_1}],\cdots,P_n[I_{S_K}]) \xrightarrow{a.s.} L_0 + C\rho(\pi_0).$$

By [A.2], for any $\delta > 0$, $L(\min_k\{1-g_{n,k}\}) \leq \max_k L(g_{n,k}) \leq L_0 + C(\rho(\pi_0)+\delta)$, when n large enough, a.s.. Below we fix some $\delta > 0$ and denote $L_0 + C(\rho(\pi_0)+\delta) := L_C$.

Let $\mathcal{F} = \{f : \mathcal{X} \to [0,1], L(f) \leq L_C\}$, where $\mathcal{X}$ is some bounded domain on $\mathbb{R}^d$ such that $\cup_k S_k \subset \mathcal{X}$. Then by Lemma 1.2,

$$\sup_{f\in\mathcal{F}}(P_n - P)[f] \to 0 \text{ a.s..}$$

Therefore $P_n[\min_k\{1-g_{n,k}\}] - P[\min_k\{1-g_{n,k}\}] \to 0$, a.s., it follows that $P[\min_k\{1-g_{n,k}\}] \to 0$.

2. For any $a \in [0,1], b \in [0,1], L > 0$, consider again the quantity (2.9) used when $K = 2$:

$$C_{S_k}(L,a,b) := \inf\{\int_{S_k} f dP : f(x_1) = a, f(x_2) = b \text{ for some } x_1, x_2 \in S_k, L(f) \leq L, f : \mathcal{X} \to [0,1]\},$$

which measures regularity of P on S_k. By sharpness of S_k, $C_{S_k}(L, a, b) = 0$ iff $a = b = 0$.

Suppose $\arg\min_j(1 - g_{n,j}(x))$ is not constant on S_k. By argument 1, there exists a point x^* such that $\min_k\{1 - g_{n,j}(x^*)\} < \frac{I_n(\tilde{g})}{P_n[I_{S_k}]}$. Denote $\arg\min_j(1 - g_{n,j}(x^*)) = k^*$. Then there exists another point $x^{**} \in S_k$ such that $\arg\min_j(1 - g_{n,j}(x^{**})) = k^{**} \neq k^*$. For some $\epsilon > 0$, let n be large enough that $\frac{I_n(\tilde{g})}{P_n[I_{S_k}]} < \epsilon$.

Consider $h(x) = (1 - g_{k^*})(x) - \min_{j \neq k^*}(1 - g_j)(x)$. Since $\sum_{j=1}^{K} g_j(x) \equiv 1$, $1 - g_{n,k^*}(x^*) < \epsilon$, we have $\min_{j \neq k^*}(1 - g_j)(x^*) = 1 - \max_{j \neq k^*} g_j(x^*) \geq 1 - \sum_{j \neq k^*} g_j(x^*) = g_{k^*}(x^*) > 1 - \epsilon$. Therefore

$$h(x^*) < \epsilon - (1 - \epsilon) = 2\epsilon - 1, h(x^{**}) > 0.$$

When $\epsilon < 1/2$, by continuity of $h(x)$ and connectedness of S_k, there is a point $\tilde{x} \in S_k$ such that $h(\tilde{x}) = 0$. This implies $g_{j^*}(\tilde{x}) = \max_{j \neq k^*} g_j(\tilde{x}), 1 = \sum_j g_j(\tilde{x}) \geq g_{k^*}(\tilde{x}) + \max_{j \neq k^*} g_j(\tilde{x}) = 2g_{k^*}(\tilde{x})$, so $g_{k^*}(\tilde{x}) \leq 1/2$.

Note that for the two points x^* and $\tilde{x}$ above, we have $\min_j(1 - g_j)(x^*) = (1 - g_{k^*})(x^*) < \epsilon$ and $\min_j(1 - g_j)(\tilde{x}) = (1 - g_{k^*})(\tilde{x}) \geq 1/2$. For any $\epsilon < \frac{1}{2}$, again, by continuity of $\min_j(1 - g_j)$ and connectedness of S_k, there are two points a and b that $\min_j(1 - g_j)$ is equal to ϵ and $1/2$, respectively. Then, by definition of C_{S_k}, $P[\min_k(1 - g_k)I_{S_k}] \geq C_{S_k}(L_C, \epsilon, \frac{1}{2}) > 0$, a contradiction to $P[\min_k(1 - g_k)] \to 0$.

3. Suppose there exists k^{**} such that $\arg\min_j(1 - g_{n,j}(x))|S_k \neq k^{**}$ for any $k = 1, \cdots, K$. Since we can always switch the roles of $k = 1, \cdots, k = K$ and in turn switch $g_1, \cdots, g_K$

accordingly, we can assume w.l.o.g that $k^{**} = 1$. We have

$$P_n[\arg\min_k(1 - g_{n,k})] < I_n(g_n) \leq I_n(\tilde{g}) := \epsilon_n.$$

Denote $k^* = \arg\min_k(1 - g_{n,k})(x), x \in S_k$. By argument 2, k^* is uniquely defined,

$$\begin{aligned} P_n[g_{n,1}] &= \sum_{k=1}^{K} P_n[g_{n,1} I_{S_k}] \\ &\leq \sum_{k=1}^{K} P_n[(1 - g_{n,k^*}) I_{S_k}] \\ &= P_n[\min_k(1 - g_{n,k})] < \epsilon_n. \end{aligned}$$

Therefore observe that when ϵ_n goes to 0,

$$\rho(P_n[g_n]) = \rho(P_n(g_{n,1}), \cdots, P_n(g_{n,K})) \approx \rho(0, x_2, \cdots, x_k),$$

for some $x_2, \cdots, x_k$. Choose some $\delta > 0$, let n be large enough that $|\rho(P_n[I_{S_1}], \cdots, P_n[I_{S_K}]) - \rho(\pi_0)| < \delta$ and $\rho(P_n[g_n]) = \rho(P_n(g_{n,1}), \cdots, P_n(g_{n,K})) > \min_{x_2, \cdots, x_k} \rho(0, x_2, \cdots, x_k) - \delta$ (replace by $\rho(P_n[g_n]) > M$ for large M if $\rho(0, x_2, \cdots, x_k) = \infty$). We have

$$\lambda_{2,n} L_0 + \lambda_{3,n}(\rho(\pi_0) + \delta) \geq \lambda_{2,n} L_0 + \lambda_{3,n} \rho(P_n[I_{S_1}], \cdots, P_n[I_{S_K}])$$

$$= I_n(\tilde{g}) \geq I_n(g_n) \geq \lambda_{3,n} \rho(P_n(g_n)).$$

On the other hand,

$$\begin{aligned}\lambda_{3,n}\rho(P_n(g_n)) - (\lambda_{2,n}L_0 + \lambda_{3,n}(\rho(\pi_0)+\delta)) &= \lambda_{3,n}(\rho(P_n(g_n)) - \rho(\pi_0) - \delta) - \lambda_{2,n}L_0 \\ &> \lambda_{3,n}(\min_{x_2,\cdots,x_k} \rho(0,x_2,\cdots,x_k) - 2\delta - \rho(\pi_0)) \\ &\quad - \lambda_{2,n}L_0,\end{aligned}$$

where the last line is non-negative as long as $\frac{\lambda_{3,n}}{\lambda_{2,n}} \geq \frac{L_0}{\min_{x_2,\cdots,x_k} \rho(0,x_2,\cdots,x_k) - \rho(\pi_0) - 2\delta}$, and a contradiction will follow. When $\rho(0,x_2,\cdots,x_k) = \infty$, the above becomes $\frac{\lambda_{3,n}}{\lambda_{2,n}} \geq \frac{L_0}{M - \rho(\pi_0) - \delta}$. The claim then follows by letting $\delta \to 0, M \to \infty$.

4.5.2 Proof of Theorem 4.2

Denote the constant 0 function by 0_X. The proof is done by contrasting $I(g_0)$ with $I(\tilde{g})$ and $I(0_X)$. We have

$$I(0_X) = \lambda_3,$$

$$
\begin{aligned}
I_1(\tilde{g}) &= \int \tilde{g} \wedge (1-\tilde{g}) d((1-\delta)P + \delta P_\epsilon) \\
&= (1-\delta) \int \tilde{g} \wedge (1-\tilde{g}) dP + \delta \int \tilde{g} \wedge (1-\tilde{g}) dP_\epsilon \\
&\leq 0 + \delta = \delta, \\
E(\tilde{g}) &= (1-\delta) \int \tilde{g} dP + \delta \int \tilde{g} dP_\epsilon \\
&= (1-\delta) \int_{S_2} 1 dP + \delta \int \tilde{g} dP_\epsilon \\
&\leq (1-\delta)\pi_2 + \delta, \\
1 - E(\tilde{g}) &\leq (1-\delta)\pi_1 + \delta \\
\max\{E[\tilde{g}], 1 - E[\tilde{g}]\} &\leq (1-\delta)\max\{\pi_1, \pi_2\} + \delta, \\
I(\tilde{g}) &\leq \delta + \lambda_2 L_0 + \lambda_3[(1-\delta)\alpha_0 + \delta].
\end{aligned}
$$

a. There exists a point x on S_k such that $g_0 \wedge (1-g_0)(x) < \lambda_3/\pi_k$. Otherwise $I(g_0) \geq E[g_0 \wedge (1-g_0) I_{S_k}] \geq \lambda_3/\pi_k \cdot \pi_k = I(0_X)$, a contradiction.

b. $g_0 \wedge (1-g_0)$ has bounded Lipschitz constant, thus continuous. This is because $\lambda_2 L(g_0) \leq I(g_0) \leq I(0_X) = \lambda_3$, so $L(g_0) \leq \lambda_3/\lambda_2$, $L(g_0 \wedge (1-g_0)) \leq L(g_0) \leq \lambda_3/\lambda_2 \leq C$.

c. Consider the two functions

$$
h_1(x) = \bar{C}_{S_k}(C, \frac{x}{\pi_k}, \frac{1}{2}), h_2(x) = \frac{1}{1-\delta} x, x \in [0, \frac{\pi_k}{2}],
$$

where $\bar{C}_{S_k}(C, \cdot, \cdot)$ is the normalized constant defined in (2.11). Since $h_1(0) > 0$, decreasing in

x and continuous, $h_2(0) = 0$, increasing in x, there is a point where $h_1(x) = h_2(x)$, denote that point by c_k. When $x < c_k, \bar{C}_{S_k}(C, \frac{x}{\pi_k}, \frac{1}{2}) > \frac{1}{1-\delta}x$.

Claim: When $\lambda_3 < c_k$, $\frac{1}{2} - g_0$ cannot change sign within each S_k.

Suppose $\frac{1}{2} - g_0$ changes sign on S_k, then by a. there is a continuous path in S_k such that g_0 continuously change from λ_3/π_k to $1/2$. By definition of $C_{S_k}(L, a, b)$ (2.9) and by b., we have

$$\begin{aligned} I(g_0) &> (1-\delta)\int_{S_k} g_0 \wedge (1-g_0)dP \\ &\geq (1-\delta)C_{S_k}(C, \frac{\lambda_3}{\pi_k}, \frac{1}{2}) \\ &\geq (1-\delta)\bar{C}_{S_k}(C, \frac{\lambda_3}{\pi_k}, \frac{1}{2}) \\ &> (1-\delta)\cdot\frac{1}{1-\delta}\lambda_3 = I(0_X), \end{aligned}$$

a contradiction.

d. Suppose $\frac{1}{2} - g_0$ have the same sign on S_1, S_2, assume w.l.o.g that $g_0 < \frac{1}{2}$.

$$\begin{aligned}
I(\tilde{g}) \geq I(g_0) &= (1-\delta)\int g_0 \wedge (1-g_0) dP + \delta \int g_0 \wedge (1-g_0) dP_\epsilon + \lambda_2 L(g_0) \\
&\quad + \lambda_3 \max\{E[g_0], 1 - E[g_0]\} \\
&= (1-\delta)\int g_0 dP + \delta \int g_0 dP_\epsilon - \delta(\int g_0 dP_\epsilon - \int g_0 \wedge (1-g_0) dP_\epsilon) \\
&\quad + \lambda_2 L(g_0) + \lambda_3 \max\{E[g_0], 1 - E[g_0]\} \\
&\geq \int g_0 d((1-\delta)P + \delta P_\epsilon) - \delta \int |g_0 - g_0 \wedge (1-g_0)| dP_\epsilon \\
&\quad + \lambda_2 L(g_0) + \lambda_3 (1 - E[g_0]) \\
&= E[g_0] - \delta + \lambda_2 L(g_0) + \lambda_3 (1 - E[g_0]).
\end{aligned}$$

so we have $E[g_0] < I(\tilde{g}) + \delta$ and $I(g_0) \geq \lambda_3(1 - E[g_0]) - \delta$. On one hand,

$$\delta + \lambda_2 L_0 + \lambda_3[(1-\delta)\alpha_0 + \delta] \geq I(\tilde{g}) \geq I(g_0) \geq \lambda_3(1 - E[g_0]) - \delta.$$

On the other hand,

$$\begin{aligned}
\lambda_3(1 - E[g_0]) - \delta - (\delta + \lambda_2 L_0 + \lambda_3[(1-\delta)\alpha_0 + \delta]) &= \lambda_3(1 - E[g_0] - (1-\delta)\alpha_0 - \delta) \\
&\quad - 2\delta - \lambda_2 L_0 \\
&\geq \lambda_3(1 - I(\tilde{g}) - \delta - (1-\delta)\alpha_0 - \delta) \\
&\quad - 2\delta - \lambda_2 L_0,
\end{aligned}$$

leading to a contradiction if the above expression is positive.

c. and d. together gives a range under which $g < \frac{1}{2}$ on one cluster and $g > \frac{1}{2}$ on the other: $\lambda_3 < \min\{c_1, c_2\}$, $\lambda_3/\lambda_2 \leq C$, $\frac{2\delta+\lambda_2 L_0}{1-C-2\delta-(1-\delta)\alpha_0} < \lambda_3 < \frac{C-\delta-\lambda_2 L_0}{(1-\delta)\alpha_0+\delta}$, where C is a constant satisfying $\delta < C < 1-2\delta-(1-\delta)\alpha_0$.

The last condition suggests that λ_2 and λ_3 should not be too small when the model includes noise: if λ_2 and λ_3 both go to 0, the last expression will converge to $-2\delta < 0$, and we would not have the above guarantee. Also for the range to exist, δ cannot be too large. The condition in the theorem $\delta < \frac{(1-\alpha_0)^2}{8\alpha_0}$ gives a rough estimate.

4.5.3 Proof of Theorem 4.3

Part 1. analysis of I_1

Claim: When the density in S_1, S_2 is bounded below by ϵ_s, and the noise density (density outside S) is bounded above by ϵ_{noise}, then

(1) $tE[g_0 I_{S_1}] + (1-t)E[g_1 I_{S_1}] - E[g_t^\circ I_{S_1}] \succeq \epsilon_s (x_1 - x_0)^2$,

(2) $|tE[(g_0 \wedge (1-g_0))I_{S^C}] + (1-t)E[(g_1 \wedge (1-g_1))I_{S^C}] - E[(g_t^\circ \wedge (1-g_t^\circ))I_{S^C}]| \preceq \epsilon_{\text{noise}}(x_1 - x_0)^2$, where the remaining constants depend on L and t.

Subproof of (1): Suppose $x_0 < x_1$. From the bipartition result in Theorem 4.2, we also

have that x_0 lies to the right of S_1 and x_1 to the left of S_2.

$$\begin{aligned}
tE[g_0 I_{S_1}] + (1-t)E[g_1 I_{S_1}] - E[g_t^\circ I_{S_1}] &= \int_{S_1} [t\max\{\frac{1}{2} + L(x-x_0), 0\} \\
&\quad + (1-t)\max\{\frac{1}{2} + L(x-x_1), 0\} \\
&\quad - \max\{\frac{1}{2} + L(x-x_t), 0\}]dP_X \\
&= \int_{[x_0 - \frac{1}{2L}, x_t - \frac{1}{2L}]} t[\frac{1}{2} + L(x-x_0)]dP_X \\
&\quad + \int_{[x_t - \frac{1}{2L}, x_1 - \frac{1}{2L}]} [tL(x_t - x_0) - (1-t)(\frac{1}{2} + L(x-x_t))] \\
&= t\int_{[x_0 - \frac{1}{2L}, x_t - \frac{1}{2L}]} [\frac{1}{2} + L(x-x_0)]dP_X \\
&\quad + (1-t)\int_{[x_t - \frac{1}{2L}, x_1 - \frac{1}{2L}]} (L(x_1 - x) - \frac{1}{2})dP_X.
\end{aligned}$$

We have $0 \leq \frac{1}{2} + L(x-x_0) \leq L(x_t - x_0) = (1-t)L(x_1 - x_0), \forall x \leq x_t - \frac{1}{2L}$, and $0 \leq L(x_1 - x) - \frac{1}{2} \leq L(x_1 - x_t) = tL(x_1 - x_0), \forall x \geq x_t - \frac{1}{2L}$. Note that (see Figure 4.1)

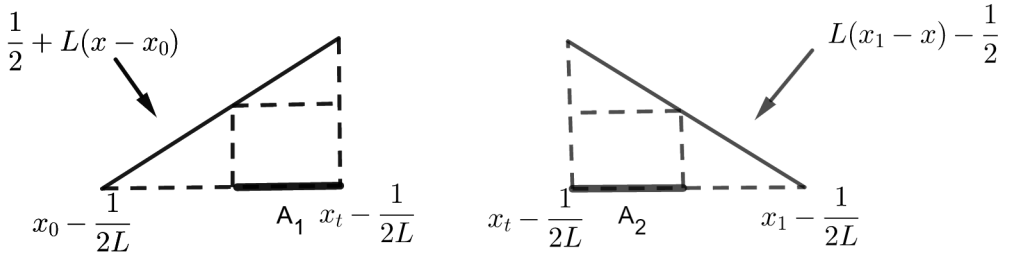


Figure 4.1: lower and upper bound for A_1 and A_2 are chosen such that the two integrands are lower bounded (by half the height of the triangle) on A_1, A_2

$$\frac{1}{2} + L(x-x_0) \geq \frac{1}{2}(1-t)L(x_1 - x_0), \forall x \in [x_0 - \frac{1}{2L} + \frac{1-t}{2}(x_1 - x_0), x_t - \frac{1}{2L}] := A_1 \subset S_1,$$

$$L(x_1 - x) - \frac{1}{2} \geq \frac{1}{2}tL(x_1 - x_0), \forall x \in [x_t - \frac{1}{2L}, x_1 - \frac{1}{2L} - \frac{t}{2}(x_1 - x_0)] := A_2 \subset S_1,$$

where $m(A_1) = \frac{1-t}{2}(x_1 - x_0), m(A_2) = \frac{t}{2}(x_1 - x_0)$.

So a lower bound for LHS of (1) is

$$\begin{aligned}
&t \cdot \frac{(1-t)L}{2}(x_1 - x_0)P(A_1) + (1-t) \cdot \frac{tL}{2}(x_1 - x_0)P(A_2) \\
&\geq \frac{t(1-t)}{2}L(x_1 - x_0)\left[\epsilon_s(P(A_1) + P(A_2))\right] \\
&= \frac{t(1-t)}{2}L(x_1 - x_0)\epsilon_s\left[\frac{1-t}{2}(x_1 - x_0) + \frac{t}{2}(x_1 - x_0)\right] \\
&\propto \epsilon_s(x_1 - x_0)^2,
\end{aligned}$$

as long as t is bounded away from 0 and 1.

Subproof of (2): Since $x_0 < x_1$, we have $S^C = R_1 \cup R_2 \cup R_3$, where

$$R_1 = \{x \in S^C : g_0(x) < \frac{1}{2}, g_1(x) < \frac{1}{2}\}, R_2 = \{x \in S^C : g_0(x) > \frac{1}{2}, g_1(x) > \frac{1}{2}\},$$

$$R_3 = \{x \in S^C : g_0(x) \geq \frac{1}{2}, g_1(x) \leq \frac{1}{2}\} = [x_0, x_1].$$

From now on, for some subset $D \subset \mathbb{R}^d$, denote $E_D[g] := E[gI_D]$. Note that all four functions $1/2 - g_0, 1/2 - g_1, 1/2 - g_t^\circ, 1/2 - g_t$ are positive on R_1, negative on R_2, and

$$E_{R_1}[g_t^\circ] \leq E_{R_1}[g_t] = tE_{R_1}[g_0] + (1-t)E_{R_1}[g_1],$$

$$E_{R_2}[g_t^\circ] \geq E_{R_2}[g_t] = tE_{R_2}[g_0] + (1-t)E_{R_2}[g_1],$$

so we have

$$E_{R_i}[g_t^\circ \wedge (1-g_t^\circ)] - tE_{R_i}[g_0 \wedge (1-g_0)] - (1-t)E_{R_i}[g_1 \wedge (1-g_1)] \leq 0,\ i=1,2.$$

Therefore,

$$\begin{aligned} &E_{S^C}[g_t^\circ \wedge (1-g_t^\circ)] - tE_{S^C}[g_0 \wedge (1-g_0)] - (1-t)E_{S^C}[g_1 \wedge (1-g_1)] \\ &\leq E_{R_3}[g_t^\circ \wedge (1-g_t^\circ)] - tE_{R_3}[g_0 \wedge (1-g_0)] - (1-t)E_{R_3}[g_1 \wedge (1-g_1)], \end{aligned}$$

that is, it suffices to control the expectation within R_3.

When restricted to R_3, g_t° is a linear combination of g_0 and g_1:

$$g_t^\circ(x) = tg_0(x) + (1-t)g_1(x), \quad x \in R_3,$$

we have

$$\begin{aligned} &tE_{R_3}[g_t^\circ \wedge (1-g_t^\circ)] - tE_{R_3}[g_0 \wedge (1-g_0)] \\ &\leq tE_{R_3}|tg_0 + (1-t)g_1 - g_0| \\ &= t(1-t)E_{R_3}|g_1 - g_0| \\ &\leq t(1-t)E_{R_3}|\frac{1}{2} + L(X-x_0) - \frac{1}{2} - L(X-x_1)| \\ &= t(1-t)LE_{R_3}|x_1 - x_0| \\ &= t(1-t)L(x_1-x_0)P_X([x_0,x_1]) \\ &\leq t(1-t)L \cdot \epsilon_{\text{noise}}(x_1-x_0)^2, \text{ (since } [x_0,x_1] \cap S = \emptyset) \end{aligned}$$

and similarly,

$$(1-t)E_{R_3}[g_t^\circ \wedge (1-g_t^\circ)] - (1-t)E_{R_3}[g_1 \wedge (1-g_1)] \leq t(1-t)L \cdot \epsilon_{\text{noise}}(x_1 - x_0)^2.$$

Therefore

$$E_{S^C}[g_t^\circ \wedge (1-g_t^\circ)] - tE_{S^C}[g_0 \wedge (1-g_0)] - (1-t)E_{S^C}[g_1 \wedge (1-g_1)] \leq 2t(1-t)L \cdot \epsilon_{\text{noise}}(x_1 - x_0)^2.$$

Part 2. analysis of I_3

The three functions g_0, g_1, g_t° have the same Lipschitz constant. It remains to control I_3 ($I_3(g) = \max\{E[g], 1 - E[g]\}$). By convexity of $I_3(g)$ in g, $I_3(g_t) \leq tI_3(g_0) + (1-t)I_3(g_1)$. Therefore

$$\begin{aligned}
tI_3(g_0) + (1-t)I_3(g_1) - I_3(g_t^\circ) &\geq I_3(g_t) - I_3(g_t^\circ) \\
&\geq -\lambda_3 |E[g_t - g_t^\circ]| \\
&= -\lambda_3 |E_{S_1}[g_t - g_t^\circ] - E_{S_2}[g_t^\circ - g_t]| \\
&= -\lambda_3 (E_{S_1}[g_t - g_t^\circ] + E_{S_2}[g_t^\circ - g_t]) \\
&\quad (g_t \geq g_t^\circ \text{ on } S_1, g_t \leq g_t^\circ \text{ on } S_2) \\
&= -\lambda_3 (tE_{S_1}[g_0] + (1-t)E_{S_1}[g_1] - E_{S_1}[g_t^\circ] \\
&\quad + E_{S_2}[g_t^\circ] - tE_{S_2}[g_0] - (1-t)E_{S_2}[g_1]).
\end{aligned}$$

Part 3. combining I_1 and I_3

The last expression appears also in the analysis of I_1, so we can combine terms when

analyzing $I(g)$:

$$
\begin{aligned}
tI(g_0) + (1-t)I(g_1) - I(g_t^\circ) &= tI_1(g_0) + (1-t)I_1(g_1) - I_1(g_t^\circ) \\
&\quad + tI_3(g_0) + (1-t)I_3(g_1) - I_3(g_t^\circ) \\
&\geq tE_{S_1}[g_0] + (1-t)E_{S_1}[g_1] - E_{S_1}[g_t^\circ] \\
&\quad + tE_{S_2}[1-g_0] + (1-t)E_{S_2}[1-g_1] - E_{S_2}[1-g_t^\circ] \\
&\quad + tE_{S^C}[g_0 \wedge (1-g_0)] + (1-t)E_{S^C}[g_0 \wedge (1-g_0)] \\
&\quad - E_{S^C}[g_t^\circ \wedge (1-g_t^\circ)] \\
&\quad - \lambda_3(tE_{S_1}[g_0] + (1-t)E_{S_1}[g_1] - E_{S_1}[g_t^\circ] \\
&\quad + E_{S_2}[g_t^\circ] - tE_{S_2}[g_0] - (1-t)E_{S_2}[g_1]) \\
&= (1-\lambda_3)(tE_{S_1}[g_0] + (1-t)E_{S_1}[g_1] - E_{S_1}[g_t^\circ) \\
&\quad + (1-\lambda_3)(E_{S_2}[g_t^\circ] - tE_{S_2}[g_0] - (1-t)E_{S_2}[g_1]) \\
&\quad + tE_{S^C}[g_0 \wedge (1-g_0)] + (1-t)E_{S^C}[g_0 \wedge (1-g_0)] \\
&\quad - E_{S^C}[g_t^\circ \wedge (1-g_t^\circ)].
\end{aligned}
$$

Let $t = 1/2$, we have

(1): $tE_{S_1}[g_0] + (1-t)E_{S_1}[g_1] - E_{S_1}[g_t^\circ] \geq \frac{1}{16}L \cdot \epsilon_s(x_1 - x_0)^2$, and similarly,

$E_{S_2}[g_t^\circ] - tE_{S_2}[g_0] - (1-t)E_{S_2}[g_1] \geq \frac{1}{16}L \cdot \epsilon_s(x_1 - x_0)^2$;

(2): $E_{S^C}[g_t^\circ \wedge (1-g_t^\circ)] - tE_{S^C}[g_0 \wedge (1-g_0)] - (1-t)E_{S^C}[g_1 \wedge (1-g_1)] \leq \frac{1}{2}L \cdot \epsilon_{\text{noise}}(x_1 - x_0)^2$,

so $tI(g_0) + (1-t)I(g_1) - I(g_t^\circ) \geq \frac{1}{8}(1-\lambda_3)L\epsilon_s(x_1 - x_0)^2 - \frac{1}{2}L\epsilon_{\text{noise}}(x_1 - x_0)^2$, and

$I(g_t^\circ) < tI(g_0) + (1-t)I(g_1)$ as long as $\epsilon_s/\epsilon_{\text{noise}} > 4/(1-\lambda_3)$.

Chapter 5: Conclusion and future work

5.1 Contribution of this work

Contributions made in this thesis beyond [65] (Lipschitz classifier) include: consideration of the population (variational) problem and (statistical) consistency issue (Theorem 1.3), two optimality conditions (Theorem 1.1, 2.1), a model (C1) which the method adapts to (Theorem 2.3), the uniqueness problem (Theorem 2.6), and further computational developments on top of the original linear program in [65] (Chapter 3). Lastly, our approach can be seen as a step forward from classification to clustering.

Contribution of the thesis to the clustering literature include the following. This work proposes a general criterion where clustering is viewed as the easiest classification problem (1.1). The corresponding data problem has natural consistency property. The Lipschitz formulation offers a novel approach for continuous clustering with good mathematical properties, but is also different from traditional model-based clustering.

Overall, we hope this work can offer necessary preparations for further mathematical analysis and algorithmic development.

5.2 Future directions

We describe some major future directions for data and ideal problem.

Ideal variational problem

By Theorem 2.1, the remaining question in finding optimal g is now a geometric variational problem: how to find an optimal surface U for a fixed Lipschitz constant L. The first step may be to use the fact that almost every level set of a Lipschitz function is $(d-1)$-rectifiable [2], in company with Corollary 2.2. There are also results specifically applied to distance functions [30] that may extend to all level sets. The next direction is to establish regularity result ([37, 52]) of U possibly under addition assumptions such as convexity of clusters.

An equally interesting question is to characterize mathematical (regularity) properties of $U_n = \{g_n = 1/2\}$ in the data problem. This may also help to determine optimal U in the ideal problem, either numerically or theoretically (asymptotically).

Algorithm

Convergence guarantee for alternating minimization An important problem about algorithmic guarantee is to explain the empirically small number of z-iterations (see, e.g., Table 3.2) in Algorithm 1, and to what extent a global optimum can be found, under certain ideal models.

There are existing general and problem-specific results for alternating minimization along this line. We refer to Theorem 4.3 and 5.5 in the review paper [25] for general results in the machine learning literature, and [4] for such result in the statistical literature, in particular for

the EM algorithm (which is a special case of the alternating minimization principle). These results have the common flavor: first, certain initialization procedures are proposed to ensure the algorithm starts in a "basin of attraction" (often come with convexity of objective function within this region), within which the global optimum can then be approached at a linear rate. This in turn implies $\log(\frac{1}{\epsilon})$ iterations are sufficient to solve the optimization problem to ϵ accuracy.

Core set/sparse representation Lipschitz extension from n data points may be determined by a much smaller set of data points. Specifically, in (3.4), we call $[s] \subset \{1, \cdots, n\}$ a "core set" for g_n if for any x,

$$g_n(x) = \frac{1}{2} \min_{i \in [s]} \{g_n(x_i) + Ld(x, x_i)\} + \frac{1}{2} \max_{i \in [s]} \{g_n(x_i) - Ld(x, x_i)\}.$$

In the 1-d case in Figure 3.1 when g_n is piecewise linear, $[s]$ can be reduced to only the two turning points of g_n. This reduction can be important for prediction at future points. For coreset in K-means clustering, see [22].

Distributional result

A (functional) limit theorem for g_n would provide a more complete understanding of this approach. For example, it will further justify the subsampling and confidence band calculation in Chapter 3. A central limit theorem for K-means clustering was proved by Pollard [44]. However, the technique relies on the parametric and differentiable nature of K-means objective function, while (1.5) is both nonparametric and nonsmooth.

Real data illustration

More real data illustrations of the computational developments in Chapter 3 will be helpful, such as using the visualization and diagnostic ideas in section 3.7. For implementation on large-scale datasets, more progress would be required on (1) making the algorithm more scalable; (2) have better understanding of the algorithm, such as the earlier discussion in this section.

Some general questions in clustering

Clusterable models As discussed in section 1.1, a theory that applies to both traditional model-based statistical methods and modern machine learning approaches for clustering would be desirable. We propose that future work should begin with clarifying the concept of "clusterable models", followed by theoretical study of clusterable models on one hand, and design of efficient clustering algorithms under corresponding models on the other hand. This will also set up the basis for a tuning parameter selection theory (see section 3.8). The meaning of "clusterable" can be more general than an identifiable statistical mixture model. Modeling considerations are important for other extensions of the clustering problem as well, such as bi-clustering [17].

High-dimensional clustering problem Intuitively speaking, high dimensional classification problem is relatively simple because in higher dimension it becomes easier to find hyperplanes to separate the classes. The remaining question in high dimensional classification is then to find an optimal hyperplane. For unlabeled data, however, the abundance of hyperplanes becomes an obstacle: hyperplane cuts are very arbitrary, so that clustering is not always possible or meaningful. Thus the more important question for high dimensional clustering is to study statistical limits under various distributional assumptions. We refer to [8, 27] for the Gaussian case.

Recent works [10, 35] consider high dimensional regimes where both the number of clusters K and number of variables are comparable to sample size n. Comprehensive study on canonical models like Gaussian mixture models will remain important for these problems.

Appendix A:

A.1 Analysis results

A.1.1 Non-smooth analysis

Results in this section can be found in Chapter 10 of [13], which is a generalization of differential calculus for smooth functions, and subdifferential calculus for convex functions. The function class considered is locally Lipschitz functions. In the main thesis, we will mostly work with (globally) Lipschitz functions, which are locally Lipschitz everywhere. This includes, in particular, distance functions to a closed set.

X denotes a Banach space. Let $f : X \to \mathbb{R}$ be Lipschitz of rank K near a given point $x \in X$, that is, for some $\epsilon > 0$, we have

$$|f(y) - f(z)| \leq K||y - z||, \forall y, z \in B(x, \epsilon).$$

That is, f is Lipschitz on some (sufficiently small) neighborhood around x, which implies continuity at x. We will mostly work with Euclidean space $X = \mathbb{R}^n$.

Definition A.1.1. The **generalized directional derivative** of f at x in the direction v, denoted

$f^\circ(x;v)$, is defined as:

$$f^\circ(x;v) = \limsup_{y \to x, t \downarrow 0} \frac{f(y+tv) - f(y)}{t},$$

where $y \in X, t > 0$.

E.g., the generalized directional derivative of a distance function d_S at x in the direction v is denoted by $d_S^\circ(x;v)$.

Definition A.1.2. The **generalized gradient** of the function f at x, denoted $\partial_C f(x)$, is the unique nonempty weak* compact convex subset of X^* whose support function is $f^\circ(x;\cdot)$, that is,

$$\zeta \in \partial_C f(x) \iff f^\circ(x;v) \geq \langle \zeta, v \rangle \; \forall v \in X,$$

$$f^\circ(x;v) = \max\{\langle \zeta, v \rangle : \zeta \in \partial_C f(x)\} \; \forall v \in X.$$

We will mostly work with the case when $X = X^* = \mathbb{R}^n$.

The concepts above are indeed a generalization of gradients in the smooth and convex case, as can be seen from the following:

Theorem A.1. *If f is continuously differentiable near x, then $\partial_C f(x) = \{f'(x)\}$. If f is convex and lower semi-continuous, and if $x \in int\ dom f$, then $\partial_C f(x) = \partial f(x)$, the subgradient of f at x.*

As in calculus, we usually work with generalized gradients through their properties, rather than from the definition.

Calculus of generalized gradients

Theorem A.2 (Sum rule)**.** *Let f and g be Lipschitz near x. Then*

$$\partial_C(f+g)(x) \subset \partial_C f(x) + \partial_C g(x).$$

If g is convex, then

$$\partial_C(f+g)(x) = \partial_C f(x) + \partial g(x).$$

Theorem A.3 (Mean value theorem)**.** *Let x and y belong to X, and suppose that f is Lipschitz on a neighborhood of the line segment $[x, y]$. Then there exists a point z in (x, y) such that*

$$f(y) - f(x) \in \langle \partial_C f(z), y - x \rangle.$$

Definition A.1.3 (convex envelope)**.** Let S be a subset of X. The **convex envelope** of S, denoted $co\ S$, is the smallest convex subset of X containing S.

The convex envelope has the following charaterization:

Lemma A.1. $co\ S = \{\sum_{i=1}^{m} t_i x_i : m \geq 1, x_i \in S, t_i \geq 0, \sum_{i=1}^{m} t_i = 1\}$.

Theorem A.4 (Gradient formula)**.** *Let $x \in \mathbb{R}^n$, and let $f : \mathbb{R}^n \to \mathbb{R}$ be Lipschitz near x. Let E_f be the set of points at which f fails to be differentiable. Then*

$$\partial_C f(x) = co\{\lim_{n\to\infty} \nabla f(x_n) : x_n \to x, x_n \notin E_f\}.$$

Remark. One property of $\partial_C f$ is that it is "blind to sets of measure 0" [13]: this means that

any measure 0 set can be ignored in the construction of all limiting sequences, without changing $\partial_C f(x)$.

A.1.2 Lipschitz functions

Many results in this section can be found in [69] and [13]. In the finishing stage of this work, we discovered that Lipschitz analysis plays a fundamental role in geometric measure theory, two good references are [37] and [52].

Lemma A.2. *For any function g, $L(g \wedge (1-g)) \leq L(g)$. In general, for K functions $g_1, \cdots, g_K$, $L(g_1 \wedge \cdots \wedge g_K) \leq \max_k L(g_k)$.*

Lemma A.3. *$L(g)$ is convex in g.*

Proof. For any two functions g_1, g_2 and any $0 \leq t \leq 1$,

$$\begin{aligned}
L(tg_1 + (1-t)g_2) &= \max_{x,y} \frac{|tg_1(x) + (1-t)g_2(x) - tg_1(y) - (1-t)g_2(y)|}{d(x,y)} \\
&\leq \max_{x,y} \frac{t|g_1(x) - g_1(y)| + (1-t)|g_2(x) - g_2(y)|}{d(x,y)} \\
&= \max_{x,y} \left\{ t\frac{|g_1(x) - g_1(y)|}{d(x,y)} + (1-t)\frac{|g_2(x) - g_2(y)|}{d(x,y)} \right\} \\
&\leq t \max_{x,y} \frac{|g_1(x) - g_1(y)|}{d(x,y)} + (1-t) \max_{x,y} \frac{|g_2(x) - g_2(y)|}{d(x,y)} \\
&= tL(g_1) + (1-t)L(g_2)
\end{aligned}$$

□

Theorem A.5 (Kirszbraun Theorem, Lipschitz extension)**.** *Let $E \subset \mathbb{R}^n$, $f : E \to \mathbb{R}^m$ such that $Lip(f, E) = L < \infty$. Then there exists $\bar{f} : \mathbb{R}^n \to \mathbb{R}^m$, $\bar{f}$ is L-Lipschitz, $\bar{f}|_E = f$.*

There are many constructions for Lipschitz extension. For example,

($m = 1$) Mcshane extension

$$\bar{f}(y) := \inf_{x \in E} \{f(x) + L|y - x|\}$$

$$\bar{\bar{f}}(y) := \sup_{x \in E} \{f(x) - L|y - x|\}$$

These two constructions generalize to metric space:

$$\bar{f}(y) := \inf_{x \in E} \{f(x) + Ld(x, y)\} \tag{A.1}$$

$$\bar{\bar{f}}(y) := \sup_{x \in E} \{f(x) - Ld(x, y)\} \tag{A.2}$$

Therefore

Lemma A.4. *Let $(\mathcal{X}, d)$ be a metric space and $E \subset \mathcal{X}$, $f : E \to \mathbb{R}$ is Lipschitz with Lipschitz constant L. Then there exists $\bar{f} : \mathcal{X} \to \mathbb{R}$, $\bar{f}$ is L-Lipschitz, $\bar{f}|_E = f$.*

The following lemma gives a family of constructions for Lipschitz extension from finite data points, which is a consequence of the two constructions (A.1) and (A.2), if we let $E = \{x_1, \cdots, x_n\}$.

Lemma A.5 (Lemma 7 from [65])**.** *Given a function f defined on a finite subset $x_1, \cdots, x_n$ of $\mathcal{X}$, there exists a function $\bar{f}$ which coincides with f on $x_1, \cdots, x_n$, is defined on the whole space $\mathcal{X}$, and has the same Lipschitz constant as f. Additionally, it is possible to explicitly construct $\bar{f}$*

in the form

$$\bar{f}(x) = \alpha \min_{i=1,\cdots,n} (f(x_i) + L(f)d(x, x_i)) + (1-\alpha) \max_{i=1,\cdots,n} (f(x_i) - L(f)d(x, x_i)),$$

for any $\alpha \in [0,1]$, *with* $L(f) = \max_{i,j=1,\cdots,n} (f(x_i) - f(x_j))/d(x_i, x_j)$.

Corollary A.1. *When* $\alpha = \frac{1}{2}$, *the construction in the above lemma also implies* $\min_{i=1,\cdots,n} f(x_i) \leq \bar{f}(x) \leq \max_{i=1,\cdots,n} f(x_i)$.

Proof. Since

$$\bar{f}(x) \leq \frac{1}{2}(f(x_{i_0}) + L(f)d(x, x_{i_0})) + \frac{1}{2}(f(x_{i_0}) - L(f)d(x, x_{i_0})) = f(x_{i_0}) \leq \max_{i=1,\cdots,n} f(x_i),$$

where i_0 is the index for which the max part is maximized: $f(x_{i_0}) - L(f)d(x, x_{i_0}) = \max_{i=1,\cdots,n} (f(x_i) - L(f)d(x, x_i))$. The other side of the inequality can be similarly shown. □

Theorem A.6 (Rademacher's theorem)**.** *Let* $f : \mathbb{R}^n \to \mathbb{R}$ *be Lipschitz. Then* f *is differentiable almost everywhere, and its gradient vector field is essentially bounded with* $|| \, |\nabla f| \, ||_\infty = L(f)$.

Corollary A.2 (integration by part)**.** *Let* $f : \mathbb{R}^n \to \mathbb{R}^m$ *be Lipschitz, then*

$$\int_{\mathbb{R}^n} \langle \nabla f, \phi \rangle = - \int_{\mathbb{R}^n} \langle f, \ \nabla \phi \rangle, \forall \phi \in C_c^1(\mathbb{R}^n),$$

where $C_c^1(\mathbb{R}^n)$ *denotes the class of continuously differentiable compactly supported functions on* $\mathbb{R}^n$. *Since integration by part always figures the right dimension,* ∇ *is understood as either* $gradient$ *or* $divergence$ *according to the dimension.*

The remaining lemmas in this section relate the global Lipschitz constant with local ones by norm of (classical or generalized) gradients.

Lemma A.6. *Suppose g is differentiable, then $L(g) = \sup_x ||\nabla g(x)||_2$.*

Lemma A.7. *Suppose g is Lipschitz, then $L(g) = \sup_{\zeta \in \partial_C g} ||\zeta||$ ($\partial_C g := \{\zeta : \zeta \in \partial_C g(x)$ for some $x\}$).*

Lemma A.8. *Suppose g is Lipschitz, then $\sup_{\zeta \in \partial_C g(x)} ||\zeta|| = \sup_{\zeta = \lim_n \nabla g(x_n), x_n \to x} ||\zeta||$ (the supremum is determined by differentiable points).*

The following is a consequence of the previous two lemmas:

Lemma A.9. *Suppose g is Lipschitz, then $L(g) = \sup_{x \text{ differentiable}} ||\nabla g(x)||$.*

Proof of Lemma A.2. It suffices to show that $|\min\{g(y), 1-g(y)\} - \min\{g(x), 1-g(x)\}|$ is bounded by $|g(y) - g(x)|$. Note that in the case when $g(y) \le 1 - g(y), g(x) \le 1 - g(x)$ or $g(y) > 1 - g(y), g(x) > 1 - g(x)$, we have $|\min\{g(y), 1-g(y)\} - \min\{g(x), 1-g(x)\}| = |g(y) - g(x)|$. When $g(y) \le 1 - g(y), g(x) > 1 - g(x)$,

$$g(y) - g(x) \le g(y) - (1 - g(x)) \le 1 - g(y) - (1 - g(x)) = g(x) - g(y),$$

$$|\min\{g(y), 1-g(y)\} - \min\{g(x), 1-g(x)\}| = |g(y) - (1-g(x))| \le |g(y) - g(x)|.$$

The other case $g(y) > 1 - g(y), g(x) \le 1 - g(x)$ is similar. Therefore $L(g \wedge (1-g)) \le L(g)$.

In general, $|\min\{g_1(y), \cdots, g_K(y)\} - \min\{g_1(x), \cdots, g_K(x)\}| \le \max_k |g_k(y) - g_k(x)|$. The lemma thus follows. □

Proof of Lemma A.6. By Taylor's theorem,

$$\begin{aligned} g(y) - g(x) &= \nabla g(x)(y-x) + o(||y-x||_2), y \to x \\ &\leq ||\nabla g(x)||_2 ||y-x||_2 + o(||y-x||), \\ \limsup_{y \to x} \frac{g(y)-g(x)}{||y-x||_2} &\leq ||\nabla g(x)||_2. \end{aligned}$$

Let $y = x + h\nabla g(x)$,

$$\begin{aligned} g(y) - g(x) &= \nabla g(x)(h\nabla g(x)) + o(h||\nabla g(x)||_2), h \to 0 \\ &= h||\nabla g(x)||_2^2 + o(h||\nabla g(x)||_2), \\ \lim_{h \to 0} \frac{g(x + h\nabla g(x)) - g(x)}{||h\nabla g(x)||_2} &= ||\nabla g(x)||_2. \end{aligned}$$

Therefore

$$\begin{aligned} \limsup_{y \to x} \frac{g(y)-g(x)}{||y-x||_2} &= ||\nabla g(x)||_2, \\ L(g) = \sup_{x \neq y} \frac{g(y)-g(x)}{||y-x||_2} &\geq \sup_x \limsup_{y \to x} \frac{g(y)-g(x)}{||y-x||_2} \\ &= \sup_x ||\nabla g(x)||_2. \end{aligned}$$

On the other hand, by mean value theorem,

$$\begin{aligned} g(y) - g(x) &= \nabla g(\xi)(y-x), \xi = x + \theta(y-x), \theta \in (0,1), \\ &\leq ||\nabla g(\xi)||_2 ||y-x||_2, \\ \frac{g(y)-g(x)}{||y-x||_2} &\leq ||\nabla g(\xi)||_2 \leq \sup_x ||\nabla g(x)||_2, \\ L(g) &\leq \sup_x ||\nabla g(x)||_2, \end{aligned}$$

so $L(g) = \sup_x ||\nabla g(x)||_2$. □

Proof of Lemma A.7. By definition, the generalized directional derivative of g at x (A.1.1) is

$$g^\circ(x,v) = \limsup_{t \downarrow 0} \frac{g(x+tv) - g(x)}{t},$$

so for any direction v, we can extract a sequence $t_n \downarrow 0$ such that $\lim_n \frac{g(x+t_n v)-g(x)}{t_n} = g^\circ(x,v)$.

The generalized gradient $\partial_C g(x)$ (A.1.2) is characterized by directional derivatives:

$$g^\circ(x,v) = \max\{\langle \zeta, v\rangle, \zeta \in \partial_C g(x)\}.$$

Choose $v = \zeta$ for some $\zeta \in \partial_C g(x)$, we have

$$\lim_n \frac{g(x+t_n\zeta) - g(x)}{t_n} = g^\circ(x,\zeta) \geq \langle \zeta, \zeta\rangle = ||\zeta||^2,$$

$$\begin{aligned}\lim_n \frac{g(x+t_n\zeta)-g(x)}{||t_n\zeta||} &= \lim_n \frac{g(x+t_n\zeta)-g(x)}{t_n}\cdot\frac{1}{||\zeta||}\\ &= g^\circ(x,\zeta)\cdot\frac{1}{||\zeta||}\\ &\geq ||\zeta||^2\cdot\frac{1}{||\zeta||}=||\zeta||.\end{aligned}$$

The local Lipschitz constant at x

$$\limsup_{||y-x||\to 0}\frac{g(y)-g(x)}{||y-x||}\geq\lim_n\frac{g(x+t_n\zeta)-g(x)}{||t_n\zeta||}\geq||\zeta||,\forall\zeta\in\partial_C g(x),$$

$$\limsup_{||y-x||\to 0}\frac{g(y)-g(x)}{||y-x||}\geq\sup_{\zeta\in\partial_C g(x)}||\zeta||.$$

On one hand, the global Lipschitz constant dominates the local ones,

$$L(g)=\sup_{x\neq y}\frac{g(y)-g(x)}{||y-x||}\geq\sup_x\limsup_{||y-x||\to 0}\frac{g(y)-g(x)}{||y-x||}\geq\sup_x\sup_{\zeta\in\partial_C g(x)}||\zeta||=\sup_{\zeta\in\partial_C g}||\zeta||;$$

On the other hand, by mean value theorem [A.3],

$$g(y)-g(x)\in\langle\partial_C g(z),y-x\rangle,$$

for some point z on the line segment $[x,y]$. Therefore

$$\begin{aligned}g(y)-g(x)&\leq\sup_{\zeta\in\partial_C g(z)}||\zeta||||y-x||\\ &\leq\sup_{\zeta\in\partial_C g}||\zeta||||y-x||,\\ \frac{g(y)-g(x)}{||y-x||}&\leq\sup_{\zeta\in\partial_C g}||\zeta||,\end{aligned}$$

so $L(g) = \sup_{x \neq y} \frac{g(y)-g(x)}{||y-x||} \leq \sup_{\zeta \in \partial_C g} ||\zeta||$. We conclude $L(g) = \sup_{\zeta \in \partial_C g} ||\zeta||$. □

Proof of Lemma A.8. Denote $E_x = \{\zeta : \zeta = \lim_n \nabla g(x_n), x_n \to x\}$. By gradient formula [A.4] and [A.1], any $\zeta \in \partial_C g(x)$ can be written as $\zeta = \sum_{i=1}^{m} c_i \zeta_i$, where $0 \leq c_i \leq 1, \sum_{i=1}^{m} c_i = 1, \zeta_1, \cdots, \zeta_m \in E_x$. It follows that

$$||\zeta|| \leq \sum_{i=1}^{m} c_m ||\zeta_i|| \leq (\sum_{i=1}^{m} c_i) \max_{i=1,\cdots,m} ||\zeta_i|| = \max_{i=1,\cdots,m} ||\zeta_i|| \leq \sup_{\zeta \in E_x} ||\zeta||.$$

□

A.1.3 Distance functions

Definition A.1.4. S denotes a nonempty closed subset of a Banach space X. The **distance function** associated with the set S is defined by

$$d_S(x) = \inf_{y \in S} ||y - x||,$$

which is globally Lipschitz of rank 1.

Lemma A.10 (Exercise 10.40, [13])**.** *S is a nonempty closed subset of $\mathbb{R}^n$, and $proj_S(x)$ denotes the set of points $u \in S$ satisfying $d_S(x) = ||x - u||$. Then*

- *d_S is Lipschitz and $L(d_S) = 1$.*
- *d_S is differentiable a.e.. For any $x \notin S$ such that $d'_S(x)$ exists, $proj_S(x)$ is unique, and*

$$\nabla d_S(x) = \frac{x - y}{||x - y||}, \text{ where } y = proj_S(x).$$

Therefore, $||\nabla d_S(x)|| = 1$ *a.e. in* S^C *(and equality holds for any* x *where* d_S *is differentiable).*

Lemma A.11. *Let* S *be a closed set. Suppose* d_S *is differentiable at* $x \notin S$*, and* $y = proj_S(x)$*, then* d_S *is differentiable on the open line segment* xy *(but may not be differentiable on the ray that leaves* x*). All points on this line segment share the same gradient, equal to* $\nabla d_S(x) = \frac{x-y}{||x-y||}$*. Suppose the line segment* xy *intersects* $\partial B_r(S)$ *at point* z *for some* $0 < r < d(x, y)$*, then* $z = proj_{B_r(S)}(x)$ *(*z *is also the closest point from* x *to* $B_r(S)$*). Therefore,* $\nabla d_{B_r(S)}(x) = \frac{x-z}{||x-z||} = \frac{x-y}{||x-y||} = \nabla d_S(x)$.

Proof of Lemma A.10. Let $y \in \text{proj}_S(x)$, which implies $\inf_{u\in S} ||x-u|| = ||x-y||$.

If $d'_S(x)$ exists, then for any $v \in \mathbb{R}^n$

$$\begin{aligned}
d'_S(x;v) &= \lim_{t\downarrow 0} \frac{d_S(x+tv) - d_S(x)}{t} \\
&= \lim_{t\downarrow 0} \frac{\inf_{u\in S} ||x+tv-u|| - \inf_{u\in S} ||x-u||}{t} \\
&\leq \lim_{t\downarrow 0} \frac{||x+tv-y|| - ||x-y||}{t} \\
&= \frac{d}{dt}||x+tv-y||\Big|_{t=0} \\
&= \frac{2\langle x-y, v\rangle}{2||x+tv-y||}\Big|_{t=0} \\
&= \langle \frac{x-y}{||x-y||}, v\rangle.
\end{aligned}$$

By property of gradient, $d'_S(x;v) \geq \langle \nabla d_S(x), v\rangle, \forall v \in \mathbb{R}^n$, we have

$$\langle \frac{x-y}{||x-y||}, v\rangle \geq \langle \nabla d_S(x), v\rangle, \forall v \in \mathbb{R}^n.$$

Replace v by $-v$, we deduce $\langle \frac{x-y}{||x-y||}, v\rangle \leq \langle \nabla d_S(x), v\rangle$, and so

$$\langle \frac{x-y}{||x-y||}, v\rangle = \langle \nabla d_S(x), v\rangle, \forall v \in \mathbb{R}^n.$$

It follows that $\nabla d_S(x) = \frac{x-y}{||x-y||}$.

From this we also see that suppose $d'_S(x)$ exists, then y must be the unique point in $\text{proj}_S(x)$ (otherwise gradient will be non-unique). □

Proof of Lemma A.11. By Lemma A.10, it suffices for the second sentence of the assertion to show that for any x' on the line segment xy (not including y), $y = \text{proj}_S(x')$. For any $y' \in S$,

$$\begin{aligned} d(x', y') &\geq d(x, y') - d(x, x') \\ &\geq d(x, y) - d(x, x') \\ &= d(x', y), \end{aligned}$$

where the second inequality follows from $y = \text{proj}_S(x)$, and equality is achieved only when $y' = y$.

For the second part, suppose $d(x, z') \leq d(x, z)$ for some $z' \in B_r(S)$. Let $y' \in \text{proj}_S(z')$

(so $d(z', y') \leq r$), and look at the triangle with endpoints x, z', y'. We have

$$\begin{aligned} d(x, y') &\leq d(x, z') + d(z', y') \\ &\leq d(x, z) + r \\ &= d(x, z) + d(z, y) \\ &= d(x, y). \end{aligned}$$

Since $y = \text{proj}_S(x)$, we have $d(x, y) \leq d(x, y')$, " = " iff $y' = y$. Thus all the inequalities should be equality, and $y' = y$, z' is on the line segment xy' (now the same as line segment xy) to achieve the first equality. Therefore, $z' = z$. □

A.1.4 Other

Theorem A.7 (Fubini's theorem)**.** *Let $(X, M, \mu), (Y, N, \nu)$ be measure spaces. If $E \subset X \times Y$, for $x \in X, y \in Y$ define the x-section E_x and $y-$section E^y of E by*

$$E_x = \{y \in Y : (x, y) \in E\}, E^y = \{x \in X : (x, y) \in E\}.$$

If $E \in M \times N$ and $\mu \times \nu(E) = 0$, then $\nu(E_x) = \mu(E^y) = 0$ for a.e. x and y.

The reference is [19].

A.2 Statistical results

Theorem A.8. *Let U_n be a U-statistic with degree d, i.e., $U_n = \frac{1}{\binom{n}{d}} \sum_c f(\xi_{i_1}, \cdots, \xi_{i_d})$ where $\xi_1, \cdots, \xi_n$ are i.i.d random variables, f is permutation symmetric in its arguments, and c denotes all combinations of d distinct elements $\{i_1, \cdots, i_d\}$ from $\{1, \cdots, n\}$. Suppose $a \leq \xi_i \leq b, i = 1, \cdots, n$, then*

$$P(U_n - E[U_n] \geq t) \leq \exp(-2\lfloor n/d \rfloor t^2/(b-a)^2).$$

The reference is (2.4) in [41] or (4.3) in [26].

Lemma A.12. *Suppose $X_n \xrightarrow{p} \mu$, and X_n's are bounded uniformly, then $E[X_n] \xrightarrow{n\to\infty} \mu$.*

Proof. We have $|X_n| \leq C$ for some constant C. For any $\epsilon > 0, P(|X_n - \mu| > \epsilon) \to 0$.

$$\begin{aligned} EX_n &= E[X_n I(|X_n - \mu| > \epsilon)] + E[X_n I(|X_n - \mu| \leq \epsilon)] \\ &\leq CP(|X_n - \mu| > \epsilon) + (\mu + \epsilon)P(|X_n - \mu| \leq \epsilon) \\ &\xrightarrow{n\to\infty} \mu + \epsilon, \end{aligned}$$

the inequality on the other side can be shown similarly, so

$$\mu - \epsilon \leq \liminf EX_n \leq \limsup EX_n \leq \mu + \epsilon.$$

Let $\epsilon \to 0$ to obtain $E[X_n] \to \mu$. □

In fact, this is a trivial consequence of the bounded convergence theorem (or dominated convergence theorem) and the fact that convergence in probability implies for every subsequence

there is a further subsequence along which a.s. convergence holds (and that common limits for the subsequences implies the original sequence has the same limit). See [16] Theorem 2.3.4.

A.2.1 Empirical process theory

In this section we will follow the notation in chapter 19 of [60].

Let $X_1, \cdots, X_n$ be a random sample from a probability distribution P on space $\mathcal{X}$. Let P_n denote the empirical measure on $X_1, \cdots, X_n$. Given a measurable function $f : \mathcal{X} \to \mathbb{R}$, write

$$Pf = \int f dP, \quad P_n f = \frac{1}{n} \sum_{i=1}^{n} f(X_i).$$

Let $\mathcal{F}$ be a class of functions.

Definition A.2.1 (bracketing number)**.** Given two functions l and u, the bracket $[l, u]$ is the set of all functions f with $l \leq f \leq u$. An ϵ-bracket in $||\cdot||$ is a bracket $[l, u]$ with $||u - l|| < \epsilon$. The **bracketing number** $N_{[]}(\epsilon, \mathcal{F}, ||\cdot||)$ is the minimum number of ϵ-brackets needed to cover $\mathcal{F}$.

A function class that satisfies uniform law of large numbers under distribution P is called P-Glivenko-Cantelli:

Definition A.2.2. A class $\mathcal{F}$ of measurable functions $f : \mathcal{X} \to \mathbb{R}$ is called **P-Glivenko-Cantelli** if

$$||P_n f - Pf||_{\mathcal{F}} = \sup_{f \in \mathcal{F}} |P_n f - Pf| \overset{a.s.}{\to} 0.$$

The following gives a user-friendly condition for a function class to be P-Glivenko-Cantelli: finite bracketing number for every $\epsilon > 0$ implies uniform law of large numbers holds.

Theorem A.9 (Glivenko-Cantelli, [60] Theorem 19.4)**.** *Every class $\mathcal{F}$ of measurable functions such that $N_{[]}(\epsilon, \mathcal{F}, ||\cdot||_{L_1(P)}) < \infty$ for every $\epsilon > 0$ is P-Glivenko-Cantelli.*

Theorem A.10 ([60] Example 19.11)**.** *Let $\mathcal{F}$ be the collection of all monotone functions $f : \mathbb{R} \to [-1, 1]$, or bigger, the set of all functions that are of variation bounded by 1. Then there exists a constant K such that, for every $r \geq 1$ and probability measure P,*

$$\log N_{[]}(\epsilon, \mathcal{F}, ||\cdot||_{L_r(P)}) \leq K(\frac{1}{\epsilon}).$$

An alternative condition for P-Glivenko-Cantelli is based on covering number:

Definition A.2.3 (covering number)**.** The covering number $N(\epsilon, \mathcal{F}, ||\cdot||)$ is minimum number of open balls $\{f : ||g - f|| < \epsilon\}$ of radius ϵ and center g needed to cover $\mathcal{F}$.

Theorem A.11. *Let $\mathcal{F}$ be a class of measurable functions with an envelope function F, i.e., $|f(x)| \leq F(x) < \infty$ for every x and f. Suppose*

$$\sup_Q N(\epsilon ||F||_{L_1(Q)}, \mathcal{F}, ||\cdot||_{L_1(Q)}) < \infty,$$

for every $\epsilon > 0$, where sup *is taken over all probability measures Q such that $||F||_{L_1(Q)} > 0$. If $PF < \infty$, then $\mathcal{F}$ is P-Gilvenko-Cantelli.*

Theorem A.12 (covering number for Lipschitz function balls, [29])**.** *For a totally bounded metric space $(\mathcal{X}, d)$ and the unit ball B of $(Lip(\mathcal{X}), ||\cdot||_L)$,*

$$N(\mathcal{X}, 4\epsilon, d) \leq \log_2 N(\epsilon, B, ||\cdot||_\infty) \leq N(\mathcal{X}, \epsilon/4, d) \log_2(2\lfloor \frac{2diam(\mathcal{X})}{\epsilon} \rfloor + 1),$$

where $N(\mathcal{X}, \epsilon, d)$ is the minimum number of balls with centers in $\mathcal{X}$ and radius d to cover $\mathcal{X}$. If, in addition, $\mathcal{X}$ is connected and centred,

$$N(\mathcal{X}, 2\epsilon, d) \leq \log_2 N(\epsilon, B, ||\cdot||_\infty) \leq N(\mathcal{X}, \epsilon/2, d) + \log_2(2\lfloor \frac{2diam(\mathcal{X})}{\epsilon} \rfloor + 1))$$

Remark. only metric properties of the underlying space $\mathcal{X}$ is involved.

A.3 Other elementary facts

Lemma A.13. *The maximum of a family of convex functions (finite or infinite) is convex.*

See, for example, [7].

Lemma A.14.

$$\max\{a_1, a_2\} - \max\{b_1, b_2\} \leq \max\{a_1 - b_1, a_2 - b_2\}.$$

In general,

$$\max\{a_1, \cdots, a_k\} - \max\{b_1, \cdots, b_k\} \leq \max\{a_1 - b_1, \cdots, a_k - b_k\}.$$

Bibliography

[1] Emmanuel Abbe. Community detection and stochastic block models: recent developments. *The Journal of Machine Learning Research*, 18(1):6446–6531, 2017.

[2] Giovanni Alberti, Stefano Bianchini, and Gianluca Crippa. Structure of level sets and Sard-type properties of Lipschitz maps. *Annali della Scuola Normale Superiore di Pisa-Classe di Scienze*, 12(4):863–902, 2013.

[3] Bryon Aragam, Chen Dan, Eric P Xing, and Pradeep Ravikumar. Identifiability of nonparametric mixture models and Bayes optimal clustering. *The Annals of Statistics*, 48(4):2277–2302, 2020.

[4] Sivaraman Balakrishnan, Martin J. Wainwright, and Bin Yu. Statistical guarantees for the EM algorithm: From population to sample-based analysis. *The Annals of Statistics*, 45(1):77 – 120, 2017.

[5] Shai Ben-David, Ulrike Von Luxburg, and Dávid Pál. A sober look at clustering stability. In *Proceedings of the 19th Annual Conference on Learning Theory (COLT)*, pages 5–19. Springer, Berlin, 2006.

[6] Charles Bouveyron, Gilles Celeux, T Brendan Murphy, and Adrian E Raftery. *Model-based clustering and classification for data science: with applications in R*, volume 50. Cambridge University Press, 2019.

[7] Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge University Press, 2004.

[8] T. Tony Cai, Jing Ma, and Linjun Zhang. CHIME: Clustering of high-dimensional Gaussian mixtures with EM algorithm and its optimality. *The Annals of Statistics*, 47(3):1234 – 1267, 2019.

[9] Tony F Chan, Selim Esedoglu, and Mila Nikolova. Algorithms for finding global minimizers of image segmentation and denoising models. *SIAM Journal of Applied Mathematics*, 66(5):1632–1648, 2006.

[10] Xin Chen and Anderson Y Zhang. Optimal clustering in anisotropic Gaussian mixture models. *arXiv preprint arXiv:2101.05402*, 2021.

[11] Yen-Chi Chen, Christopher R Genovese, and Larry Wasserman. Density level sets: Asymptotics, inference, and visualization. *Journal of the American Statistical Association*, 112(520):1684–1696, 2017.

[12] Eric C Chi and Kenneth Lange. Splitting methods for convex clustering. *Journal of Computational and Graphical Statistics*, 24(4):994–1013, 2015.

[13] Francis Clarke. *Functional Analysis, Calculus of Variations and Optimal Control*. Graduate Texts in Mathematics. Springer London, 2013.

[14] Michael B Cohen, Sam Elder, Cameron Musco, Christopher Musco, and Madalina Persu. Dimensionality reduction for k-means clustering and low rank approximation. In *Proceedings of the forty-seventh annual ACM symposium on Theory of computing*, pages 163–172, 2015.

[15] Luc Devroye, László Györfi, and Gábor Lugosi. *A probabilistic theory of pattern recognition*. Springer, New York, 1996.

[16] Rick Durrett. *Probability: Theory and Examples*. Cambridge University Press, USA, 4th edition, 2010.

[17] Soheil Feizi, Hamid Javadi, and David Tse. Tensor biclustering. *Advances in Neural Information Processing Systems*, page 1311–1320, 2017.

[18] Thomas S Ferguson. *Mathematical statistics: A decision theoretic approach*. Academic, New York, 1967.

[19] Gerald B Folland. *Real analysis: modern techniques and their applications*. John Wiley & Sons, 1999.

[20] Tom Goldstein, Xavier Bresson, Stan Osher, and Antonin Chambolle. Global minimization of Markov random fields with applications to optical flow. *Inverse Problems and Imaging*, 6(4):623–644, 2012.

[21] László Györfi, Michael Köhler, Adam Krzyżak, and Harro Walk. *A distribution-free theory of nonparametric regression*. Springer-Verlag, New York, 2002.

[22] Sariel Har-Peled and Soham Mazumdar. On coresets for k-means and k-median clustering. In *Proceedings of the thirty-sixth annual ACM symposium on Theory of computing*, pages 291–300, 2004.

[23] Wolfgang Karl Härdle and Léopold Simar. *Applied multivariate statistical analysis*. Springer Nature, 2019.

[24] John A. Hartigan. *Clustering Algorithms*. John Wiley & Sons, 1975.

[25] Prateek Jain, Purushottam Kar, et al. Non-convex optimization for machine learning. *Foundations and Trends® in Machine Learning*, 10(3-4):142–363, 2017.

[26] Svante Janson. Large deviations for sums of partly dependent random variables. *Random Structures & Algorithms*, 24(3):234–248, 2004.

[27] Jiashun Jin, Zheng Tracy Ke, and Wanjie Wang. Phase transitions for high dimensional clustering and related problems. *The Annals of Statistics*, 45(5):2151 – 2189, 2017.

[28] Leonard Kaufman and Peter J Rousseeuw. *Finding groups in data: an introduction to cluster analysis*. John Wiley & Sons, 2009.

[29] A. N. Kolmogorov and V. M. Tihomirov. ε-entropy and ε-capacity of sets in functional spaces. *American Mathematical Society Translations (2)*, 17:277–364, 1961.

[30] Daniel Kraft. Measure-theoretic properties of level sets of distance functions. *The Journal of Geometric Analysis*, 26:2777–2796, 2016.

[31] Amit Kumar, Yogish Sabharwal, and Sandeep Sen. A simple linear time $(1+\varepsilon)$-approximation algorithm for k-means clustering in any dimensions. In *45th Annual IEEE Symposium on Foundations of Computer Science*, pages 454–462. IEEE, 2004.

[32] Jing Lei and Alessandro Rinaldo. Consistency of spectral clustering in stochastic block models. *The Annals of Statistics*, 43(1):215 – 237, 2015.

[33] Anna V Little, Mauro Maggioni, and James M Murphy. Path-based spectral clustering: Guarantees, robustness to outliers, and fast algorithms. *The Journal of Machine Learning Research*, 21, 2020.

[34] Stuart Lloyd. Least squares quantization in pcm. *IEEE transactions on information theory*, 28(2):129–137, 1982.

[35] Matthias Löffler, Anderson Y Zhang, and Harrison H Zhou. Optimality of spectral clustering in the Gaussian mixture model. *The Annals of Statistics*, 49(5):2506–2530, 2021.

[36] Vince Lyzinski, Daniel L. Sussman, Minh Tang, Avanti Athreya, and Carey E. Priebe. Perfect clustering for stochastic blockmodel graphs via adjacency spectral embedding. *Electronic Journal of Statistics*, 8(2):2905 – 2922, 2014.

[37] Francesco Maggi. *Sets of Finite Perimeter and Geometric Variational Problems: An Introduction to Geometric Measure Theory*. Cambridge Studies in Advanced Mathematics. Cambridge University Press, 2012.

[38] Enno Mammen and Wolfgang Polonik. Confidence regions for level sets. *Journal of Multivariate Analysis*, 122:202–214, 2013.

[39] Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of machine learning*. MIT press, 2018.

[40] Gabriel Peyré, Marco Cuturi, et al. Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019.

[41] Dimitris N Politis and Joseph P Romano. Large sample confidence regions based on subsamples under minimal assumptions. *The Annals of Statistics*, pages 2031–2050, 1994.

[42] Dimitris N Politis, Joseph P Romano, and Michael Wolf. *Subsampling*. Springer Science & Business Media, 1999.

[43] David Pollard. Strong consistency of k-means clustering. *The Annals of Statistics*, pages 135–140, 1981.

[44] David Pollard. A central limit theorem for k-means clustering. *The Annals of Probability*, 10(4):919–926, 1982.

[45] David Pollard. *Convergence of stochastic processes*. Springer-Verlag, New York, 1984.

[46] Carey E Priebe, Youngser Park, Joshua T Vogelstein, John M Conroy, Vince Lyzinski, Minh Tang, Avanti Athreya, Joshua Cape, and Eric Bridgeford. On a two-truths phenomenon in spectral graph clustering. *Proceedings of the National Academy of Sciences*, 116(13):5995–6000, 2019.

[47] William M. Rand. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, 66(336):846–850, 1971.

[48] Alessandro Rinaldo and Larry Wasserman. Generalized density clustering. *The Annals of Statistics*, 38(5):2678–2722, 2010.

[49] Robert J Serfling. *Approximation theorems of mathematical statistics*. John Wiley & Sons, New York, 1980.

[50] Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge University Press, 2014.

[51] Jun Shao. Linear model selection by cross-validation. *Journal of the American Statistical Association*, 88(422):486–494, 1993.

[52] Leon Simon et al. *Lectures on geometric measure theory*. The Australian National University, Mathematical Sciences Institute, Centre for Mathematics & its Applications, 1983.

[53] Mervyn Stone. Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society: Series B (Methodological)*, 36(2):111–133, 1974.

[54] Werner Stuetzle and Rebecca Nugent. A generalized single linkage method for estimating the cluster tree of a density. *Journal of Computational and Graphical Statistics*, 19(2):397–418, 2010.

[55] Arthur Szlam and Xavier Bresson. Total variation, cheeger cuts. In *ICML*, volume 10, pages 1039–1046, 2010.

[56] Michael J Todd. The many facets of linear programming. *Mathematical Programming*, 91(3):417–436, 2002.

[57] Gregor Traven, Gal Matijevič, Tomaz Zwitter, M Žerjal, Janez Kos, Martin Asplund, Joss Bland-Hawthorn, Andrew R Casey, Gayandhi De Silva, Kenneth Freeman, et al. The galah survey: classification and diagnostics with t-SNE reduction of spectral information. *The Astrophysical Journal Supplement Series*, 228(2):24, 2017.

[58] Nicolás García Trillos, Franca Hoffmann, and Bamdad Hosseini. Geometric structure of graph Laplacian embeddings. *The Journal of Machine Learning Research*, 22(1):2934–2988, 2021.

[59] Alexander B Tsybakov. Optimal aggregation of classifiers in statistical learning. *The Annals of Statistics*, 32(1):135–166, 2004.

[60] A. W. van der Vaart. *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 1998.

[61] Mark J Van Der Laan and Jenny Bryan. Gene expression analysis with the parametric bootstrap. *Biostatistics*, 2(4):445–461, 2001.

[62] Ulrike Von Luxburg. A tutorial on spectral clustering. *Statistics and Computing*, 17:395–416, 2007.

[63] Ulrike Von Luxburg, Mikhail Belkin, and Olivier Bousquet. Consistency of spectral clustering. *The Annals of Statistics*, pages 555–586, 2008.

[64] Ulrike Von Luxburg and Shai Ben-David. Towards a statistical theory of clustering. In *Pascal workshop on statistics and optimization of clustering*, pages 20–26. London, UK, 2005.

[65] Ulrike von Luxburg and Olivier Bousquet. Distance-based classification with Lipschitz functions. *The Journal of Machine Learning Research*, 5(Jun):669–695, 2004.

[66] Ulrike Von Luxburg et al. Clustering stability: an overview. *Foundations and Trends® in Machine Learning*, 2(3):235–274, 2010.

[67] Abraham Wald. *Statistical decision functions*. Wiley, 1961.

[68] Junhui Wang. Consistent selection of the number of clusters via crossvalidation. *Biometrika*, 97(4):893–904, 2010.

[69] Nik Weaver. *Lipschitz algebras*. World Scientific, 2018.

[70] Yuhong Yang. Comparing learning methods for classification. *Statistica Sinica*, 16(2):635–657, 2006.

[71] Yuhong Yang. Consistency of cross validation for comparing regression procedures. *The Annals of Statistics*, 35(6):2450 – 2473, 2007.