Journal of Machine Learning Research 26 (2025) 1-36 Submitted 11/23; Revised 12/25; Published 12/25

Step and Smooth Decompositions as Topological Clustering

Luciano Vinas LUCIANOVINAS@G.UCLA.EDU
Department of Statistics
University of California, Los Angeles
Los Angeles, CA 90095-1554, USA

Arash A. Amini AAAMINI@STAT.UCLA.EDU
Department of Statistics
University of California, Los Angeles
Los Angeles, CA 90095-1554, USA

Editor: Krishnakumar Balasubramanian

Abstract

We investigate a class of recovery problems for which observations are a noisy combination of continuous and step functions. These problems can be seen as non-injective instances of non-linear ICA with direct applications to image decontamination for magnetic resonance imaging. Alternately, the problem can be viewed as clustering in the presence of structured (smooth) contaminant. We show that a global topological property (graph connectivity) interacts with a local property (the degree of smoothness of the continuous component) to determine conditions under which the components are identifiable. Additionally, a practical estimation algorithm is provided for the case when the contaminant lies in a reproducing kernel Hilbert space of continuous functions. Algorithm effectiveness is demonstrated through a series of simulations and real-world studies.

Keywords: Non-convex Decompositions, Identifiability, RKHS, Clustering, Topological Analysis

1. Introduction

The prototypical recovery problem is nonparametric regression where we observe an unknown function corrupted by additive white noise: $y_i = f^*(x_i)+\varepsilon_i$, for $i = 1, \ldots, n$, where f^* belongs to some function class $\mathcal{F}$ and ε_i is the measurement noise. Important to the recovery is the structure of $\mathcal{F}$ and how it can be leveraged to differentiate observations from noise. Examples of previously explored structures in nonparametric regression include: smoothness (Tsybakov, 2009), sparsity (Wainwright, 2009; Bickel et al., 2009), homogeneity (Ke et al., 2015), and piecewise simplicity (Kim et al., 2009; Tibshirani, 2014). In each of these problems, there is a particular interest in uncovering the structure-specific recovery conditions under which a finite-sample, data-estimate $\widehat{f}$ eventually recovers the optimal, data-generating f^*.

Another flavor of recovery problems include decompositions of the form

$$y_i = f^*(x_i) + g^*(x_i) + \varepsilon_i, \tag{1}$$

where the recovery quantities of interest include both f^* and g^*. Naturally, this type of recovery problem, with its multiple recoverable quantities, is more difficult than basic

nonparametric regression. Examples of such decompositions with provable recovery guarantees are rare but some notable examples include the case of sparse plus low-rank matrix recovery (Chandrasekaran et al., 2009; Bahmani and Romberg, 2016; Tanner and Vary, 2023) and compressed sensing in a pair of orthogonal bases (Donoho and Kutyniok, 2013).

In this paper, we consider a nonparametric decomposition of the form (1) where the signal is a combination of continuous and step functions. We provide identifiability conditions for the continuous and step functions f^* and g^* in terms of the modulus of continuity of f^* and the height between steps in g^*. Analysis of f^* and g^* will be sufficiently general, where each function is considered to be a mapping from a metric space $(\mathcal{X},\, d)$ to a normed vector space $(\mathcal{Y},\, \|\cdot\|)$.

In its simplest formulation, we consider f^* to be real-valued and continuous, lying in a Hilbert-norm R-ball of a reproducing kernel Hilbert space (RKHS). For this scenario, a practical estimation algorithm is proposed with consistency guarantees given in terms of spectral quantities related to the observed kernel matrix of the RKHS.

As in most regression analysis, we conduct our analysis under finite sampling constraints. For g^* which attains at most M unique values within a given sample, the composite observations will be re-expressed as

$$y_i = f^*(x_i) + \mu^*_{z^*_i} + \varepsilon_i, \quad \text{for } i = 1, \dots, n \tag{2}$$

where $\boldsymbol{\mu}^* \in \mathbb{R}^M$ is a vector of values referred to as the levels of g^*, and $z^*_i \in [M]$ are labels to the corresponding levels of g^*. Our main goal is to recover the labels z^*_i correctly, with a secondary goal of recovering the levels $\boldsymbol{\mu}^*$ and the continuous function f^*. For our finite sample setting, recovery of $f^* \in \mathcal{F}$ will be relaxed to finding an element of the equivalence class

$$[f^*]_n = \big\{ f \in \mathcal{F} :\ f(x_i) = f^*(x_i),\ \forall i \in [n] \big\}. \tag{3}$$

This recovery condition may be refined to instead selecting a representative solution from $[f^*]_n$, such as a minimum-norm solution. An approach of this sort will depend on the regularity available in the function space $\mathcal{F}$ and will not be a topic of focus in our forthcoming analysis.

1.1 Applications

To motivate the problem, let us give some concrete applications of the step and smooth decomposition model (2).

Decompositions in Non-linear ICA

Non-linear independent component analysis (ICA) (Hyvärinen and Pajunen, 1999) provides a general framework to describe signal mixing problems. In non-linear ICA, the mixed observation $\boldsymbol{y} = \psi(\boldsymbol{s})$ is generated using independent, latent sources $\boldsymbol{s} \in \mathbb{R}^n$ and a non-linear, mixing function $\psi : \mathbb{R}^n \to \mathbb{R}^n$. In other ICA formulations (Hyvarinen et al., 2019), joint independence of $\boldsymbol{s}$ is relaxed to a conditional independence given some auxiliary information $\boldsymbol{u} \in \mathbb{R}^n$. That is,

$$\log p(\boldsymbol{s}|\boldsymbol{u}) = \sum_{i=1}^{n} q_i(s_i, \boldsymbol{u})$$

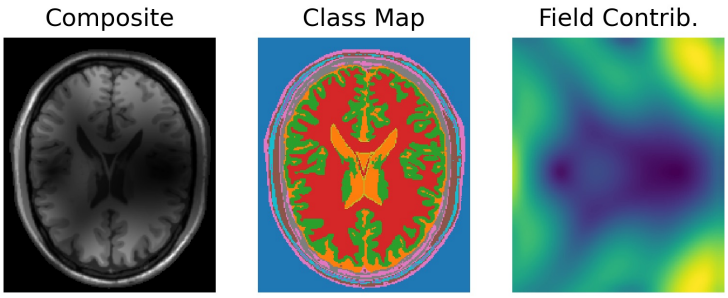


Figure 1: Example of a prominent bias field modifying the BrainWeb (Cocosco et al., 1997) phantom. Leftmost image is the MRI of a synthetic brain which has been perturbed by a contaminant field. Middle image is a tissue categorization for the synthetic brain. Rightmost image is the contaminant field in ambient space (in absence of the synthetic brain).

for appropriately defined densities q_i.

Decomposition (2) can be understood in terms of a self-mixing, non-linear ICA problem. In the simplest scenario, we may consider sources $s_i = (x_i, u_i)$ with auxiliary information $u_i \sim \text{Unif}[0, 1)$ and mixing defined by

$$\psi(s_i) = f^*(x_i) + \mu^*_{\phi(x_i,u_i)} \quad \text{where} \quad \phi(x_i, u_i) = \lfloor u_i \cdot M \rfloor + 1. \tag{4}$$

Generalizations to (4) may consider different cut-off functions $\phi(x_i, u_i)$ which also incorporate sample spatial information x_i in their cut-offs.

In contrast to traditional ICA problems, the mixing function defined in (4) is not necessarily injective on $\mathbb{R}^n$ for all choices of f^* and ϕ. This a recovery setting not covered in recent non-linear ICA literature (Hyvarinen et al., 2019; Khemakhem et al., 2020; Zheng et al., 2022) and one we are interested in exploring in this paper. In particular, when given partial information $\{(x_i, y_i)\}_i$, which properties of the data, if any at all, can help overcome the non-injectivity of a general f^* and μ^*_ϕ?

Decompositions in Medical Image Correction

In magnetic resonance imaging (MRI), image quality can be affected by factors ranging from radiofrequency coil setup to patient positioning and geometry (Asher et al., 2010). Dependent on these factors, MRI images may be contaminated with a spatially smooth, multiplicative field, known as the bias field. Figure 1 illustrates an example of a contaminated MRI image.

The MRI bias field problem admits the following multiplicative formulation (Vovk et al., 2007),

$$y(x) = f^*(x) \cdot \mu^*(x), \quad \text{for } x \in \mathcal{X} \tag{5}$$

where f^* is a positive smooth field on $\mathcal{X}$, and $\mu^*(x)$ are, by convention, positive tissue values at locations $x \in \mathcal{X}$. Given a fixed number of tissues classes M, process (5) can be reformulated as (2) under a log-transformation.

In supervised learning tasks, the visual inconsistencies caused by MRI bias fields present significant challenges, as they prevent the acquisition of accurate ground truth signal

information from patient scans. This issue parallels the earlier discussed problem of non-linear ICA, where, again, learning is hampered due to partial information and concerns regarding injectivity.

1.2 Prior Work

To the authors' best knowledge, the closest work on the theory of continuous and step decompositions is Kim and Tagare (2014), where they provide a characterization of the set of viable functions given an observed composite signal h^*. The composite $h^* = f^* \cdot g^*$ is assumed to be the product of a positive continuous function f^* and a positive step-wise function g^*. Assuming knowledge of the tissue ratios $\{\mu_k/\mu_{k+1}\}_{k=1}^{M-1}$, Kim and Tagare (2014) have shown that one there are scalars $\{a_k\}_{k=1}^{M}$ such that the set

$$\widetilde{\mathcal{F}} = \big\{ f : \mathcal{X} \to \mathbb{R} : \forall x \in \mathcal{X},\, f(x) \in \{a_k h^*(x)\}_{k=1}^{M} \big\}$$

contains a unique scalar multiple of f^*. This result is then followed by a practical algorithm which optimizes over a soft-label surrogate of $\widetilde{\mathcal{F}}$.

The theoretical result of Kim and Tagare (2014) is interesting since it dramatically reduces the search space for a viable f, esp. when $\mathcal{X}$ is finite. What this result does not tell us is how to identify f^* in the set $\widetilde{\mathcal{F}}$, and whether f^* is identifiable at all. This issue becomes readily apparent in finite sample scenarios, where there may be multiple ways to construct observations h^* from different smooth-and-step pairs $(\tilde{f}, \tilde{g})$. In short, the work of Kim and Tagare (2014) does not address the question of identifiability which is a focus of our work. Moreover, when no level information is available, $\widetilde{\mathcal{F}}$ itself is unknown. In this regime, attempts to approximate the set $\widetilde{\mathcal{F}}$ would ultimately be sensitive to initialization choice for scale parameters $\{a_k\}_k$.

2. Identifiability Theory

We consider the problem of identifying components $(f^*, \boldsymbol{\mu}^*, \boldsymbol{z}^*)$ from observations

$$y_i = f^*(x_i) + \mu^*_{z_i^*}, \quad \text{for } i = 1, \dots, n, \tag{6}$$

where f^* belongs to a class of smooth functions, from a metric space $(\mathcal{X}, d)$ to a normed space $(\mathcal{Y}, \|\cdot\|)$. Specifically, we assume that $f^* \in \mathcal{F}_\omega(\mathcal{X})$, the set of uniformly continuous functions with *modulus of continuity* $\omega : [0, \infty) \to [0, \infty)$, that is,

$$\mathcal{F}_\omega(\mathcal{X}) := \big\{ f : \mathcal{X} \to \mathcal{Y} :\ \|f(x) - f(x')\| \le \omega(d(x, x')),\ \forall x, x' \in \mathcal{X} \big\}. \tag{7}$$

For ease of presentation we will assume $\mathcal{Y} = (\mathbb{R}, |\cdot|)$. Proofs of the results for a general normed space $\mathcal{Y}$ can be found in Appendix A.

A model is identifiable if the ground-truth parameters $(f^*, \boldsymbol{\mu}^*, \boldsymbol{z}^*)$ can be unambiguously recovered from observed samples $\{y_i\}_i$ following (6). For our recovery procedure we consider solving the optimization

$$(\widehat{f}, \widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{z}}) = \operatorname*{argmin}_{\substack{f \in \mathcal{F}_\omega(\mathcal{X}), \\ \boldsymbol{\mu} \in \mathbb{R}^M,\, \boldsymbol{z} \in [M]^n}} \frac{1}{n} \sum_{i=1}^{n} (y_i - \mu_{z_i} - f(x_i))^2, \tag{8}$$

as well as *zero-mean* version where a constraint is added ensuring that f is empirically zero-mean, i.e., $\sum_{i=1}^n f(x_i) = 0$. This zero-mean constraint addresses issues analogous to the scalar multiple problem described by Kim and Tagare (2014). Note that, in our case, correctly recovering $(\boldsymbol{\mu}^*, \boldsymbol{z}^*)$ will also recover $[f^*]_n$ since $f^*(x_i) = y_i - \mu^*_{z_i^*}$. Then, by providing conditions under which (8) unambiguously recovers the sampled clusters $\{\mu^*_{z_i^*}\}_i$, we will have shown identifiability for the step and smooth decomposition (6).

2.1 Topological Clustering

To motivate our forthcoming topological definitions, consider the following failure case for step and smooth identifiability:

Example 1 *Consider the two cluster case ($M = 2$) on $\mathcal{X} = (\mathbb{R}^d, \|\cdot\|_2)$ and take $\omega(t) = t$, that is, $\mathcal{F}_\omega(\mathcal{X})$ contains all 1-Lipschitz functions on $\mathcal{X}$. Let $f^*(\boldsymbol{x}) = 0$ and $\mu_1^* = -\mu_2^* = 1$ with linearly separable clusters $\mathcal{C}_k = \{i \in [n] : z_i^* = k\}$; that is, there exists unit-norm $\boldsymbol{w} \in \mathbb{R}^d$ and $c_1, c_2 \in \mathbb{R}$ such that*

$$\boldsymbol{w}^T \boldsymbol{x}_i \le c_1 < c_2 \le \boldsymbol{w}^T \boldsymbol{x}_j$$

for all $i \in \mathcal{C}_1$ and $j \in \mathcal{C}_2$.

Consider the piecewise cluster-interpolating function $\widetilde{f}(\boldsymbol{x}) = -\min\{\max\{\boldsymbol{w}^T\boldsymbol{x} - c_1, 0\}, 2\}$. Clearly, $\widetilde{f}$ is 1-Lipschitz, that is $\widetilde{f} \in \mathcal{F}_\omega(\mathcal{X})$. Now, if $\mathcal{C}_1$ and $\mathcal{C}_2$ are sufficiently separated so that $c_2 - c_1 \ge 2 = \mu_1^* - \mu_2^*$, then

$$\widetilde{f}(\boldsymbol{x}_i) + \mu_1^* = f^*(\boldsymbol{x}_i) + \mu^*_{z_i^*}$$

for every $i \in [n]$. That is, rather than two cluster, we can put all points in a single cluster (Cluster 1) and explain the variation in y_i entirely by the smooth function $\widetilde{f}$, or we can put the points in two clusters and explain the residual variation (which turns out to be zero in this case) by the original f^*. In other words, in this case the observations $y_i = f^*(\boldsymbol{x}_i) + \mu^*_{z_i^*}$ do not allow for an unambiguous recovery of $(f^*, \boldsymbol{\mu}^*, \boldsymbol{z}^*)$.

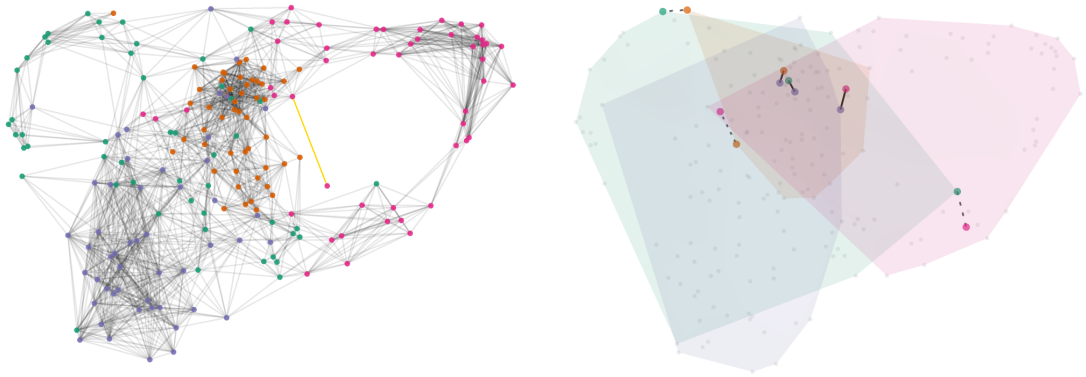
Example 1 shows that without a constraint on how far apart samples in the point cloud $\{x_i\}_{i=1}^n$ are placed, it is always possible to construct examples where one can interpolate between different clusters using a smooth $\widetilde{f}$. Stated differently, the distance between samples, compared to the difference in cluster levels μ_k^* for $k \in [M]$, must be within the scale allowed by the modulus of continuity ω. It turns out the the proper way to measure the separation of samples is through the connectivity of the ρ-neighbor graph, which we recall next:

Definition 1 (Neighbor Graph) *The ρ-neighbor graph $G_\rho(X)$ of point cloud $X = \{x_i\}$ is the graph with vertex set $[n]$ and edge set*

$$E = \big\{(i,j) \in [n]^2 : i \ne j \text{ and } d(x_i, x_j) \le \rho\big\}.$$

The ρ-neighbor graph captures some aspect of the topology of the point cloud. Paired with the modulus of continuity ω, this graph allows us to quantify long-range variation of a particular $f^* \in \mathcal{F}_\omega(\mathcal{X})$ via its local variations along the edges.

In this sense, every point cloud X has a minimum, necessary communication length $\rho_{\min}(X)$, such that all long-range variations in $\{d(x, x')\}_{x,x' \in X}$ can be bounded in terms of local ones $\{d(x_i, x_j)\}_{(i,j)\in E}$ where $G_{\rho_{\min}}(X) = ([n], E)$. More formally:



(a) ρ-neighbor connectivity graph (b) Cluster connectivity graph

Figure 2: A ρ-neighbor and cluster connectivity graph on UMAP-reduced features for four topics from the "20 Newsgroups" classification dataset. Highlighted in gold in the left subfigure is an edge with length equal to connectivity $\rho_{\min}$. Drawn in black in the right subfigure are the corresponding cluster distance edges $d(\mathcal{C}_k, \mathcal{C}_\ell)$. Cluster graph edges which are larger than δ_{lbd} are drawn in dashed. The final cluster graph $G_{\delta_{\text{lbl}}}(\mathcal{C})$ is a tree with a connecting hub at the blue colored cluster.[1]

Definition 2 (Connectivity) *For a point cloud X, the connectivity is defined as*

$$\rho_{\min}(X) := \inf\big\{\rho > 0 : \ G_\rho(X) \ \textit{is connected}\big\}.$$

So far, we have defined a connectivity parameter, $\rho_{\min}(X)$, such that deviations in f^* can be translated into traversals between neighboring nodes in X. Let us define a similar concept for the deviations of the step component $\mu^*_{(\cdot)}$. Let $\mathcal{C} = \{\mathcal{C}_k\}_{k\in[M]}$ be the set of M clusters with $\mathcal{C}_k = \{i \in [n] : z^*_i = k\}$. For each pair of clusters, there is a corresponding notion of *cluster distance* given by:

$$d(\mathcal{C}_k, \mathcal{C}_\ell) := \min_{i\in\mathcal{C}_k,\, j\in\mathcal{C}_\ell} d(x_i, x_j). \tag{9}$$

Then, for an associated tolerance parameter $\delta > 0$, let us construct the δ-neighbor graph $G_\delta(\mathcal{C})$ with vertex set $\mathcal{C}$ and edge set

$$\big\{(\mathcal{C}_k, \mathcal{C}_\ell) : \ k \neq \ell \text{ and } \ d(\mathcal{C}_k, \mathcal{C}_\ell) \leq \delta\big\}.$$

Similar to the smooth case, there is a minimum necessary communication length δ_{lbl}, depending on both the point cloud X and the set of labels $\boldsymbol{z}^*$, such that all deviations of $\mu^*_{(\cdot)}$ can be translated to traversals on $G_\delta(\mathcal{C})$.

Definition 3 (Label distance) *The label distance for paired data $(X, \boldsymbol{z}^*)$ is*

$$\delta_{\text{lbl}}(X, \boldsymbol{z}^*) := \inf\big\{\delta > 0 : \ G_\delta(\mathcal{C}) \ \textit{is connected}\big\}. \tag{10}$$

1. For interactive 3D network representation: https://github.com/lucianoAvinas/topological-clustering-plots.

When it is clear from context, we omit dependence on sample $(X, \boldsymbol{z}^*)$ for the previously defined topological quantities. An example of these quantities in real-world data is given Figure 2.

Finally, we also need the following simple condition on the labels:

Definition 4 (Label saturation) *A label vector $\boldsymbol{z}^* \in [M]^n$* saturates *$[M]$ if every label in $[M]$ is present in $\boldsymbol{z}^* = (z_1^*, \ldots, z_n^*)$, that is, for every $\ell \in [M]$, there is $i \in [n]$, with $z_i^* = \ell$.*

This condition is needed to avoid the trivial case where some of the levels $\{\mu_k^*\}$ are redundant (i.e. label k does not appear until some $n > N$). It is clearly necessary for the identifiability of the levels and labels in the uncontaminated model. It is always possible to redefine $\boldsymbol{z}^*$ to be saturated by relabeling, and dropping redundant levels.

2.2 Identifiability Results

Our main result is the following cluster recovery guarantee:

Theorem 5 (Cluster recovery) *Let $X = \{x_i\}_{i=1}^n$ be a point cloud in a metric space $(\mathcal{X}, d)$ and let $\{y_i\}_{i=1}^n$ follow model* (6) *with $f^* \in \mathcal{F}_\omega(\mathcal{X})$ and $\boldsymbol{z}^* \in [M]^n$ that saturates $[M]$. If the connectivity $\rho_{\min}$ of X satisfies*

$$\omega(\rho_{\min}) < \frac{1}{2M} \min_{k \neq \ell} |\mu_k^* - \mu_\ell^*|, \tag{11}$$

then, the labels $\widehat{\boldsymbol{z}}$ produced by (8) *have zero misclassification error relative to $\boldsymbol{z}^*$.*

Our next result is an error bound on the recovered levels $\widehat{\boldsymbol{\mu}}$:

Proposition 6 (Level recovery) *Under the assumptions of Theorem 5, let $(\widehat{f}, \widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{z}})$ be the solution of the zero-mean version of problem* (8)*. Then, we have*

$$\max_{k \in [M]} |\mu_k^* - \widehat{\mu}_k| \leq 2(M-1)\, \omega(\delta_{\text{lbl}}) + \Big| \frac{1}{n} \sum_{i=1}^n f^*(x_i) \Big|. \tag{12}$$

In essence, both Theorem 5 and Proposition 6 provide deviation bounds under specific connectivity constraints. The quantities $\rho_{\min}$ and δ_{lbl} gauge the minimum jump distances at which the induced graphs of $\{x_i\}_{i=1}^n$ and $\{\mathcal{C}_k\}_{k=1}^M$ remain connected. The modulus $\omega(\cdot)$ then translates these jumps in distances into equivalent jumps in levels, observed indirectly through $\{y_i\}_i$.

Theorem 5 says that perfect cluster recovery $\widehat{\boldsymbol{z}} \equiv \boldsymbol{z}^*$ is attainable if this translated jump is roughly below the minimum resolution of the true levels $\{\mu_k^*\}$. Proposition 6 has a similar theme but now in the context of level recovery where, unlike $\boldsymbol{z}^* \in [M]^n$, the levels $\mu_k^* \in \mathbb{R}$ lie on a continuum. This leads to a gradual reduction in error as outlined in Proposition 6, contrasting with the sharp recovery of discrete labels $z_i^* \in [M]$ in Theorem 5.

The remainder term $|\frac{1}{n} \sum_i f^*(x_i)|$ in (12) highlights the scalar-shift ambiguity inherent in the components of model (6), where for any scalar $c \in \mathbb{R}$, it is possible to rewrite (6) as

$$y_i = (f^*(x_i) - c) + (\mu_{z_i^*}^* + c).$$

As such, the two components are only identifiable up to a scalar shift. More generally, problem (8) can be extended to include a constraint $\frac{1}{n}\sum_{i=1}^{n} f(x_i) = \overline{f^*}$ for some select mean value $\overline{f^*}$, in which case Proposition 6 holds with the remainder term $|\frac{1}{n}\sum_i f^*(x_i) - \overline{f^*}|$.

As an immediate corollary to Proposition 6, one can show that, under mild regularity on the sampling of $(X, \boldsymbol{z}^*)$, the zero-mean recovery problem (8) achieves asymptotic identifiability of $(\boldsymbol{\mu}^*, \boldsymbol{z}^*)$. As this corollary references multiple sets of samples, the notation $(\cdot)^{(n)}$ will be used to differentiate parameters belonging to different sets of observations $\{y_i\}_i$. We also allow the number of observed levels M_n to grow with n. We say that a condition is *eventually satisfied* if it holds for all $n \geq N$ for some $N \in \mathbb{N}$.

Corollary 7 *Consider a sequence of point clouds $\{X^{(n)}\}$, with corresponding true labels $\{\boldsymbol{z}^{*(n)}\}$ and class levels $\boldsymbol{\mu}^{*(n)} \in \mathbb{R}^{M_n}$. Let $\delta_{\mathrm{lbl}}^{(n)}$ be the label distance for $(X^{(n)}, \boldsymbol{z}^{*(n)})$. Assume that the connectivity condition* (11) *is eventually satisfied, and as $n \to \infty$,*

$$\omega(\delta_{\mathrm{lbl}}^{(n)}) = o(M_n^{-1}), \quad \frac{1}{n}\sum_{x \in X^{(n)}} f^*(x) = o(1).$$

Then for any solution $(\widehat{f}^{(n)}, \widehat{\boldsymbol{\mu}}^{(n)}, \widehat{\boldsymbol{z}}^{(n)})$ of the zero-mean version of problem (8)*,*

$$\lim_{n\to\infty} \max_{k \in M_n} \left|\mu_k^{*(n)} - \widehat{\mu}_k^{(n)}\right| = 0.$$

According to Corollary 7, when $\{M_n\}$ is bounded, a set of sufficient conditions for recovery of both clusters and levels is:

$$\rho_{\min}^{(n)} = o(1), \quad \delta_{\mathrm{lbl}}^{(n)} = o(1), \quad \text{and} \quad \Delta_n := \min_{k \neq \ell} |\mu_k^{*(n)} - \mu_\ell^{*(n)}| = \Omega(1),$$

i.e., minimum level gap is bounded below. When $\{M_n\}_n$ is unbounded, both the connectivity $\rho_{\min}$ and the label distance δ_{lbl} must decrease more rapidly. For example, when the smooth component is Lipschitz (i.e., $\omega(t) = Lt$), a set of sufficient conditions are

$$\rho_{\min}^{(n)} = o(\Delta_n / M_n), \quad \delta_{\mathrm{lbl}}^{(n)} = o(1/M_n).$$

Note that, while Corollary 7 is a deterministic result, it can be translated to a high probability version given appropriate assumptions on the sampling distribution of $(X, \boldsymbol{z})$.

The identifiability results of this section are intuitive and are described in terms of easily understood topological quantities. It is worth emphasizing that, prior to our analysis, obtaining a perfect classification result similar to Theorem 5 is not immediately clear for a general context. That is, irrespective of the placements of labels $\boldsymbol{z}^*$ on the point cloud X, and regardless of the dimension of the space carrying X, we have shown that one can globally control $\widehat{\boldsymbol{z}}$ using only a scalar parameter of the point cloud, namely, the radius of connectivity of its associated neighbor graphs $G_\rho(X)$.

3. Methods and Optimization

For practical estimation, we consider estimating functions $f^* \in \mathbb{H}$ lying in the Hilbert-norm R-ball of an RKHS. The following example shows that this case can be treated as a special case of (7) with a linear modulus $\omega(t) = O(t)$.

Example 2 *Consider the case where f^* lies in RKHS $\mathbb{H}$ with kernel $\mathcal{K} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. The natural metric to consider on $\mathcal{X}$ is the so-called* kernel metric

$$d_{\mathcal{K}}(x, x') := \|\mathcal{K}(x, \cdot) - \mathcal{K}(x', \cdot)\|_{\mathbb{H}} = \sqrt{\mathcal{K}(x, x) - 2\mathcal{K}(x, x') + \mathcal{K}(x', x')}. \tag{13}$$

Using the Cauchy–Schwarz inequality, it is straightforward to show the following Lipschitz property: For any $f \in \mathbb{H}$, we have

$$|f(x) - f(x')| \leq \|f\|_{\mathbb{H}}\, d_{\mathcal{K}}(x, x')$$

for all $x, x' \in \mathcal{X}$. Letting ω_f denote a modulus of continuity of function f, the above shows that one can take $\omega_f(t) = \|f\|_{\mathbb{H}} \cdot t$ for all $f \in \mathbb{H}$. If we further assume $\|f^*\|_{\mathbb{H}} \leq R$ for some constant R, then $\omega(t) = O(t)$.

AltMin Algorithm For our estimation procedure, we propose a blockwise coordinate descent with alternating updates on $(\boldsymbol{\mu}, \boldsymbol{z})$ and f. More specfically, in each iteration, the current estimates $(\widehat{f}, \widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{z}})$ are updated to the new ones $(\widehat{f}^+, \widehat{\boldsymbol{\mu}}^+, \widehat{\boldsymbol{z}}^+)$ by

$$\widehat{f}^+ = \operatorname*{argmin}_{f \in \mathbb{H}} \frac{1}{n} \sum_{i=1}^{n} (y_i - \widehat{\mu}_{\widehat{z}_i} - f(x_i))^2 + \tau \|f\|_{\mathbb{H}}^2, \tag{14}$$

$$(\widehat{\boldsymbol{\mu}}^+, \widehat{\boldsymbol{z}}^+) = \operatorname*{argmin}_{\boldsymbol{\mu} \in \mathbb{R}^M,\, \boldsymbol{z} \in [M]^n} \frac{1}{n} \sum_{i=1}^{n} (y_i - \mu_{z_i} - \widehat{f}^+(x_i))^2, \tag{15}$$

with τ and M being values which may be determined through a cross-validation procedure.

This procedure constitutes a block coordinate descent on the joint *regularized* objective function:

$$J(f, \boldsymbol{\mu}, \boldsymbol{z}) = \frac{1}{n} \sum_{i=1}^{n} (y_i - \mu_{z_i} - f(x_i))^2 + \tau \|f\|_{\mathbb{H}}^2. \tag{16}$$

Specifically, update (14) minimizes J with respect to f while keeping $(\boldsymbol{\mu}, \boldsymbol{z})$ fixed. Subsequently, update (15) minimizes the sum-of-squares term with respect to $(\boldsymbol{\mu}, \boldsymbol{z})$; since the regularization penalty $\tau \|f\|_{\mathbb{H}}^2$ is constant with respect to these parameters, this step is equivalent to minimizing J. Consequently, the value of the objective $J(f, \boldsymbol{\mu}, \boldsymbol{z})$ is non-increasing at every iteration. Since the objective is bounded below by zero, the algorithm is guaranteed to converge in objective value, and we define the stopping criterion based on the relative change in J.

For fixed $\widehat{f}$, optimization (15) can be solved through a k-means procedure. For RKHS $\mathbb{H}$ equipped with kernel $\mathcal{K} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, optimization (14) has the following representer solution

$$\widehat{f}^+ = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \widehat{\alpha}_i^+ \mathcal{K}(x_i, \cdot), \;\; \widehat{\boldsymbol{\alpha}}^+ := (K + \tau I_n)^{-1} (\boldsymbol{y} - \widehat{Z}\widehat{\boldsymbol{\mu}})/\sqrt{n} \tag{17}$$

where K is the $n \times n$ kernel matrix with entries $K_{ij} = \mathcal{K}(x_i, x_j)/n$ and $\widehat{Z} \in \{0, 1\}^{n \times L}$ is the one-hot encoding label matrix for previous label estimate $\widehat{\boldsymbol{z}}$.

3.1 Consistency of the One-Step Procedure

In this section, we analyze the statistical properties of the AltMin updates. Specifically, we focus on the "one-step" performance: given observations y_i generated from the composite model (1), does a single application of the update equations (14) and (15) yield consistent estimates for f^* and the clustering parameters?

We answer this in the affirmative, even without prior knowledge of the true step function g^*, by analyzing the KRR update (14) initialized with a trivial estimate $\widehat{g} = 0$ (equivalently, $\widehat{\boldsymbol{\mu}} = \mathbf{0}$). The resulting estimator for f^* is:

$$\widehat{f} = \operatorname*{argmin}_{f \in \mathbb{H}} \frac{1}{n} \sum_{i=1}^{n} (y_i - f(x_i))^2 + \tau \|f\|_{\mathbb{H}}^2. \tag{18}$$

The consistency of this estimator relies on the spectral separation between the smooth function $f^* \in \mathbb{H}$ and the step function g^*, which lies outside the RKHS. To quantify this separation, we look at the projection of

$$\boldsymbol{g}^* = \big(g^*(x_1), \dots, g^*(x_n)\big) \in \mathbb{R}^n$$

onto the eigenbasis of the (normalized) kernel matrix $K = (\mathcal{K}(x_i, x_j)/n) \in \mathbb{R}^{n \times n}$. Let $K = V \Lambda V^{\mathsf{T}}$ be the eigendecomposition with eigenvalues $\lambda_1 \geq \cdots \geq \lambda_n$. Let $\breve{\boldsymbol{g}} = V^{\mathsf{T}} \boldsymbol{g}^*$ be the coefficients of the step signal in this basis.

We introduce a *spectral tail condition* on g^*. Let $S_{g^*}(t)$ be the spectral survival function of the step component:

$$S_{g^*}(t) := \sum_{i=1}^{n} \frac{\breve{g}_i^2}{n} \cdot 1\{\lambda_i > t\}. \tag{19}$$

Let $r_n = \max\{i \in [n] : \breve{g}_i^2 > 0\}$ be the index of the last non-zero component, and let β_n be the largest exponent $\beta \geq 0$ satisfying the tail bound:

$$S_{g^*}(t) \;\leq\; \|\boldsymbol{g}^*\|_\infty^2 \cdot \Big(\frac{\lambda_{r_n}}{t}\Big)^{\beta}, \quad \text{for all } t > 0. \tag{20}$$

Such a tail bound always exists, since the trivial case $\beta = 0$ reduces to $\|\boldsymbol{g}^*/\sqrt{n}\|_2^2 \leq \|\boldsymbol{g}^*\|_\infty^2$. Intuitively, parameters r_n and β_n capture how "high-frequency" the step signal g^* is relative to the kernel K. A larger β_n implies that the energy of g^* decays faster as eigenvalues decrease, while $r_n \sim n$ implies the energy is not solely concentrated near $\lambda = \lambda_1$.

Our first main result establishes the consistency of the KRR estimator $\widehat{\boldsymbol{f}}$ in the mean square sense despite the presence of the unmodeled step component g^*.

Theorem 8 (MSE Consistency) *Consider a Mercer kernel and i.i.d. samples $\{x_i\}_i$. Let the regularization parameter τ_n be chosen such that $\tau_n \to 0$ and $n\tau_n \to \infty$. Assume the step component satisfies the spectral tail condition* (20) *with* $\liminf \beta_n > 0$ *and* $\liminf r_n/n > 0$, *and that $\xi_n := \lambda_{r_n}/\tau_n = o(1)$. Then,*

$$\lim_{n \to \infty} \mathbb{E}_{\boldsymbol{\varepsilon}} \operatorname{MSE}(\boldsymbol{f}^*, \widehat{\boldsymbol{f}}) = 0.$$

This MSE decay can be made explicit given knowledge of the population eigenvalue decay. For example, in the case of Sobolev kernels, we obtain the following rates:

Corollary 9 (Sobolev Rates) *Suppose $\mathcal{K}$ is the kernel for the Sobolev-α RKHS on $[0,1]$, where $\lambda_i \asymp i^{-2\alpha}$. If g^* satisfies the spectral tail condition with decay parameter β_n, there exists a selection of τ_n such that $\xi_n = n^{\frac{-2\alpha}{2\alpha+1}}$ and*

$$\mathbb{E}_{\boldsymbol{\varepsilon}} \operatorname{MSE}(\widehat{\boldsymbol{f}}) \lesssim n^{-\frac{2\alpha}{2\alpha+1}(\beta_n \wedge 2)}.$$

Theorem 8 confirms that f^* can be consistently estimated even when treating g^* as unmodeled noise, provided g^* is sufficiently rough (high-frequency) relative to the kernel. Consequently, the residuals $\boldsymbol{y} - \widehat{\boldsymbol{f}}$ become a reliable proxy for the noisy step signal $\boldsymbol{g}^* + \boldsymbol{\varepsilon}$.

This leads to our second result: the clustering step performed on these residuals converges to the optimal clustering of the uncontaminated k-means problem. Let $\widehat{\boldsymbol{\mu}}^{(n)}$ be the minimizers of the empirical objective $\widetilde{L}_n(\boldsymbol{\mu}) = \frac{1}{n}\sum_i \min_k |\mu_k + \widehat{f}(x_i) - y_i|^2$.

Theorem 10 (Clustering Consistency) *Under the assumptions of Theorem 8, the minimizers $\{\widehat{\boldsymbol{\mu}}^{(n)}\}_n$ converge in probability to the minimizers of the population objective*

$$L_*(\boldsymbol{\mu}) = \int \min_{k\in[M]} |\mu_k - (g^*(x) + \varepsilon)| \, \mathbb{P}(dx \times d\varepsilon). \tag{21}$$

Furthermore, the misclassification rate between the estimated labels and the optimal labels converges to zero in probability. Here, the optimal labels are nearest assignments of the samples $\{g^*(x_i) + \varepsilon_i\}_{i=1}^n$ to the minimizers of (21).

A detailed proof and precise statement of Theorem 10 are provided in Appendix D. These results theoretically validate the AltMin procedure: the trivial initialization $\widehat{\boldsymbol{g}} = \mathbf{0}$ yields a consistent f^*, which in turn allows the clustering step to recover the true parameters of the uncontaminated k-means problem.

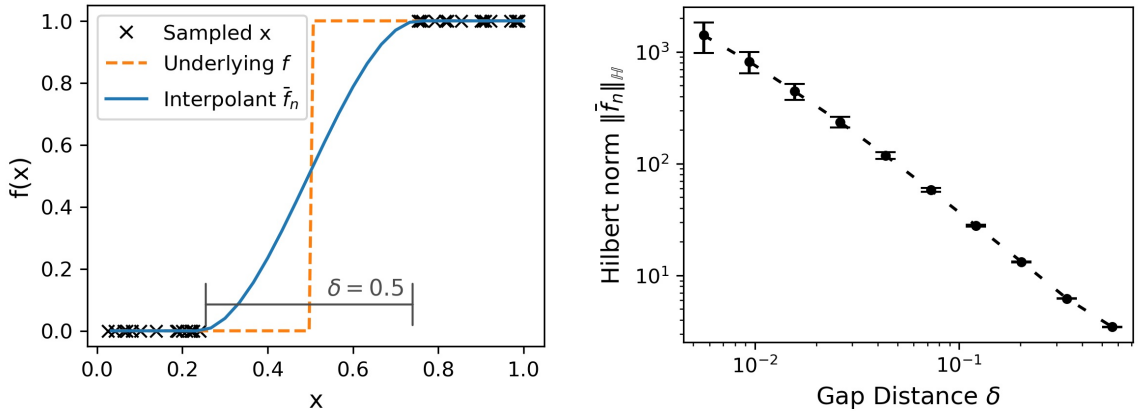
3.2 Spectral Analysis and Filtering

In this section, we provide the technical derivations and intuition behind the consistency results. We analyze the KRR estimator via a bias-variance decomposition that explicitly accounts for the misspecification introduced by g^*.

Let $\Gamma_\tau = \Lambda(\Lambda + \tau I)^{-1}$ be the spectral filter operator associated with KRR. The MSE of the estimator $\widehat{\boldsymbol{f}}$ defined in (18) decomposes as:

$$\begin{aligned}\frac{1}{n}\mathbb{E}_{\boldsymbol{\varepsilon}}\|\boldsymbol{f}^* - \widehat{\boldsymbol{f}}\|_2^2 &= \frac{1}{n}\mathbb{E}_{\boldsymbol{\varepsilon}}\|(I_n - \Gamma_\tau)V^\top \boldsymbol{f}^* - \Gamma_\tau V^\top \boldsymbol{g}^* - \Gamma_\tau V^\top \boldsymbol{\varepsilon}\|_2^2 \\ &\le \underbrace{\frac{2}{n}\|(I_n - \Gamma_\tau)\breve{\boldsymbol{f}}\|_2^2}_{\text{Bias}(f^*)} + \underbrace{\frac{2}{n}\|\,\Gamma_\tau \breve{\boldsymbol{g}}\,\|_2^2}_{\text{Leakage}(g^*)} + \underbrace{\frac{\sigma^2}{n}\operatorname{tr}(\Gamma_\tau^2)}_{\text{Variance}}\end{aligned} \tag{22}$$

where $\breve{\boldsymbol{f}} = V^\top \boldsymbol{f}^*$ and $\boldsymbol{f}^* = \big(f^*(x_1), \ldots, f^*(x_n)\big)$. The first and third terms are standard KRR bias and variance, which vanish as $n \to \infty$ for appropriate $\tau = \tau_n$. The critical term is



(a) RKHS interpolant with pathological sampling.

(b) $\mathbb{H}$-norm as distance to discontinuity decreases.

Figure 3: Spectral drift illustration: As sampled points $\{x_i\}_i$ approach a discontinuity of g^*, the Hilbert norm of the interpolant diverges, indicating energy shift to high frequencies.

the middle "Leakage" term:

$$\frac{1}{n}\|\Gamma_\tau \breve{\boldsymbol{g}}\|_2^2 = \frac{1}{n}\sum_{i=1}^{n} \underbrace{\frac{\lambda_i^2}{(\lambda_i+\tau)^2}}_{h(\lambda_i;\tau)} \breve{g}_i^2. \tag{23}$$

This term represents the portion of the step signal g^* that "leaks" into the estimate of f^*. For consistency, the KRR operator must act as a low-pass filter, suppressing $\boldsymbol{g}^*$. This occurs effectively only if the energy of $\boldsymbol{g}^*$ is concentrated in the high-frequency components (small eigenvalues) where $h(\lambda_i;\tau) \approx (\lambda_i/\tau)^2$ is small.

Spectral Drift. We first observe that step functions g^* are not elements of the continuous RKHS $\mathbb{H}$. Consequently, their projection onto the RKHS basis exhibits a phenomenon we call *spectral drift*.

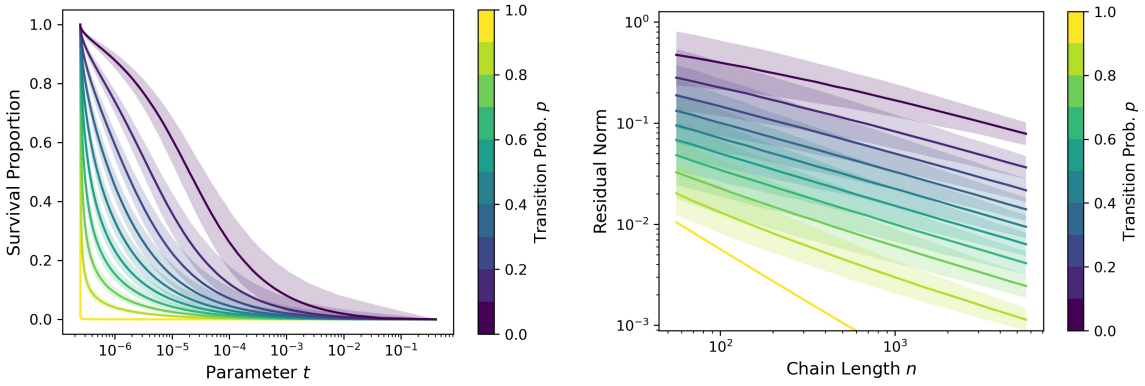
Proposition 11 *Let $\mathcal{H}$ be an RKHS of continuous functions with a continuous kernel. Let $\bar{f}_n$ be the minimum $\mathbb{H}$-norm interpolant of g^* on $\{x_i\}_{i=1}^n$, i.e.,*

$$\bar{f}_n = \operatorname*{argmin}_{f\in\mathbb{H}:\, f(x_i)=g^*(x_i)} \|f\|_{\mathbb{H}}.$$

If $g^* \notin \mathbb{H}$, and the sampling points $\{x_i\}$ are dense in the domain, then the norm of the interpolant diverges:

$$\|\bar{f}_n\|_{\mathbb{H}}^2 = \frac{1}{n}\sum_{i=1}^{n} \frac{\breve{g}_i^2}{\lambda_i} \to \infty \quad \text{as } n \to \infty. \tag{24}$$

(See Appendix B for proof). Proposition 11 implies that the energy $\breve{g}_i^2$ of a step function cannot decay too rapidly relative to λ_i. In fact, significant energy must reside in the tail of the spectrum (high indices).



(a) Survival functions of $\breve{\boldsymbol{g}}$ for length $n = 1000$.

(b) Residual norm $\frac{1}{n}\|\Gamma_\tau \breve{\boldsymbol{g}}\|_2^2$ for $\tau_n = \sqrt{\lambda_n}$.

Figure 4: Experiment results for the 2-state, p-probability Markov chain. 10000 chains were simulated for each $p \in \{k/10\}_{k=1}^{10}$. Shown in subfigures are median results with 95% probability intervals shaded in the corresponding colors. In the case of $p = 1$, there is no shading.

Figure 3 illustrates the necessity of the density assumption using a Sobolev-2 RKHS and a step function g^* discontinuous at 0.5. If samples are adversarially constructed to maintain a gap δ around the discontinuity, the interpolant norm remains bounded. However, as the gap closes ($\delta \to 0$), the norm diverges (Figure 3(b)), consistent with the proposition. Under standard i.i.d. sampling, the sequence is almost surely dense, ensuring this drift occurs.

Effective Filtering. While spectral drift guarantees energy is pushed toward higher frequencies, it is not sufficient on its own to ensure the leakage term (23) vanishes. It is theoretically possible for a fixed fraction of energy to remain in low frequencies. The spectral tail condition (Equation (20)) and the assumption $r_n \to \infty$ ensure that the mass of g^* is pushed sufficiently far into the tail such that the filter $h(\lambda_i; \tau)$ can eliminate it.

The following proposition quantifies this decay.

Proposition 12 (Filtering Bound) *Let $\xi_n := \lambda_{r_n}/\tau_n$. Under the tail condition (20), the leakage term satisfies:*

$$\frac{1}{n}\sum_{i=1}^{n} h(\lambda_i; \tau_n)\, \breve{g}_i^2 \;\lesssim\; \max\{\xi_n^2,\, \xi_n^{\beta_n}\}. \tag{25}$$

Since $\lambda_{r_n} \to 0$ (as $r_n = \Omega(n)$ and eigenvalues decay), we can choose τ_n such that $\xi_n = o(1)$, driving the leakage term to zero. For example, if $\beta_n \geq 2$, the convergence rate is $O((\lambda_n/\tau_n)^2)$.

To build intuition for the spectral tail decay β_n, we analyze the behavior of the leakage term for a specific class of stochastic step signals.

Example 3 (Markov Chain Step Signals) *Consider a step signal $\boldsymbol{g}^* \in \{-1, 1\}^n$ generated by a 2-state Markov chain with transition probability p. We estimate f^* using the min-kernel $\mathcal{K}(x, x') = \min(x, x')$, corresponding to the Sobolev-1 RKHS.*

Intuitively, chains with higher p produce "rougher" signals (more frequent steps), resulting in energy concentrated in higher frequencies (larger β_n). This is corroborated in Figure 4: as p increases, the spectral survival function $S_{g^*}(t)$ decays more sharply (Fig. 4a), leading to steeper decay in the residual norm (Fig. 4b).

For the min-kernel, $\lambda_n \approx (4n)^{-1}$. With regularization $\tau_n = \sqrt{\lambda_n}$, the fastest rate guaranteed by Proposition 12 is $O(\xi_n^2) = O(n^{-1})$. This linear slope is empirically attained in the log-log plot of Figure 4b for $p = 1$, providing evidence that the spectral conditions required for consistency are met by general stochastic step signals.

Synthesis. Combining the decomposition (22) with Proposition 12 proves Theorem 8. The spectral drift (Prop. 11) motivates why g^* lives in the high frequencies, and the tail bound (Prop. 12) ensures the KRR regularization parameter τ can be tuned to filter g^* out while preserving f^*.

Remark 13 *Although we analyzed the initialization $\widehat{\boldsymbol{g}} = \mathbf{0}$, the result holds for any initial estimate $\widehat{\boldsymbol{g}}$ provided the difference $\boldsymbol{g}^* - \widehat{\boldsymbol{g}}$ satisfies the spectral tail conditions. Since $\boldsymbol{g}^*$ is rough and any reasonable initialization is likely regular or zero, the difference retains the rough, high-frequency characteristics necessary for filtering.*

4. Experiments

We now provide experimental results on the performance of the AltMin algorithm. First, we consider simulated data from an M-class data generating process on $\mathcal{X} = [0, 1]$ where data $X^{(n)} = \{i/n\}_{i=1}^n$ is equispaced, cluster labels $z_i^* \sim \mathrm{Unif}([M])$ are uniformly distributed, and the step and smooth components follow

$$\mu_k^* = k - \frac{M+1}{2}, \qquad f_\beta^*(x) = \frac{3}{4}\sin(2\pi\beta x). \tag{26}$$

The min kernel from Example 3 was chosen for estimation due to its sinusoidal eigenfunctions. Given the equispaced data, the smallest radius ρ that guarantees the connectivity condition of Theorem 5 is

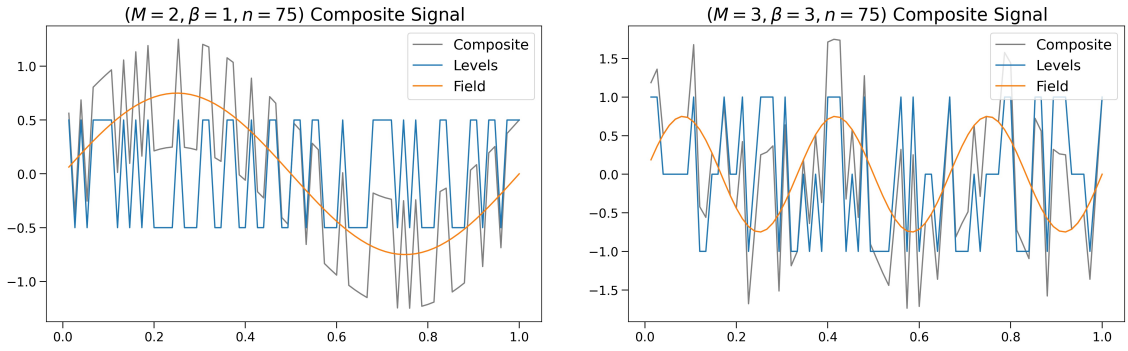
$$\min_{i \neq j} d_{\mathcal{K}}(x_i, x_j) = \sqrt{(i+1)/n - 2i/n + i/n} = n^{-1/2},$$

where the kernel-metric $d_{\mathcal{K}}(x, x')$ was defined in Example 2. The Hilbert-norm of f_β^* can be computed using inner product $\langle f, g\rangle_{\mathbb{H}^1} = \int_0^1 \partial_x f(x)\, \partial_x g(x)\, dx$. Evaluating this norm gives the following worst-case bound on the modulus of continuity of f_β^*,

$$\omega(\rho_{\min}) \leq \|f_\beta^*\|_{\mathbb{H}^1} \cdot \rho_{\min} \leq \frac{3\sqrt{2}}{4}\pi\beta \cdot n^{-1/2}. \tag{27}$$

Finally, a noisy recovery setting will be considered where i.i.d. noise $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$ is added to mixed observations $f_\beta^*(x_i) + \mu_{z_i^*}^*$.

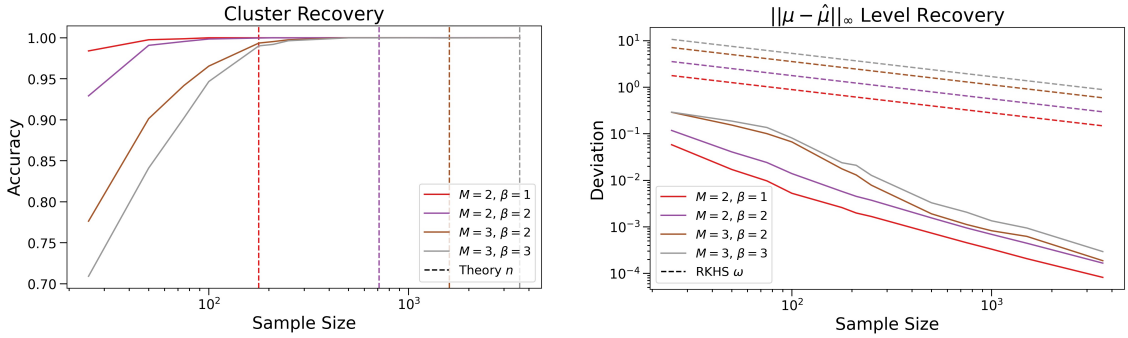
In both recovery settings, sample size is grown in roughly exponential manner starting from $n = 25$ to $n = 3600$. At each sample size n, a total of 100 datasets $(X^{(n)}, \boldsymbol{y})$ were simulated. Accuracy and deviation results at each n were calculated using the mean score of the 100 datasets.



(a) Smoother field

(b) Rougher field

Figure 5: Signal, field and composite observation simulated from (26) for two and three classes. Here, a composite observation is the pointwise sum between levels and field.



(a) (Noiseless) Classification curves

(b) (Noiseless) Deviation curves

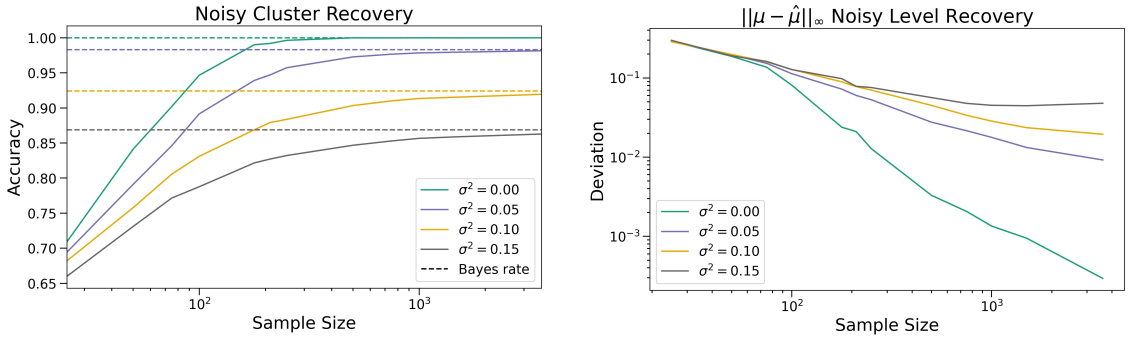
Figure 6: AltMin recovery results for a noiseless simulated setting. Worst-case theory bounds are shown as dashed lines for each of the different settings.

4.1 Simulation Experiments

Four settings were considered for noiseless recovery: $(M, \beta) \in \{(2,1), (2,2), (3,2), (3,3)\}$. Cluster recovery and deviation results for the four settings can be found in Figure 6. For each setting of the optimization problem (8), worst-case recovery bounds, shown dashed in Figure 6, were calculated using Theorem 5 and Proposition 6. The AltMin algorithm stays well within these worst-case bounds, demonstrating the effectiveness of the simple blockwise updates for specific problem settings.

For noisy recovery, the setting with $M = \beta = 3$ was considered at noise levels $\sigma^2 \in \{0, 0.05, 0.1, 0.15\}$. Cluster recovery and deviation results for these four settings can be found in Figure 7. In each of the noisy settings, AltMin approaches the Bayes error of what is expected for a perfect classifier.

We note that the rate at which AltMin approaches Bayes error seems faster for the cases where σ^2 is low. This may suggest that the AltMin algorithm is well-suited for smooth field, cluster recovery problems which experience low amounts of background noise.



(a) (Noisy) Classification curves

(b) (Noisy) Deviation curves

Figure 7: AltMin recovery results for a noisy simulated setting. Bayes error rates for classification are shown as dashed lines for the various noise levels.

4.2 MRI Decontamination

For application, we return to the motivating MRI bias field problem. This is a real-world example where the magnitude of the inhomogeneity f^* and the tissue intensity g^* are much larger than the scale of the background noise σ^2 (Asher et al., 2010). As we have seen in Section 4.1, this is a type of problem which is a good candidate for the AltMin algorithm.

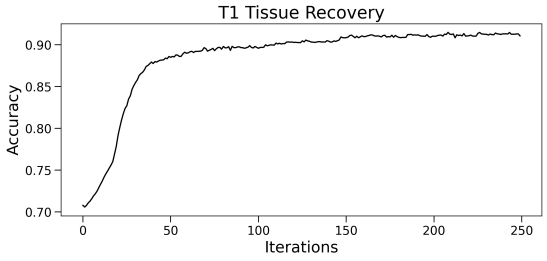
To make our experiment quantitative, we consider a 4-class, strongly-biased variant of the BrainWeb (Cocosco et al., 1997) phantom. The field estimation step (14) is carried out using a Python spline routine `csaps` (Prilepin, 2023). This routine uses an RKHS tensor product of univariate smoothing splines to fit the multidimensional data. Relevant `csaps` smoothing parameters were selected using a post-fitting process. In practice, smoothing parameters would be selected using a validation set of data which corresponds to a specific coil cluster or MRI scanner.

For implementation, we consider modeling the bias field for both single sequence and multi-sequence scans. In a multi-sequence scan, it is understood that the bias field does not vary much between sequences (Belaroussi et al., 2006). For this reason, we consider the following general p-sequence data model

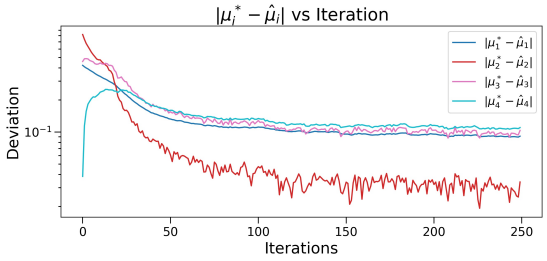
$$\boldsymbol{y}(x) = f^*(x) \cdot \boldsymbol{\mu}^*(x), \quad \text{for } x \in \mathcal{X}$$

where levels $\boldsymbol{\mu}^*(x)$ take value in $\mathbb{R}^p$ and bias $f^*(x)$ is still a scalar function.

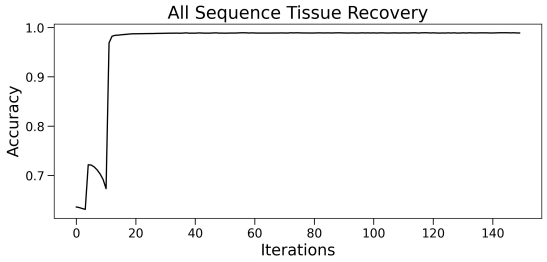
Bias field and tissue decomposition results for the single and multi-sequence setting can be found in Figure 9 with the respective AltMin optimization results found in Figure 8. The presence of redundant sequencing data, albeit at different intensity scalings, seems to significantly improve AltMin convergence as shown in Figures 8c-8d. This also translate to an improved performance, as many of the anomalous tissue patches seen in Figure 9a no longer occur in Figure 9b. Additional experiments comparing AltMin to other medical debiasing methods can be found in Appendix C.



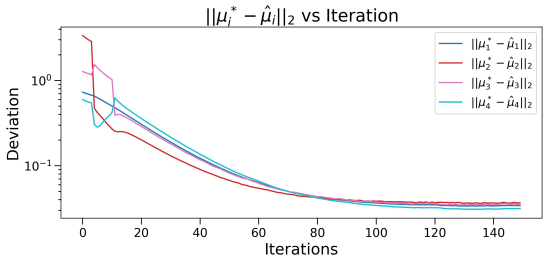
(a) (Single) Accuracy vs. iteration



(b) (Single) Deviation vs. iteration

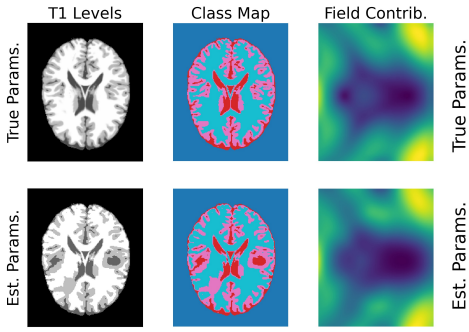


(c) (Multi) Accuracy vs. iteration

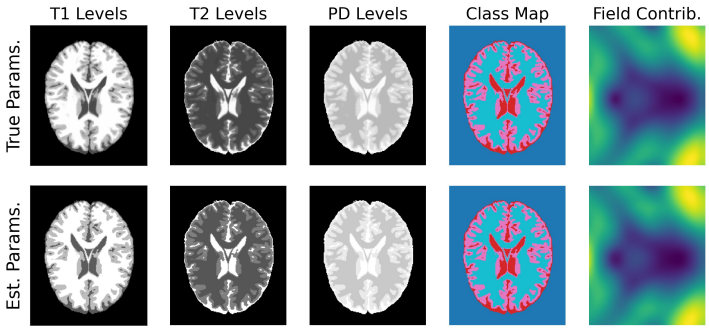


(d) (Multi) Deviation vs. iteration

Figure 8: Cluster and level accuracy of the AltMin algorithm on the biased BrainWeb phantom. Final accuracies for single and multi-sequence settings are 91.07% and 98.91% respectively. Level deviations in the multi-sequence setting are calculated with respect to the vector 2-norm.



(a) Single sequence recovery



(b) Multi-sequence recovery

Figure 9: AltMin decomposition for the biased BrainWeb dataset. Class maps of the single sequence setting show anomalous tissue patches in areas where the field changes most rapidly.

5. CONCLUSION

In this paper, we defined the problem of composite signal decomposition for continuous contaminants and step-wise signals. We outlined recovery conditions that leverage the local and global topology of the data including: connectivity, minimum true level deviation, and the degree of oscillation of the contaminant. These quantities are natural, and their roles in recovery intuitively clear, allowing for a high-level understanding to be easily derived from our theoretical finding.

Besides identifiability, we developed a practical algorithm AltMin for handling contaminants that reside within an RKHS. This algorithm can be viewed as an extension of both kernel ridge regression (KRR) and k-means, with updates to each being performed alternately. MSE bounds for the algorithm were provided in terms of the spectral properties of the data, leading to a "one-step" consistency result in the large sample limit.

We evaluated AltMin empirically on both simulated and real-world data. In the case of simulated data, AltMin operated well within the worst-case theory bounds outlined in Section 2. When the data was further corrupted by noise, AltMin approached the best possible classification rates for the given data generating process. In the real-world study, we conducted an MRI tissue recovery experiment, illustrating how tensor products of smoothing splines can be employed to estimate contaminant MRI bias fields. Given redundant data on the same bias field, AltMin significantly enhanced clustering performance and overall optimization stability.

These empirical studies, alongside the identifiability theory of Section 2, suggest that step and smooth decompositions are attainable within worst-case optimality guarantees. Regarding application, the alternating optimization of AltMin appears well-suited for data-dense tasks, especially when data is spatially uniform and low in noise. In this context, decomposition problems akin to MRI multi-sequence recovery could be promising avenues for further applications of the AltMin algorithm.

References

Kambiz A Asher, Neal K Bangerter, Ronald D Watkins, and Garry E Gold. Radiofrequency coils for musculoskeletal magnetic resonance imaging. *Top. Magn. Reson. Imaging*, 21(5): 315–323, October 2010.

Sohail Bahmani and Justin Romberg. Near-optimal estimation of simultaneously sparse and low-rank matrices from nested linear measurements. *Information and Inference: A Journal of the IMA*, 5(3):331–351, 05 2016.

Boubakeur Belaroussi, Julien Milles, Sabin Carme, Yue Min Zhu, and Hugues Benoit-Cattin. Intensity non-uniformity correction in MRI: Existing methods and their validation. *Medical Image Analysis*, 10(2):234–246, 2006.

Peter J Bickel, Ya'acov Ritov, and Alexandre B Tsybakov. Simultaneous analysis of Lasso and Dantzig selector. *The Annals of statistics*, 37(4):1705–1732, 2009.

Andrea Braides. *Gamma-Convergence for Beginners*. Oxford University Press, 07 2002.

Venkat Chandrasekaran, Sujay Sanghavi, Pablo A. Parrilo, and Alan S. Willsky. Sparse and low-rank matrix decompositions. In *2009 47th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 962–967, 2009.

Chris A. Cocosco, Vasken Kollokian, Remi K.-S. Kwan, G. Bruce Pike, and Alan C. Evans. BrainWeb: Online interface to a 3D MRI simulated brain database. *NeuroImage*, 5:425, 1997.

David Donoho and Gitta Kutyniok. Microlocal analysis of the geometric separation problem. *Communications on Pure and Applied Mathematics*, 66(1):1–47, 2013.

Aapo Hyvarinen, Hiroaki Sasaki, and Richard Turner. Nonlinear ICA using auxiliary variables and generalized contrastive learning. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pages 859–868. PMLR, 16–18 Apr 2019.

Aapo Hyvärinen and Petteri Pajunen. Nonlinear independent component analysis: Existence and uniqueness results. *Neural Networks*, 12(3):429–439, 1999.

Zheng Tracy Ke, Jianqing Fan, and Yichao Wu. Homogeneity pursuit. *Journal of the American Statistical Association*, 110(509):175–194, 2015.

Ilyes Khemakhem, Diederik Kingma, Ricardo Monti, and Aapo Hyvarinen. Variational autoencoders and nonlinear ICA: A unifying framework. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 2207–2217. PMLR, 26–28 Aug 2020.

Seung-Jean Kim, Kwangmoo Koh, Stephen Boyd, and Dimitry Gorinevsky. ℓ_1 trend filtering. *SIAM Review*, 51(2):339–360, 2009.

Yunho Kim and Hemant D. Tagare. Intensity nonuniformity correction for brain mr images with known voxel classes. *SIAM Journal on Imaging Sciences*, 7(1):528–557, 2014.

Ken Lang. Newsweeder: Learning to filter netnews. In *Proceedings of the Twelfth International Conference on Machine Learning*, pages 331–339, 1995.

Jüri Lember. On minimizing sequences for k-centres. *Journal of Approximation Theory*, 120 (1):20–35, 2003.

Leland McInnes, John Healy, Nathaniel Saul, and Lukas Großberger. Umap: Uniform manifold approximation and projection. *Journal of Open Source Software*, 3(29):861, 2018.

Eugene Prilepin. csaps. `https://github.com/espdev/csaps`, 2023.

Jared Tanner and Simon Vary. Compressed sensing of low-rank plus sparse matrices. *Applied and Computational Harmonic Analysis*, 64:254–293, 2023.

Matthew Thorpe, Florian Theil, Adam M. Johansen, and Neil Cade. Convergence of the k-means minimization problem using Γ-convergence. *SIAM Journal on Applied Mathematics*, 75(6):2444–2474, 2015.

Ryan J. Tibshirani. Adaptive piecewise polynomial estimation via trend filtering. *The Annals of Statistics*, 42(1):285 – 323, 2014.

Alexandre B Tsybakov. *Introduction to Nonparametric Estimation*. Springer, 2009.

Nicholas J. Tustison, Brian B. Avants, Philip A. Cook, Yuanjie Zheng, Alexander Egan, Paul A. Yushkevich, and James C. Gee. N4ITK: Improved N3 bias correction. *IEEE Transactions on Medical Imaging*, 29(6):1310–1320, 2010.

Luciano Vinas, Arash A. Amini, Jade Fischer, and Atchar Sudhyadhom. LapGM: A multisequence MR bias correction and normalization model, 2022.

Uro Vovk, Franjo Pernus, and Botjan Likar. A review of methods for correction of intensity inhomogeneity in MRI. *IEEE Transactions on Medical Imaging*, 26(3):405–421, 2007.

Martin J. Wainwright. Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (Lasso). *IEEE Transactions on Information Theory*, 55(5):2183–2202, 2009.

Martin J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2019.

Yujia Zheng, Ignavier Ng, and Kun Zhang. On the identifiability of nonlinear ICA: Sparsity and beyond. In *Advances in Neural Information Processing Systems*, 2022.

Appendix A. Identifiability Proofs

Any optimal candidate solution $(\widehat{f}, \widehat{\mu}, \widehat{z})$ to (8) which is fit to data $\{y_i\}$ generated from (6) must satisfy

$$f^*(x_i) + \mu^*_{z_i^*} = \widehat{f}(x_i) + \widehat{\mu}_{\widehat{z}_i}, \quad \text{for all } i \in [n]. \tag{28}$$

Since $\widehat{f} - f^* \in \mathcal{F}_{2\omega}(\mathcal{X})$, we may instead analyze the discrepancy

$$g(x_i) = \mu^*_{z_i^*} - \widehat{\mu}_{\widehat{z}_i}, \quad \text{for all } i = 1, \dots, n,$$

for $g \in \mathcal{F}_{2\omega}(\mathcal{X})$. In addition, we will assume that function $g : \mathcal{X} \to \mathcal{Y}$ takes values on a normed vector space $(\mathcal{Y}, \|\cdot\|)$. As a result, the modulus of continuity ω will be related to the induced norm-metric as $\|g(x) - g(x')\| \le \omega\big(d(x, x')\big)$.

The following result is the main ingredient in the proof of Theorem 5:

Theorem 14 *Suppose for $g \in \mathcal{F}_{2\omega}(\mathcal{X})$ we have $g(x_i) = \mu^*_{z_i^*} - \widehat{\mu}_{\widehat{z}_i}$ for all $i \in [n]$ where $\boldsymbol{z}^* = (z_i^*)$ and $\widehat{\boldsymbol{z}} = (\widehat{z}_i)$ both belong to $[M]^n$. Assume the following holds:*

(a) $\|\mu_k^* - \mu_\ell^*\| \ge \gamma$ for all $k \ne \ell$.

(b) $G_\rho(X)$ is connected for some ρ with $2\,\omega(\rho) < \gamma/M$.

Then for all $i, j \in [n]$ we have

$$\widehat{z}_i = \widehat{z}_j \implies z_i^* = z_j^*. \tag{29}$$

Proof Start by considering the induction hypothesis that, for any path $\mathcal{P} \subseteq G_\rho(X)$ of length T, all element pairs $i, j \in \mathcal{P}$ satisfy (29). The base case of $T = 0$ holds trivially with $i = j$.

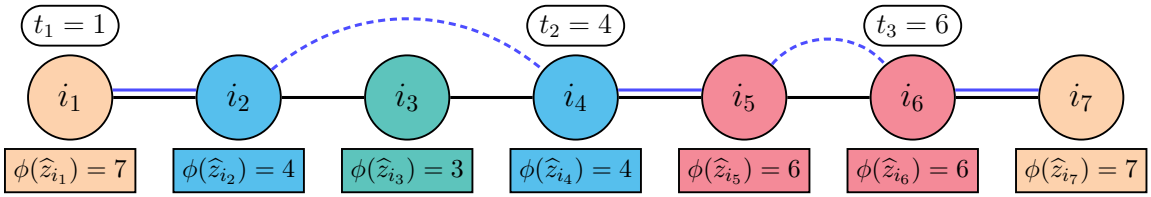


Figure 10: Example use of the $\phi(r)$ function and $\{(u_q, v_q)\}_{q=1}^{Q}$ sequence, shown in blue above, for a 4-class path of length 7. The colors denote the estimated cluster labels. This example has $\{(u_q, v_q)\}_{q=1}^{Q} = \{(i_1, i_2), (i_4, i_5), (i_6, i_7)\}$ with $Q = 3$.

Throughout the proof, by the label of a node i, we mean its estimated label $\widehat{z}_i$. Consider a general path $\mathcal{P} = \{i_t\}_{t=1}^{T+1}$ of length $T+1$ inside $G_\rho(X)$. As both $\{i_t\}_{t=1}^{T}$ and $\{i_t\}_{t=2}^{T+1}$ are paths of length T, we only need to verify (29) for i_1 and i_{T+1}. Therefore, for our induction step it is sufficient to show that $\widehat{z}_{i_1} = \widehat{z}_{i_{T+1}}$ and $z^*_{i_1} \neq z^*_{i_{T+1}}$ cannot simultaneously hold for the given assumptions (a) and (b).

For the sake of contradiction, assume $\widehat{z}_{i_1} = \widehat{z}_{i_{T+1}}$ and $z^*_{i_1} \neq z^*_{i_{T+1}}$. Under this assumption the induction hypothesis guarantees

$$\widehat{z}_{i_t} \neq \widehat{z}_{i_1} \quad \text{for } 1 < t < T+1. \tag{30}$$

Note that if this was not the case with

$$\widehat{z}_{i_1} = \widehat{z}_{i_t} = \widehat{z}_{i_{T+1}} \quad \text{for some } 1 < t < T+1,$$

then the condition $z^*_{i_1} \neq z^*_{i_{T+1}}$ would have caused a contradiction at the earlier induction step $\max\{(T+1) - t, t - 1\}$.

Next let $\mathcal{R}$ be the set of labels $\widehat{z}_{i_t}$ on path $\mathcal{P}$. Function $\phi(r)$ will be the index of the last node we see on the path from i_1 to i_{T+1} that has label r, that is,

$$\phi(r) = \max_{t \in [T+1]} \{t : \ \widehat{z}_{i_t} = r\}.$$

We construct an edge sequence $\{(u_q, v_q)\}_{q=1}^{Q}$—where Q is determined by the construction—recursively as follows: Let $(u_1, v_1) = (i_1, i_2)$ and for $q = 2, \ldots, Q$,

$$(u_q, v_q) = (i_{t_q}, i_{t_q+1}) \quad \text{where} \quad t_q = \phi(\widehat{z}_{v_{q-1}}).$$

The construction continues until $t_Q = T$, so that $(u_Q, v_Q) = (i_T, i_{T+1})$. See Figure 10 for a concrete example. By construction, the labels of v_{q-1} and u_q are the same, while the labels of v_{q-1} and v_q are necessarily different. By this latter property, the labels of $v_1, \ldots, v_{Q-1}$ are distinct elements of $\mathcal{R}$. The added uniqueness condition of (30) gives that the label of v_Q is also distinct from $v_1, \ldots, v_{Q-1}$, hence $Q \leq |\mathcal{R}|$.

Using $\widehat{z}_{v_{q-1}} = \widehat{z}_{u_q}$, we obtain the decomposition

$$\widehat{\mu}_{\widehat{z}_{u_1}} - \widehat{\mu}_{\widehat{z}_{v_Q}} = \sum_{q=1}^{Q} (\widehat{\mu}_{\widehat{z}_{u_q}} - \widehat{\mu}_{\widehat{z}_{v_q}}). \tag{31}$$

From the induction hypothesis, $\widehat{z}_{v_{q-1}} = \widehat{z}_{u_q}$ implies $z^*_{v_{q-1}} = z^*_{u_q}$ for $2 \le q \le Q$. This gives the decomposition

$$\mu^*_{z^*_{u_1}} - \mu^*_{z^*_{v_Q}} = \sum_{q=1}^{Q} (\mu^*_{z^*_{u_q}} - \mu^*_{z^*_{v_q}}). \tag{32}$$

Moreover, since u_q and v_q are adjacent on the path, they satisfy $d(x_{u_q}, x_{v_q}) \le \rho$, which by assumption (b) implies

$$\|(\mu^*_{z^*_{u_q}} - \mu^*_{z^*_{v_q}}) - (\widehat{\mu}_{\widehat{z}_{u_q}} - \widehat{\mu}_{\widehat{z}_{v_q}})\| = \|g(x_{u_q}) - g(x_{v_q})\| < \gamma/M. \tag{33}$$

By assumption, $\widehat{z}_{u_1} = \widehat{z}_{v_Q}$, hence the LHS of (31) is zero. Then, subtracting decomposition (31) from (32) and using the triangle inequality, we get

$$\|\mu^*_{z^*_{u_1}} - \mu^*_{z^*_{v_Q}}\| \le \sum_{q=1}^{Q} \|(\mu^*_{z^*_{u_q}} - \mu^*_{z^*_{v_q}}) - (\widehat{\mu}_{\widehat{z}_{u_q}} - \widehat{\mu}_{\widehat{z}_{v_q}})\| < Q\,\gamma/M,$$

where the second inequality is by (33). If at the same time $z_{i_1} \ne z_{i_{T+1}}$ then $\mu^*_{z^*_{u_1}} \ne \mu^*_{z^*_{v_Q}}$, and by assumption (a), $\gamma \le \|\mu^*_{z^*_{u_1}} - \mu^*_{z^*_{v_Q}}\|$. Hence,

$$\gamma < Q\,\gamma/M \le |\mathcal{R}|\,\gamma/M.$$

Since $|\mathcal{R}| \le M$, we arrive at a contradiction. This completes the induction step. Applying our induction claim to the connected $G_\rho(X)$ completes the proof. ■

Theorem 14 shows that $\widehat{z}$ is a *refinement* of z^*. But since $\widehat{z}$ has at most M classes and z^* has exactly M classes—due to being saturated by assumption—the classes of $\widehat{z}$ should, in fact, coincide with those of z^*. This proves Theorem 5.

Let us now prove Proposition 6. Under the assumptions of Theorem 5, we can relabel the classes of $(\widehat{\boldsymbol{z}}, \widehat{\boldsymbol{\mu}})$ so that $\widehat{\boldsymbol{z}} = \boldsymbol{z}^*$. Then, it follows from (28) that

$$\widehat{f}(x_i) - f^*(x_i) = \mu^*_{z^*_i} - \widehat{\mu}_{z^*_i} \quad \text{for all } i \in [n]. \tag{34}$$

A.1 Proof of Proposition 6

For $\delta \ge \delta_{\mathrm{lbl}}$, the neighbor graph $G_\delta(\mathcal{C})$ is connected such that every $k, \ell \in [M]$ has a series of edges $\{(x_{i_t}, x_{j_t})\}_{t=1}^T$ with $d(x_{i_t}, x_{j_t}) \le \delta$ such that $z^*_{i_1} = k$, $z^*_{j_T} = \ell$ and

$$z^*_{j_{t-1}} = z^*_{i_t} \ne z^*_{j_t}.$$

In particular, the condition $z^*_{i_t} \ne z^*_{j_t}$ ensures $T \le M-1$. Let $\delta = \delta_{\mathrm{lbl}}$ and with the shorthands $g = \widehat{f} - f^*$ and $\Delta_k = \mu^*_k - \widehat{\mu}_k$, we have $g(x_i) = \Delta_{z^*_i}$ for all $i \in [n]$. Then, the following inequality holds for all $k, \ell \in [M]$,

$$\begin{aligned} \|\Delta_k - \Delta_\ell\| &\le \sum_{t=1}^{T} \|g(x_{i_t}) - g(x_{j_t})\| \\ &\le T \cdot 2\omega(\delta_{\mathrm{lbl}}) \le 2(M-1) \cdot \omega(\delta_{\mathrm{lbl}}). \end{aligned}$$

Letting $\pi_k = |\mathcal{C}_k|/n$ be the proportion of class k, then

$$\left\|\Delta_k - \frac{1}{n}\sum_{i=1}^{n} g(x_i)\right\| = \left\|\Delta_k - \sum_{\ell=1}^{M} \pi_\ell \Delta_\ell\right\| \leq \sum_{\ell=1}^{M} \pi_\ell \|\Delta_k - \Delta_\ell\|.$$

Since $\widehat{f}$ is assumed zero-mean, $\frac{1}{n}\|\sum_{i=1}^n g(x_i)\| = \frac{1}{n}\|\sum_{i=1}^n f^*(x_i)\|$. Putting the pieces together, using the triangle inequality and noting that $\sum_\ell \pi_\ell = 1$ finishes the proof.

Appendix B. Supplement to Section 3.1

B.1 Proof of Proposition 11

Since $K(\cdot,\cdot)$ is continuous, all the functions in the RKHS are also continuous with respect to the metric topology of $\mathcal{X}$. Moreover, by the definition of the RKHS, the evaluation functional δ_x, given by $\delta_x f = f(x)$ for any $f \in \mathbb{H}$, is a continuous linear functional on $\mathbb{H}$ relative to $\|\cdot\|_{\mathbb{H}}$ for any $x \in \mathcal{X}$.

We prove the result by contradiction. Assume that $\|\bar{f}_n\|_{\mathbb{H}}$ does not converge to ∞. Then, there is a subsequence of $\bar{f}_n$ that is bounded in $\mathbb{H}$. Without loss of generality, let us pass to this subsequence for simplicity. Hence, we have $\|\bar{f}_n\|_{\mathbb{H}} \leq C$ for some $C > 0$ and all $n \geq 1$. Since $\mathbb{H}$ is a Hilbert space, the closed ball $\{f \in \mathbb{H} : \ \|f\|_{\mathbb{H}} \leq C\}$ is weakly compact. This follows from Kakutani's theorem: in a Banach space, the closed unit ball is weakly compact if and only if the Banach space is reflexive. Thus, there is a subsequence $\{\bar{f}_{n_k}\}_{k\geq 1}$ that weakly converges to some $f \in \mathbb{H}$. In particular, $\delta_x \bar{f}_{n_k} \to \delta_x f$, that is $\bar{f}_{n_k}(x) \to f(x)$ for all $x \in \mathbb{H}$. This implies that for any $i \geq 1$, $\bar{f}_{n_k}(x_i) \to f(x_i)$, and since $g^*(x_i) = \bar{f}_{n_k}(x_i)$, by the definition of the interpolant, it follows that $g^*(x_i) = f(x_i)$ for all $i \geq 1$.

Consider the case where g^* is continuous with respect to the metric topology of $\mathcal{X}$. Since $\{x_i\}_{i\geq 1}$ is a dense subset of the Hausdorff space $(\mathcal{X}, d)$ and since f is continuous, it follows that $g^*(x) = f(x)$ for all $x \in \mathcal{X}$. But this is a contradiction since $g^* \notin \mathbb{H}$ and $f \in \mathbb{H}$.

On the other hand, if g^* is discontinuous at a point $x_0 \in \mathcal{X}$, we can find two subsequences of $\{x_i\}$ converging to x_0, along which g^* converges to two different values. But since f matches g^* on $\{x_i\}$, it means that f converges to different values along those same subsequences. This contradict continuity of f at x_0, proving that we must have $\|\bar{f}_n\|_{\mathbb{H}} \to \infty$.

Finally, it is well-known (Wainwright, 2019, Section 12.5) that the minimum-norm interpolant can be written as

$$\bar{f}_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \bar{\boldsymbol{\alpha}}_i \mathcal{K}(x_i, \cdot)$$

where $\bar{\boldsymbol{\alpha}} = K^{-1}\boldsymbol{g}^*/\sqrt{n}$ and $\boldsymbol{g}^* = (g^*(x_1), \ldots, g^*(x_n))$ and K has entries $K_{ij} = \mathcal{K}(x_i, x_j)/n$. It follows that

$$\|\bar{f}_n\|_{\mathbb{H}}^2 = \bar{\boldsymbol{\alpha}}^T K \bar{\boldsymbol{\alpha}} = (\boldsymbol{g}^*/\sqrt{n})^T K^{-1} (\boldsymbol{g}^*/\sqrt{n}) = \frac{1}{n}\sum_{i=1}^{n} \breve{g}_i^2/\lambda_i, \tag{35}$$

and the proof is complete.

B.2 Proof of Proposition 12

Let $\mathbb{\lambda}$ be the discrete random variable defined by

$$\mathbb{\lambda} = \begin{cases} \lambda_i, & \text{w.p. } \breve{g}_i^2/(n\|\boldsymbol{g}^*\|_\infty^2) \\ 0, & \text{w.p. } 1 - \|\boldsymbol{g}^*\|_2^2/(n\|\boldsymbol{g}^*\|_\infty^2). \end{cases} \tag{36}$$

Further define $\psi(\lambda) = (1 + \tau/\lambda)^{-2}$, then

$$\frac{1}{n}\|\Gamma_\tau \breve{\boldsymbol{g}}\|_2^2 = \sum_{i=1}^n \frac{\breve{g}_i^2}{n}\psi(\lambda_i) \le \|\boldsymbol{g}^*\|_\infty^2 \, \mathbb{E}[\psi(\mathbb{\lambda})].$$

Let $r = r_n$ and $\beta = \beta_n$ as defined in the main text. For brevity we drop the n subscript. Function $\psi(\cdot)$ is non-negative and monotone on $[0, \infty)$ so

$$\begin{aligned} \mathbb{E}[\psi(\mathbb{\lambda})] &= \int_0^\infty \Pr(\psi(\mathbb{\lambda}) > t)\, dt \\ &= \psi(\lambda_r) + \int_{\psi(\lambda_r)}^{\psi(\lambda_1)} \Pr(\mathbb{\lambda} > \psi^{-1}(t))dt \\ &\le \psi(\lambda_r) + \int_{\psi(\lambda_r)}^{\psi(\lambda_1)} \left(\frac{\lambda_r}{\psi^{-1}(t)}\right)^\beta dt \end{aligned} \tag{37}$$

Denote the last right-hand side integral as $\mathcal{I}(\beta)$. Integral $\mathcal{I}(\beta)$ is monotone decreasing with $\mathcal{I}(\beta) < \mathcal{I}(\beta')$ for $0 \le \beta' < \beta$. Next, the inverse function ψ^{-1} can be lower bounded as

$$\psi^{-1}(t) = \frac{\tau}{t^{-1/2} - 1} \ge \tau\, t^{1/2}(1-t)^{-1/2}. \tag{38}$$

Restricting focus to $\beta \in [0, 2)$ and applying (38) to integral $\mathcal{I}(\beta)$ yields

$$\begin{aligned} \mathcal{I}(\beta) &\le (\lambda_r/\tau)^\beta \int_{\psi(\lambda_r)}^{\psi(\lambda_1)} t^{-\beta/2}(1-t)^{\beta/2}\, dt \\ &\le (\lambda_r/\tau)^\beta \int_0^1 t^{-\beta/2}(1-t)^{\beta/2}\, dt \\ &= (\lambda_r/\tau)^\beta \, \frac{\Gamma(1-\beta/2)\,\Gamma(1+\beta/2)}{\Gamma(2)}. \end{aligned}$$

Identity $\Gamma(z) = \Gamma(1+z)/z$ can be used with $\beta \in [0, 2)$ to get

$$\frac{\Gamma(1-\beta/2)\,\Gamma(1+\beta/2)}{\Gamma(2)} \le \frac{2}{2-\beta}.$$

Lastly since $\psi(\lambda) \le (\lambda/\tau)^2$ and $\mathcal{I}(\beta)$ is monotone decreasing in β, we have

$$\frac{1}{n}\|\Gamma_\tau \breve{\boldsymbol{g}}\|_2^2 \le 2\|\boldsymbol{g}^*\|_\infty^2 \cdot \max\left\{(\lambda_r/\tau)^2, \inf_{\beta' \in [0,\beta)} \frac{2}{2-\beta'} \cdot (\lambda_r/\tau)^{\beta'}\right\}. \tag{39}$$

Let $\xi \coloneqq \lambda_r/\tau$. For $\xi < 1$, function $h(x) = \xi^x/(2-x)$ achieves global minimum at $x^* = 2 - (\log \frac{1}{\xi})^{-1}$. This minimum is non-negative for $\xi < e^{-1/2}$. That is, when $\beta = 2$, we have β' approaching 2 as λ_r/τ approaches 0. More specifically, we obtain the following simplification to (39)

$$\frac{1}{n}\|\Gamma_\tau \breve{\boldsymbol{g}}\|_2^2 \le 4\|\boldsymbol{g}^*\|_\infty^2 \log(1/\xi)\, \xi^{2-(\log\frac{1}{\xi})^{-1}}. \tag{40}$$

B.3 Proof of Theorem 8

The leakage term in the MSE decomposition (22) goes to zero by Proposition 12. The bias and variance terms go to zero by assumptions $\tau_n \to 0$ and $n\tau_n \to \infty$, respectively, as the following argument shows.

Recall that $K \in \mathbb{R}^{n\times n}$ is the normalized kernel matrix with entries $\mathcal{K}(x_i, x_j)/n$ and eigenvalues $\{\lambda_i\}_{i=1}^n \geq 0$. Since $\mathcal{K}$ is a Mercer kernel, we have $\sup_x \mathcal{K}(x,x) \leq \kappa^2$ for some κ, hence $\operatorname{tr}(K) \leq \kappa^2$.

The empirical variance term is $V_n(\tau) = \frac{\sigma^2}{n}\sum_{i=1}^n \frac{\lambda_i^2}{(\lambda_i+\tau)^2}$. Using $\frac{\lambda^2}{(\lambda+\tau)^2} \leq \frac{\lambda}{\lambda+\tau} \leq \frac{\lambda}{\tau}$ for $\lambda, \tau > 0$,

$$V_n(\tau) \leq \frac{\sigma^2}{n\tau}\sum_{i=1}^n \lambda_i = \frac{\sigma^2}{n\tau}\operatorname{tr}(K) \leq \frac{\sigma^2\kappa^2}{n\tau}.$$

Therefore $n\tau_n \to \infty$ implies $V_n(\tau_n) \to 0$. The empirical bias term is

$$B_n(\tau) = \frac{2}{n}\sum_{i=1}^n \frac{\tau^2}{(\tau+\lambda_i)^2}\check{f}_i^2.$$

Since by assumption f is in the RKHS, its Hilbert norm is bounded, hence,

$$\frac{1}{n}\sum_{i=1}^n \frac{\check{f}_i^2}{\lambda_i} \leq C, \qquad \text{(with the convention } \check{f}_i^2/\lambda_i = 0 \text{ if } \lambda_i = 0)$$

for some constant C, since the LHS apporaches $\|f\|_{\mathbb{H}}^2$ as $n \to \infty$. Letting $a_i := \check{f}_i^2/\lambda_i$,

$$B_n(\tau) = \frac{2}{n}\sum_{i=1}^n a_i \frac{\tau^2\lambda_i}{(\tau+\lambda_i)^2}.$$

For $\lambda, \tau > 0$, letting $u = \lambda/\tau$ gives

$$\frac{\tau^2\lambda}{(\tau+\lambda)^2} = \tau\frac{u}{(1+u)^2} \leq \frac{\tau}{4},$$

hence

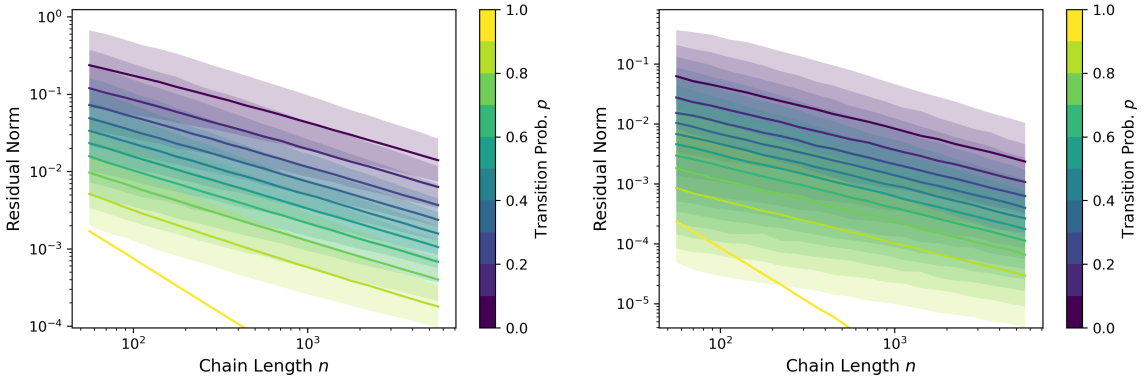
$$B_n(\tau) \leq \frac{\tau}{4}\cdot\frac{2}{n}\sum_{i=1}^n a_i \leq \frac{C}{2}\tau.$$

Therefore $\tau_n \to 0$ implies $B_n(\tau_n) \to 0$. The proof is complete.

B.3.1 Proof of Corollary 9

The Sobolev-α RKHS on $[0,1]$ has kernel defined by inner product

$$\langle f, g\rangle_{\mathbb{H}^\alpha} = \sum_{k=0}^{\alpha}\int_0^1 f^{(k)}(x)\, g^{(k)}(x)\, dx.$$



(a) Sobolev-1 with $\tau = n^{-2/3}$.

(b) Sobolev-2 with $\tau = n^{-4/5}$.

Figure 11: Residual norm decays $\frac{1}{n}\|\Gamma_\tau \breve{\boldsymbol{g}}\|_2^2$, based on the optimal τ_n selections, for different Sobolev-α kernels. Slight numerical inaccuracies are shown in the norm decay of the Sobolev-2 kernel.

This Mercer kernel has eigenvalue decay $\lambda_i = i^{-2\alpha}$. For the standard KRR problem, a minimax optimal selection of the regularization parameter is given by $\tau_n \asymp n^{\frac{-2\alpha}{2\alpha+1}}$. When plugged into the MSE expression,

$$\overline{\mathrm{MSE}}(\tau) = \frac{1}{n}\sum_{i=1}^{n} \frac{\tau^2 \breve{f}_i^2}{(\lambda_i+\tau)^2} + \frac{\sigma^2}{n}\sum_{i=1}^{n} \frac{\lambda_i^2}{(\lambda_i+\tau)^2}, \tag{41}$$

we obtain $\overline{\mathrm{MSE}}(\tau_n) \asymp n^{\frac{-2\alpha}{2\alpha+1}}$ which decays to zero as $n \to \infty$. The resulting rate for $\xi_n = \lambda_n/\tau_n$ is

$$\xi_n \asymp n^{\frac{-4\alpha^2}{2\alpha+1}} = o(1)$$

which satisfies the condition of Theorem 8. Tying back to Example 3, Figure 11 shows norm decay plots for $\frac{1}{n}\|\Gamma_\tau \breve{\boldsymbol{g}}\|_2^2$, when using τ_n minimax selections on the different Sobolev-α kernels.

Similar to the case of the min-kernel in Figure 4, as the signal $\boldsymbol{g}^*$ becomes more rough, i.e. p becomes larger, we see a quicker decay in filtered norm for the different Sobolev examples. Furthermore, these contributions are filtered at a faster rate for Sobolev kernels that are smoother, that is those with larger α values. This faster decay is not only intuitive but expected from our earlier derived ξ_n decay rate.

Appendix C. Additional Experiments

We compare AltMin to other MRI debiasing techniques using the same biased phantom as Section 4.2. For comparison, we consider a standard debiasing technique N4ITK (Tustison et al., 2010) and a Bayesian modeling approach LapGM (Vinas et al., 2022). Hyperparameters for all methods, including AltMin, were selected using the same post-fitting process. Specific to N4ITK, bias estimates were calculated on the T1-sequence information and clusterings were calculated using an additional k-means estimation at the end of the debiasing procedure.

Method	# Seqs.	BrainWeb 4-class		BrainWeb 10-class	
		Acc. [%]	Max Dev. [1]	Acc. [%]	Max Dev. [1]
k-means	1	73.62	8.21×10^{-1}	37.69	7.08×10^{0}
	3	74.38	3.41×10^{0}	44.21	1.19×10^{1}
N4ITK + k-means	1	74.10	1.17×10^{0}	40.14	6.01×10^{0}
	3	74.41	3.91×10^{0}	48.22	1.11×10^{1}
LapGM	1	76.14	2.87×10^{0}	50.16	$\mathbf{2.47 \times 10^{0}}$
	3	87.28	4.08×10^{0}	78.38	$\mathbf{4.19 \times 10^{0}}$
AltMin	1	**91.07**	$\mathbf{1.10 \times 10^{-1}}$	**56.27**	4.49×10^{0}
	3	**98.91**	$\mathbf{3.67 \times 10^{-2}}$	**82.36**	7.84×10^{0}

Table 1: Clustering results for different debiasing methods for single and multi-sequence settings.

Performance of each method for the various recovery settings can be found in Table 1. In each recovery setting, AltMin either meets or exceeds the classification and level accuracies of the other tested methods. We highlight that, for all debias methods, recovery is significantly more difficult in the 10-class setting. Methods which eventually scored well in this setting were those which could effectively leverage multi-sequence information during debias and clustering. This emphasizes the importance of replicated information for practical step and smooth recovery implementations.

Appendix D. AltMin Consistency Through Γ-Convergence Techniques

The AltMin algorithm is a blockwise optimization procedure that iteratively estimates parameter intermediaries through kernel ridge regression (KRR) and k-means update steps. Under the appropriate assumptions, each update step can be shown to be consistent with respect to their own parameter subset. However, a question remains of whether AltMin can achieve consistency in case of perturbed optimization steps.

In-sample consistency for the perturbed KRR step was already shown in Section 3.1 for samples with x-variates belonging to the set

$$\mathcal{B} \coloneqq \big\{(x_i)_{i=1}^{\infty} : \liminf_{n\to\infty} r_n/n > 0 \text{ and } \liminf_{n\to\infty} \beta_n > 0\big\}. \tag{42}$$

When it is clear from context we will let $\mathcal{B} \coloneqq \mathcal{B} \times \mathbb{R}^{\infty}$, since event $\mathcal{B}$ does not depend on noise ε. Let $\mathrm{MSE}_n : ((x_i)_{i=1}^{\infty}, (\varepsilon_i)_{i=1}^{\infty}) \mapsto \mathrm{MSE}(\boldsymbol{f}^*, \widehat{\boldsymbol{f}})$ denote the n-sample loss for finite sequence $(x_i, \varepsilon_i)_{i=1}^{n}$. Then, through Markov's inequality

$$\begin{aligned}
\lim_{n\to\infty} \mathbb{P}\big(\{\mathrm{MSE}_n(x,\varepsilon) \geq \delta\} \cap \mathcal{B}\big) &\leq \delta^{-1} \lim_{n\to\infty} \mathbb{E}_{(x,\varepsilon)}[\mathrm{MSE}_n(x,\varepsilon) \cdot 1_{\mathcal{B}}(x)] \\
&\leq \delta^{-1} \lim_{n\to\infty} \mathbb{E}_x[\mathbb{E}_{\boldsymbol{\varepsilon}}[\mathrm{MSE}(\boldsymbol{f}^*, \widehat{\boldsymbol{f}})] \cdot 1_{\mathcal{B}}(x)] \\
&\lesssim \delta^{-1} \lim_{n\to\infty} \mathbb{E}_x[e_n(x)]
\end{aligned}$$

where $e_n(x)$ is an error term which is almost-surely bounded and tending to zero. Then by dominated convergence theorem, for every $\delta > 0$,

$$\lim_{n\to\infty} \mathbb{P}\big(\{\mathrm{MSE}_n(x,\varepsilon) \geq \delta\} \cap \mathcal{B}\big) = 0 \tag{43}$$

Similarly, we will show that the AltMin clustering parameters, $(\widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{z}})$, converge to population minima in probability. More rigorously, let $L_* : \mathbb{R}^M \to \mathbb{R}$ be the population k-means clustering objective defined by

$$L_*(\boldsymbol{\mu}) = \int \min_{k\in[M]} |\mu_k - (g^*(x) + \varepsilon)|^2 \, \mathbb{P}(dx \times d\varepsilon). \tag{44}$$

The minima of L_*, defined formally as the following set,

$$\mathcal{M} \coloneqq \{\overline{\boldsymbol{\mu}} \in \mathbb{R}^M : \, L_*(\overline{\boldsymbol{\mu}}) = \inf_{\boldsymbol{\mu}\in\mathbb{R}^M} L_*(\boldsymbol{\mu})\} \tag{45}$$

each admit a nearest label partition $\mathcal{V} = (\mathcal{V}_k)_{k=1}^M$, where

$$\mathcal{V}_k \coloneqq \{u \in \mathbb{R} : \, \operatorname*{argmin}_{\ell\in[M]} |\overline{\mu}_\ell - u| = k\} \tag{46}$$

and a nearest label sequence $\overline{z}_i = \operatorname{argmin}_{k\in[M]} |\overline{\mu}_k - (g^*(x_i) + \varepsilon_i)|$ for data pairs (x_i, ε_i).

Individual deviations from the minima set $\mathcal{M} \subseteq (\mathbb{R}^M, \|\cdot\|)$ can be calculated by

$$\operatorname{Dist}(\boldsymbol{\mu}, \mathcal{M}) \coloneqq \inf_{\overline{\boldsymbol{\mu}}\in\mathcal{M}} \|\boldsymbol{\mu} - \overline{\boldsymbol{\mu}}\| \tag{47}$$

and sample misclassification relative to a sequence $\overline{z}$ can be calculated as

$$\operatorname{Miss}_n(\boldsymbol{z}, \, \overline{z}) = \frac{1}{n}\sum_{i=1}^n 1\{z_i = \overline{z}_i\} \tag{48}$$

This brings us to the following consistency result for the AltMin clustering step:

Theorem 15 *Let $\widehat{f}_n$ be the KRR estimate for $(y_i)_{i=1}^n$ and let $\widehat{\boldsymbol{\mu}}_n$ be a minimizer for*

$$\widetilde{L}_n(\boldsymbol{\mu}) = \frac{1}{n}\sum_{i=1}^n \min_{k\in[M]} |\mu_k + \widehat{f}_n(x_i) - y_i|^2$$

with corresponding label estimates $\widehat{\boldsymbol{z}}_n \in [M]^n$. Suppose $(x_i, \varepsilon_i) \overset{i.i.d.}{\sim} \mathbb{P}$. If all μ_k^* are distinct and $(x_i)_{i=1}^\infty \in \mathcal{B}$, then the minimizer sequence $\{\widehat{\boldsymbol{\mu}}_n\}_n$ converges in probability in the sense

$$\lim_{n\to\infty} \mathbb{P}(\{\operatorname{Dist}(\widehat{\boldsymbol{\mu}}_n, \mathcal{M}) > \delta)\} \cap \mathcal{B}) = 0 \quad \textit{for every } \delta > 0.$$

Furthermore, each converging subsequence $\widehat{\boldsymbol{\mu}}_m \to \overline{\boldsymbol{\mu}} \in \mathcal{M}$ has

$$\lim_{m\to\infty} \operatorname{Miss}_m(\widehat{\boldsymbol{z}}_m, \overline{z}) \to 0,$$

where $\overline{z} = (\overline{z}_i)_{i=1}^\infty$ is the label sequence associated to $\overline{\boldsymbol{\mu}}$ and $(x_i, \varepsilon_i)_{i=1}^\infty$.

That is to say, under event $\mathcal{B}$, AltMin *consistently recovers cluster and nearest label estimates for the uncontaminated k-means clustering problem $\{g(x_n) + \varepsilon_n\}_n$.*

Elaborating on the convergence shown in Theorem 15, a deviation quantity $\Delta_n \in \mathbb{R}$ converges in probability for $\mathcal{B}$ if

$$\lim_{n\to\infty} \mathbb{P}(\{\Delta_n > \delta\} \cap \mathcal{B}) = 0 \quad \text{for every } \delta > 0.$$

Equivalently stated, every subsequence of $\{\Delta_n\}_n$ has a sub-subsquence $\{\Delta_m\}_m$ such that $\Delta_m \to 0$ almost-surely in $\mathcal{B}$, that is,

$$\mathbb{P}\big(\big\{\lim_{m\to\infty} \Delta_m = 0\big\} \cap \mathcal{B}\big) = \mathbb{P}(\mathcal{B}).$$

Armed with these definitions, we begin by fixing any countable, subsequence index set $\mathcal{I} \subseteq \mathbb{N}$. For brevity, we simply refer to $\mathcal{I}$ as an index set. Then, by (43), there exists a sub-index set $\mathcal{I}' \subseteq \mathcal{I}$ such that $\mathrm{MSE}_m(x, \varepsilon) \to 0$ for $m \in \mathcal{I}'$ and $\mathbb{P}$-almost all $(x_i, \varepsilon_i)_{i=1}^\infty \in \mathcal{B}$. Without loss of generality let $\mathcal{I}' = \mathcal{I} = [n]$, it suffices to show

$$\lim_{m\to\infty} \mathrm{Dist}(\widehat{\boldsymbol{\mu}}_m, \mathcal{M}) = 0 \quad \text{for } \mathbb{P}\text{-a.e. } (x_i, \varepsilon_i)_{i=1}^\infty \in \mathcal{B}. \tag{49}$$

for m belonging to some sub-index set $\mathcal{I}'' \subseteq \mathcal{I}'$.

Our convergence result will be proven using techniques belonging to Γ-convergence. To this end, we begin with an overview Γ-convergence and how it pertains to the k-means objective and its minimizers. After this introduction, we show how the AltMin objective maintains key properties of the k-means objective plus a bracketing from the negligible MSE_n term. Lastly, we show how the empirical minimizers of the AltMin's k-means update must converge to population minimizers of the uncontaminated k-means objective.

D.1 Convergence of Empirical Minimizers

In this section, we cover the basics of Γ-convergence and how it pertains to the convergence of empirical minimizers. We refer the reader to Braides (2002) for an overview of Γ-convergence and its applications.

For functionals $F_n : \mathcal{U} \to \mathbb{R}$ defined on a common metric space $\mathcal{U}$, the Γ-convergence of a functional sequence $\{F_n\}_n$ admits the following characterization:

Definition 16 *A sequence of functionals $\{F_n\}_n$ is said to Γ-converge to Γ-limit F_* if*

$$F_*(u) \le \liminf_{n\to\infty} F_n(u_n) \qquad \textit{for every } u_n \to u,$$

and there exists at least one $u_n \to u$ such that

$$\limsup_{n\to\infty} F_n(u_n) \le F_*(u).$$

To properly characterize the behavior of minimizers for $\{F_n\}_n$, some additional regularity is needed regarding the optimization behavior of F_n. In particular, we will be interested in functional sequences that are *equi-mildly coercive*.

Definition 17 *A sequence of functionals $\{F_n\}_n$ defined on a common metric space $\mathcal{U}$ is equi-mildly coercive if there exists a non-empty, compact subset $\mathcal{A} \subset \mathcal{U}$ such that*

$$\inf_{u\in\mathcal{U}} F_n(u) = \inf_{u\in\mathcal{A}} F_n(u) \qquad \textit{for all } n.$$

Equipped with these definitions, we can state the first result for convergence of minimizers.

Theorem 18 (Braides (2002), Theorem 1.21) *Let $\{F_n\}_n$ be a sequence of equi-mildly coercive functionals on metric space $\mathcal{U}$ with Γ-limit F_*. Then $\min_{u\in\mathcal{U}} F_*(u)$ exists and equals $\lim_{n\to\infty} \min_{u\in\mathcal{U}} F_n(u)$. Furthermore, if the F_n-minimizer sequence $\{\widehat{u}_n\}_n$ is precompact, then all subsequence limits of $\{\widehat{u}_n\}_n$ are minimizers in F_*.*

Bounded sequences are necessarily precompact for any metric space $\mathcal{U}$ over the reals.

A stronger version of the coercivity condition for F_n can be stated where, for some compact $\mathcal{A} \subset \mathcal{U}$, the sequence $\{F_n\}_n$ satisfies

$$\inf_{u\in\mathcal{A}} F_n(u) < \inf_{u\in\mathcal{U}\setminus\mathcal{A}} F_n(u) \qquad \text{for all } n.$$

Under this modified coercivity condition, the sequence of F_n-minimizers $\{\widehat{u}_n\}_n$ is bounded and, as such, must contain a convergent subsequence of minimizers for real-valued $\mathcal{U}$.

D.2 k-means Consistency

In this section, we present the k-means consistency result for a specific normed space $\mathcal{U} = (\mathbb{R}, |\cdot|)$. Proofs for the following results and their subsequent generalizations can be found in Thorpe et al. (2015).

The k-means algorithm takes a sample $(u_1, u_2, \ldots, u_n) \in \mathbb{R}^n$ and optimizes for the closest centers $\widehat{\boldsymbol{\mu}} \coloneqq (\widehat{\mu}_k)_{k=1}^M$ according to the distance based objective

$$L_n(\boldsymbol{\mu}) \coloneqq \frac{1}{n}\sum_{i=1}^{n} \min_{k\in[M]} |\mu_k - u_i|_2^2. \tag{50}$$

The population analog of (50) is defined with respect to the probability law $\mathbb{P}$ under which marginal u_1 is sampled, that is,

$$L_*(\boldsymbol{\mu}) \coloneqq \int \min_{k\in[M]} |\mu_k - u|_2^2 \, \mathbb{P}(du). \tag{51}$$

The k-means optimization (50) is considered *consistent* if the sequence of empirical minimizers $\{\widehat{\boldsymbol{\mu}}_n\}_n$ of $\{L_n\}_n$ has a limit (or a subsequence limit) which minimizes (51).

Going forward, sample outcomes will be denoted as $\omega \coloneqq (u_i)_{i=1}^{\infty}$. Similarly, sample quantities like (50) which depend on ω will be made more specific using the notation $L_n^{(\omega)}$.

The first result of Thorpe et al. (2015) is a $\mathbb{P}$-almost everywhere equivalence for the Γ-limit of L_n. For the specific normed space $\mathcal{U} = (\mathbb{R}, |\cdot|)$, no additional assumptions are needed.

Theorem 19 (Thorpe et al. (2015), Theorem 3.2) *Let L_n and L_* be defined as they are in* (50) *and* (51)*. Further let each u_i be independently drawn from $\mathbb{P}$. Then, for $\mathbb{P}$-almost every ω, L_* is the Γ-limit of $\{L_n^{(\omega)}\}_n$.*

The statement "$\mathbb{P}$-almost every ω" refers to the law induced by $\mathbb{P}$ for ω, that is $\mathbb{P}^{\infty} \coloneqq \prod_{i=1}^{\infty} \mathbb{P}$.

Equi-mild coercivity of objective for $\{L_n\}_n$ can be shown if optimizing with fewer than M distinct centers is suboptimal for the population objective L_*. This can be guaranteed with the following support assumption for $\mathbb{P}$:

Assumption 1 *There exists M distinct points $\mu_1, \mu_2, \ldots, \mu_M \in \mathcal{U}$ such that, for all $R > 0$,*

$$\min_{k \in [M]} \mathbb{P}(\{u : |\mu_k - u| < R\}) > 0.$$

In the case $\mathbb{P}$ is a mixture distribution, it suffices for $\mathbb{P}$ to have at least M relative modes (e.g. for $M = 2$, a bimodal distributions with two distinct centers).

Next, provided by Thorpe et al. (2015), is a coercivity result for the k-means algorithm. In fact, this claim is stronger than the usual equi-mild coercivity.

Proposition 20 (Thorpe et al. (2015) Proposition 3.3) *Under Assumption 1 and there exists a $\delta > 0$ such that, for $\mathbb{P}$-almost every ω, there is a $R > 0$ where*

$$\inf_{\|\boldsymbol{\mu}\| \leq R} L_n^{(\omega)}(\boldsymbol{\mu}) \leq \inf_{\|\boldsymbol{\mu}\| > R} L_n^{(\omega)}(\boldsymbol{\mu}) - \delta \tag{52}$$

holds for all sufficiently large n.

Proposition 20 can also be extended to all ϵ_n-almost minimizers of $L_n^{(\omega)}$ (Lember, 2003, Lemma 2.1).

Finally, by the Γ-convergence theorem of minimizers (Theorem 18), we achieve the following consistency result for the k-means algorithm:

Corollary 21 *Let $\{\widehat{\boldsymbol{\mu}}_n\}_n$ be a sequence of minimizers for $\{L_n\}_n$. Then, under Assumption 1, any limit or subsequence limit of $\{\widehat{\boldsymbol{\mu}}_n\}_n$ must be a minimizer for L_*, $\mathbb{P}$-almost surely.*

Under this result, it is valid for $\{\widehat{\boldsymbol{\mu}}_n\}_n$ to alternate between different empirical minimizers. Corollary 21 stipulates that each of these empirical minimizers is also a minimizer for L_*.

D.3 Cluster Consistency for Contaminated Objectives

Recall that a set of samples $\{y_i\}_{i=1}^n \subset \mathbb{R}$ is step-and-smooth on $\mathcal{X}$ if

$$y_i = f(x_i) + u_i = f(x_i) + g(x_i) + \varepsilon_i$$

for continuous f and bounded g. For the KRR estimate $\widehat{f}_n$ of f, the new perturbed optimization of interest is

$$\widetilde{L}_n(\boldsymbol{\mu}) \coloneqq \frac{1}{n} \sum_{i=1}^n \min_{k \in [M]} |\mu_k + \widehat{f}_n(x_i) - y_i|^2. \tag{53}$$

Define $\mathrm{MSE}_n^{(\omega)} \coloneqq \frac{1}{n} \sum_{i=1}^n |f(x_i) - \widehat{f}(x_i)|^2$ where $\omega_i = (x_i, \varepsilon_i)$. It can be shown that the perturbed objective (53) follows a triangle inequality. More generally, consider minimization

over a general set $\mathcal{A}$, then

$$\begin{aligned}\sum_{i=1}^{n}\min_{a\in\mathcal{A}}\|a+b_i+c_i\|^2 &\le \sum_{i=1}^{n}\min_{a\in\mathcal{A}}(\|a+b_i\|+\|c_i\|)^2\\ &= \sum_{i=1}^{n}\big((\min_{a\in\mathcal{A}}\|a+b_i\|)+\|c_i\|\big)^2\\ &= \sum_{i=1}^{n}(\min_{a\in\mathcal{A}}\|a+b_i\|)^2+2\sum_{i=1}^{n}\|c_i\|\min_{a\in\mathcal{A}}\|a+b_i\|+\sum_{i=1}^{n}\|c_i\|^2\\ &\le \Big(\Big(\sum_{i=1}^{n}\min_{a\in\mathcal{A}}\|a+b_i\|^2\Big)^{1/2}+\Big(\sum_{i=1}^{n}\|c_i\|^2\Big)^{1/2}\Big)^2.\end{aligned}$$

For the reverse-inequality, express $b_i \coloneqq b_i'+c_i'$ and $c_i \coloneqq -c_i'$ to obtain

$$\Big(\sum_{i=1}^{n}\min_{a\in\mathcal{A}}\|a+b_i'\|^2\Big)^{1/2} \le \Big(\sum_{i=1}^{n}\min_{a\in\mathcal{A}}\|a+b_i'+c_i'\|^2\Big)^{1/2}+\Big(\sum_{i=1}^{n}\|c_i'\|^2\Big)^{1/2}.$$

Specified to the loss in (53), one can rewrite these inequalities in terms of the MSE and (50)

$$\big(L_n(\boldsymbol{\mu})\big)^{1/2}-\big(\mathrm{MSE}_n\big)^{1/2} \le \big(\widetilde{L}_n(\boldsymbol{\mu})\big)^{1/2} \le \big(L_n(\boldsymbol{\mu})\big)^{1/2}+\big(\mathrm{MSE}_n\big)^{1/2},\quad \forall\boldsymbol{\mu}\in\mathbb{R}^M. \tag{54}$$

This perturbation inequality is important as it is the key to inheriting both the Γ-convergence and the equi-mild coercivity properties from $L_n^{(\omega)}$.

Recall that for sequences $\{a_n\}_n$ and $\{b_n\}_n$ where a_n is potentially divergent and $\lim_n b_n = b$, the following limits hold with equality

$$\liminf_{n\to\infty}(a_n+b_n)=\liminf_{n\to\infty}a_n+b \quad\text{and}\quad \limsup_{n\to\infty}(a_n+b_n)=\limsup_{n\to\infty}a_n+b. \tag{55}$$

Equalities (55) paired with Definition 16 yields the following proposition for Γ-convergence.

Proposition 22 *Let Q be the probability distribution given by transformation $u = g(x)+\varepsilon$. Suppose $\mathrm{MSE}_n^{(\omega)}\to 0$. Then, for almost every ω, objective $\widetilde{L}_n^{(\omega)}$ Γ-converges to*

$$L_*(\boldsymbol{\mu}) = \int \min_{k\in[M]}|\mu_k-u|^2\,Q(du). \tag{56}$$

Proof It suffices to show that $\widetilde{L}_n^{(\omega)}$ and $L_n^{(\omega)}$ have the same Γ-limit whenever $\mathrm{MSE}_n^{(\omega)}\to 0$. The pushforward distribution Q inherits coordinate-wise independence from $\mathbb{P}$ so, by Theorem 19, the Γ-limit of $L_n^{(\omega)}$ equals (56) for almost every ω.

We square-root transform $\widetilde{L}_n^{(\omega)}$ and apply (54) to obtain

$$\begin{aligned}\liminf_{n\to\infty}\widetilde{L}_n^{(\omega)}(\boldsymbol{\mu}_n) &= \big(\liminf_{n\to\infty}\big(\widetilde{L}_n^{(\omega)}(\boldsymbol{\mu}_n)\big)^{1/2}\big)^2\\ &\ge \big(\liminf_{n\to\infty}\big\{\big(L_n^{(\omega)}(\boldsymbol{\mu}_n)\big)^{1/2}-\big(\mathrm{MSE}_n^{(\omega)}\big)^{1/2}\big\}\big)^2\\ &= \big(\liminf_{n\to\infty}\big(L_n^{(\omega)}(\boldsymbol{\mu}_n)\big)^{1/2}-\lim_{n\to\infty}\big(\mathrm{MSE}_n^{(\omega)}\big)^{1/2}\big)^2\\ &= \liminf_{n\to\infty}L_n^{(\omega)}(\boldsymbol{\mu}_n)\end{aligned}$$

where on the last line (55) was used. Similarly for the recovery sequence inequality

$$\begin{aligned}\limsup_{n\to\infty} \widetilde{L}_n^{(\omega)}(\boldsymbol{\mu}_n) &= \big(\liminf_{n\to\infty} \big(\widetilde{L}_n^{(\omega)}(\boldsymbol{\mu}_n)\big)^{1/2}\big)^2 \\ &\leq \big(\limsup_{n\to\infty} \big\{\big(L_n^{(\omega)}(\boldsymbol{\mu}_n)\big)^{1/2} + \big(\mathrm{MSE}_n^{(\omega)}\big)^{1/2}\big\}\big)^2 \\ &= \limsup_{n\to\infty} L_n^{(\omega)}(\boldsymbol{\mu}_n).\end{aligned}$$

Reviewing Definition 16, we see that any Γ-limit of $L_n^{(\omega)}$ must also be a Γ-limit of $\widetilde{L}_n^{(\omega)}$. ■

Next, we show equi-mild coercivity (Definition 17) for the perturbed objective $\widetilde{L}_n$.

Proposition 23 *Suppose* $\mathrm{MSE}_n^{(\omega)} \to 0$. *If* $L_n^{(\omega)}$ *satisfies* (52), *then* $\{\widetilde{L}_n^{(\omega)}\}_n$ *is a sequence of equi-mild coercive objectives.*

Proof It suffices to show that the sequence of $\widetilde{L}_n^{(\omega)}$-minimizers, $\{\widetilde{\boldsymbol{\mu}}_n\}_n$, are bounded for ω satisfying (52). Suppose, for the sake of contradiction, $\|\widetilde{\boldsymbol{\mu}}_n\| \to \infty$. By the diverging nature of $\widetilde{\boldsymbol{\mu}}_n$, every $R > 0$ can be associated with a $N \in \mathbb{N}$ such that, for all $n > N_1$,

$$\widetilde{L}_n^{(\omega)}(\widetilde{\boldsymbol{\mu}}_n) = \inf_{\|\boldsymbol{\mu}\|>R} \widetilde{L}_n^{(\omega)}(\boldsymbol{\mu}).$$

Let R be defined as in Proposition 20 where, when paired with $\delta > 0$,

$$\inf_{\boldsymbol{\mu}\in\mathcal{U}} \big(L_n^{(\omega)}(\boldsymbol{\mu})\big)^{1/2} \leq \big(\inf_{\|\boldsymbol{\mu}\|>R} L_n^{(\omega)}(\boldsymbol{\mu}) - \delta\big)^{1/2}$$

Additionally by (54),

$$\begin{aligned}\inf_{\boldsymbol{\mu}\in\mathcal{U}} \big(L_n^{(\omega)}(\boldsymbol{\mu})\big)^{1/2} &\geq \inf_{\boldsymbol{\mu}\in\mathcal{U}} (\widetilde{L}_n^{(\omega)}(\boldsymbol{\mu}))^{1/2} - \big(\mathrm{MSE}_n^{(\omega)}\big)^{1/2} \\ &= \inf_{\|\boldsymbol{\mu}\|>R} (\widetilde{L}_n^{(\omega)}(\boldsymbol{\mu}))^{1/2} - \big(\mathrm{MSE}_n^{(\omega)}\big)^{1/2}, \quad \text{for suff. large } n.\end{aligned}$$

Stringing both inequalities together and further upperbounding $L_n^{(\omega)}$ by (54) yields

$$\inf_{\|\boldsymbol{\mu}\|>R} (\widetilde{L}_n^{(\omega)}(\boldsymbol{\mu}))^{1/2} \leq \big(\inf_{\|\boldsymbol{\mu}\|>R} \widetilde{L}_n^{(\omega)}(\boldsymbol{\mu}) + \big(\mathrm{MSE}_n^{(\omega)}\big)^{1/2} - \delta\big)^{1/2} + \big(\mathrm{MSE}_n^{(\omega)}\big)^{1/2}.$$

However, since $\mathrm{MSE}_n^{(\omega)} \to 0$, there exists $N_2 \in \mathbb{N}$ such that, for all $n > N_2$,

$$\inf_{\|\boldsymbol{\mu}\|>R} (\widetilde{L}_n^{(\omega)}(\boldsymbol{\mu}))^{1/2} \leq \big(\inf_{\|\boldsymbol{\mu}\|>R} \widetilde{L}_n^{(\omega)}(\boldsymbol{\mu}) - \delta/2\big)^{1/2}$$

which is a contradiction for all $\delta > 0$. This completes the proof. ■

Combining Propositions 22 and 23 to apply Theorem 18, we show that, for almost-every $\omega \in \mathcal{B}$, the sequence $\{\widehat{\boldsymbol{\mu}}_n\}_n$ minimizes L_* in the limit. In particular, there exists a sub-index set $\mathcal{I}'' \subseteq \mathcal{I}'$ such that $\widehat{\boldsymbol{\mu}}_m \to \overline{\boldsymbol{\mu}} \in \mathcal{M}$ with $m \in \mathcal{I}''$. This shows the desired claim (49) for any initial choice of index set $\mathcal{I} \subset \mathbb{N}$.

D.4 Nearest Label Consistency from Cluster Convergence

The convergence of cluster centers $\{\widehat{\boldsymbol{\mu}}_m\}_m \to \overline{\boldsymbol{\mu}}$ guarantees a convergence in the nearest label estimates $\{\widehat{\boldsymbol{z}}_m\}_m$ to limiting labels $\overline{z}$. These convergences are abundant in $\mathcal{B}$ in the sense that every subsequence of $\{\widehat{\boldsymbol{\mu}}_n\}_n$ has a further subsequence which converges to an element in $\mathcal{M}$.

Fix one such limiting center $\overline{\boldsymbol{\mu}} \in \mathcal{M}$ and let $\{\mathcal{V}_k\}_{k=1}^M$ be the Voronoi-partition on $\mathbb{R}$ according to $\{\overline{\mu}_k\}_{k=1}^M$. Refer to (46) for a definition of $\mathcal{V}_k$. In the case when centers $\{\overline{\mu}_k\}_{k=1}^M$ are distinct, each $\mathcal{V}_k$ can be expressed as the intersection of M half-spaces $\mathcal{V}_k = \bigcap_{\ell=1}^M H_{k\ell}(\overline{\boldsymbol{\mu}})$ where

$$H_{k\ell}(\boldsymbol{\mu}) \coloneqq \big\{u \in \mathbb{R} : \big(u - (\mu_k + \mu_\ell)/2\big) \cdot (\mu_k - \mu_\ell) \geq 0\big\}. \tag{57}$$

By Assumption 1, any minimizer of L_* must have M distinct values. As such, the map $\{\overline{\mu}_k\}_k \mapsto \{\mathcal{V}_k\}_k$ is well-defined for all minimizers of (56).

Let $\mathring{A}$ denote the interior of a set A. To obtain nearest labels $\overline{z}$ according to $\overline{\boldsymbol{\mu}}$, we consider the following routine:

1. If a point u_i lies in the interior of a cell $\mathring{\mathcal{V}}_k$ then $\overline{z}_i = k$.
2. Otherwise $\overline{z}_i$ is selected arbitrarily from $[M]$.

Note that the nearest labels $\overline{z}$ are not guaranteed to agree with the generating labels z^*. This is to be expected whenever $(x, \varepsilon) \mapsto g^*(x) + \varepsilon$ is not injective over $\operatorname{supp}(\mathbb{P})$. In the case of separable clusters, that is, when noise ε is bounded and small relative to the levels of $g^*(x)$, one indeed has, up to a label permutation, $\overline{z}_i = z_i^*$ for all generated samples $(x_i, \varepsilon_i)_{i=1}^\infty$.

The nearest center labeling routine can also be extended to any estimated centers $\widehat{\boldsymbol{\mu}}_m \in \mathbb{R}^M$. Let $\{\mathcal{V}_{k,m}\}_k$ be the Voronoi partition of $\widehat{\boldsymbol{\mu}}$ and let $\widehat{\boldsymbol{z}}_n$ be the corresponding nearest labels. Similar to the population case, partition $\{\mathcal{V}_{k,m}\}_{k=1}^M$ is well-defined whenever $\widehat{\boldsymbol{\mu}}_m \to \overline{\boldsymbol{\mu}}$ and m is sufficiently large. Note, as a consequence of the coordinate-wise convergence $\|\widehat{\boldsymbol{\mu}}_m - \overline{\boldsymbol{\mu}}\|$, it is sufficient to use the identity permutation to align nearest labels $\widehat{\boldsymbol{z}}$ and $\overline{z}$.

When calculating misclassification, we will avoid ambiguity by measuring misclassification relative to the interior of different Voronoi cells. With this in mind, the misclassification (48) for nearest labels $\widehat{\boldsymbol{z}}_m$ and $\overline{z}$ can be expressed as

$$\operatorname{Miss}_m(\widehat{\boldsymbol{z}}_m, \overline{z}) = \sum_{k \neq \ell} C_{k\ell}, \tag{58}$$

where

$$C_{k\ell} \coloneqq \sum_{k \neq \ell} \frac{1}{n} \sum_{i=1}^n 1_{\mathring{\mathcal{V}}_{k,n}}(u_i) \cdot 1_{\mathring{\mathcal{V}}_\ell}(u_i) \tag{59}$$

and $u_i = g^*(x_i) + \varepsilon_i$ are sample coordinates with joint distribution Q.

Next, we present the following almost-sure misclassification convergence result for any sequence of centers which satisfies $\widehat{\boldsymbol{\mu}}_m \to \overline{\boldsymbol{\mu}}$.

Theorem 24 *Let $\{\widehat{\boldsymbol{z}}_m\}_m$ and $\overline{z}$ be defined as before. Suppose $g^*(x_i) + \varepsilon_i = u_i \overset{i.i.d.}{\sim} Q$ and $\widehat{\boldsymbol{\mu}}_m \to \overline{\boldsymbol{\mu}}$. Then, for almost every $(x_i, \varepsilon_i)_{i=1}^\infty \in \mathcal{B}$,*

$$\lim_{m \to \infty} \operatorname{Miss}_m(\widehat{\boldsymbol{z}}_m, \overline{z}) = 0.$$

Proof It suffices to show $C_{k\ell} \to 0$ for all $k \neq \ell$. Recall that $\operatorname{int}\big(\bigcap_i A_i\big) \subseteq \bigcap_i \mathring{A}_i$. Applying the interior intersection relation to the Voronoi cells $\mathcal{V}_{k,m}$ and $\mathcal{V}_\ell$ yields

$$\mathring{\mathcal{V}}_{k,m} \subseteq \bigcap_{\ell=1}^{M} \mathring{H}_{k\ell}(\widehat{\boldsymbol{\mu}}_m) \subseteq \mathring{H}_{k\ell}(\widehat{\boldsymbol{\mu}}_m) \qquad \text{and} \qquad \mathring{\mathcal{V}}_\ell \subseteq \bigcap_{k=1}^{M} \mathring{H}_{\ell k}(\overline{\boldsymbol{\mu}}) \subseteq \mathring{H}_{\ell k}(\overline{\boldsymbol{\mu}}).$$

And as a consequence,

$$C_{k\ell} \leq \frac{1}{m} \sum_{i=1}^{m} 1_{\mathring{H}_{k\ell}(\widehat{\boldsymbol{\mu}}_m)}(u_i) \cdot 1_{\mathring{H}_{\ell k}(\overline{\boldsymbol{\mu}})}(u_i).$$

To decouple from $\widehat{\mu}_n$ we define the following δ-silhouette about $\bar{\mu}$ for half-spaces $H_{k\ell}(\cdot)$,

$$\mathring{H}^{\delta}_{k\ell}(\overline{\boldsymbol{\mu}}) \coloneqq \bigcup_{\|\boldsymbol{\mu}-\overline{\boldsymbol{\mu}}\|<\delta} \mathring{H}_{k\ell}(\boldsymbol{\mu}). \tag{60}$$

By the convergence of centers $\widehat{\boldsymbol{\mu}}_m \to \overline{\boldsymbol{\mu}}$, one eventually has $\mathring{H}_{k\ell}(\widehat{\boldsymbol{\mu}}_m) \subseteq \mathring{H}^{\delta}_{k\ell}(\overline{\boldsymbol{\mu}})$ for all $k \neq \ell$ and any fixed $\delta > 0$. Therefore,

$$\begin{aligned}
C_{k\ell} &\leq \frac{1}{m} \sum_{i=1}^{m} 1_{\mathring{H}^{\delta}_{k\ell}(\overline{\boldsymbol{\mu}})}(u_i) \cdot 1_{\mathring{H}_{\ell k}(\overline{\boldsymbol{\mu}})}(u_i) \\
&\leq \frac{1}{m} \sum_{i=1}^{m} |1_{\mathring{H}^{\delta}_{k\ell}(\overline{\boldsymbol{\mu}})}(u_i) - 1_{H_{k\ell}(\overline{\boldsymbol{\mu}})}(u_i)| \cdot 1 + \frac{1}{m} \sum_{i=1}^{m} |1_{H_{k\ell}(\overline{\boldsymbol{\mu}})}(u_i)| \cdot 1_{\mathring{H}_{\ell k}(\overline{\boldsymbol{\mu}})}(u_i) \\
&= \frac{1}{m} \sum_{i=1}^{m} |1_{\mathring{H}^{\delta}_{k\ell}(\overline{\boldsymbol{\mu}})}(u_i) - 1_{H_{k\ell}(\overline{\boldsymbol{\mu}})}(u_i)| + 0,
\end{aligned}$$

where the last line follows from the fact $H_{k\ell}(\overline{\boldsymbol{\mu}}) \cap \mathring{H}_{\ell k}(\overline{\boldsymbol{\mu}}) = \varnothing$. Going forward we will suppress all dependence on $\overline{\boldsymbol{\mu}}$ for half-space sets $H_{k\ell}$ and $\mathring{H}^{\delta}_{k\ell}$.

Next, note that $H_{k\ell} \subseteq \mathring{H}^{\delta}_{k\ell}$ for all $\delta > 0$. Clearly by construction $\mathring{H}_{k\ell} \subseteq \mathring{H}^{\delta}_{k\ell}$. So, for $u \in H_{k\ell} \setminus \mathring{H}_{k\ell}$, consider an arbitrarily small perturbation $\epsilon \propto \operatorname{sgn}(\mu_k - \mu_\ell)$. Since u satisfies

$$\big(u - (\bar{\mu}_k + \bar{\mu}_\ell)/2\big) \cdot (\bar{\mu}_k - \bar{\mu}_\ell) = 0,$$

one can shift $\bar{\mu}$ as $\mu = \bar{\mu} + \epsilon$ to obtain

$$\big(u - (\bar{\mu}_k + \bar{\mu}_\ell + 2\epsilon)/2\big) \cdot (\bar{\mu}_k - \bar{\mu}_\ell) = \epsilon \cdot (\bar{\mu}_k - \bar{\mu}_\ell) \propto |\bar{\mu}_k - \bar{\mu}_\ell|,$$

where the RHS is strictly positive whenever $\bar{\mu}_k \neq \bar{\mu}_\ell$.

As such, we can combine indicators and have

$$\begin{aligned}
\lim_{m\to\infty} C_{k\ell} &\leq \lim_{m\to\infty} \frac{1}{m} \sum_{i=1}^{n} 1_{\mathring{H}^{\delta}_{k\ell} \setminus H_{k\ell}}(u_i) \\
&= Q(\mathring{H}^{\delta}_{k\ell} \setminus H_{k\ell}),
\end{aligned} \tag{61}$$

where the last line holds by independence for almost every $(u_i)_{i=1}^{\infty}$. By similar reasoning, measure convergence (61) holds, simultaneously, the countable family of sets $\big\{\mathring{H}_{k\ell}^{\delta_p} \setminus H_{k\ell}\big\}_{p=1}^{\infty}$. Let δ_p be any positive real sequence with $\delta_p \to 0^+$, then, by continuity of measure,

$$\begin{aligned}\lim_{m\to\infty} C_{k\ell} &\leq \inf_{p\to\infty} Q\big(\mathring{H}_{k\ell}^{\delta_p} \setminus H_{k\ell}\big) \\ &= Q\Big(\bigcap_{p=1}^{\infty} \big(\mathring{H}_{k\ell}^{\delta_p} \setminus H_{k\ell}\big)\Big) \\ &= Q\Big(\Big(\bigcap_{p=1}^{\infty} \mathring{H}_{k\ell}^{\delta_p}\Big) \setminus H_{k\ell}\Big) \\ &= Q(\varnothing) = 0.\end{aligned}$$

To prove the empty set assertion, note the following contrapositive statement

$$u \notin H_{k\ell} \implies u \notin \bigcap_{p=1}^{\infty} \mathring{H}_{k\ell}^{\delta_p}.$$

Indeed, fix u and define the continuous function $t_u(\mu) \coloneqq \big(u - (\mu_k + \mu_\ell)/2\big) \cdot (\mu_k - \mu_\ell)$. If $u \notin H_{k\ell}$ then $t_u(\bar{\mu}) < 0$ and, by continuity, there exists some $\epsilon_u > 0$ such that

$$\|\mu - \bar{\mu}\| < \epsilon_u \implies t_u(\mu) < 0.$$

Therefore, for every m satisfying $\delta_m < \epsilon_u$, one has $u \notin \mathring{H}_{k\ell}^{\delta_p}$. ■

As a last comment, we note that the proof of Theorem 24 can be straight-forwardly extended for samples u_i belonging to a general Hilbert space $\mathbb{H}$. In the case where $\mathbb{H}$ is a Banach space, modifications must be made to the half-space representation of $\mathcal{V}_k$.