GREEDY MMD WITHOUT REPLACEMENT: SELECTING AND STOPPING FOR MIXTURE SEPARATION

Anonymous authors
Paper under double-blind review

ABSTRACT

This paper studies how to split an unlabelled sample drawn from a mixture $\theta P+(1-\theta)Q$ when a separate reference sample from P is available. Greedy maximum mean discrepancy (MMD) selection without replacement chooses distinct observations to approximate P, leaving a remainder intended to approximate Q. Success depends on order and stopping: early choices may remove observations needed later, while a wrong stopping size can leave a poor remainder even when the selected observations fit P. An error bound tracks the effect of earlier selections. Under two geometric conditions and appropriate stopping, convergence rates are established for the MMD errors of the selected group relative to P and the remainder relative to Q. Simulations and real-data experiments illustrate the effect of stopping and support the predicted convergence behaviour.

1 INTRODUCTION

Mixture data problems arise broadly in machine learning. Specifically, suppose that reference samples are drawn from a target distribution P, while the candidate pool is drawn from the mixture $\mathsf{M} := \theta P + (1-\theta)Q$, where $\theta \in (0,1)$. This observation model is also central to mixture proportion estimation (MPE), which seeks to identify or estimate the mixture parameter θ under suitable identifiability conditions (Blanchard et al., 2010; Scott, 2015; Ramaswamy et al., 2016; du Plessis et al., 2017; Bekker & Davis, 2018; Ivanov, 2020; Garg et al., 2021; Zhu et al., 2023). A related but distinct line of research is positive–unlabeled (PU) learning (Elkan & Noto, 2008; Kiryo et al., 2017; Bekker & Davis, 2020). PU learning uses labeled positive samples from P together with unlabeled samples from M to learn a classifier that distinguishes the latent positive and negative classes. In contrast to the instance-level classification objective of PU learning, we consider distribution-level separation, partitioning the mixture sample into two disjoint groups whose empirical distributions approximate P and Q, respectively.

As a discrepancy measure between probability distributions, maximum mean discrepancy (MMD) embeds distributions into a reproducing kernel Hilbert space (RKHS) and provides a natural criterion for distributional matching through kernel mean embeddings. This criterion has been widely used for representative subset selection, where points are selected from a candidate pool so that their empirical distribution approximates a target distribution. Kernel herding greedily selects points to approximate a target mean embedding and can be interpreted as a Frank–Wolfe algorithm over a fixed convex domain in the RKHS (Chen et al., 2010; Bach et al., 2012). For a finite candidate pool, MMD-Critic (Kim et al., 2016) greedily selects distinct representative prototypes by matching the empirical mean embedding of the candidate pool itself. Greedy MMD selection has also been studied for external targets, including greedy quantization over a fixed candidate pool with repeated selections allowed (Teymur et al., 2021) and related convergence analyses over fixed domains (Pronzato, 2023). More recently, Shao et al. (2024) study generalized kernel herding without replacement (GKHR), where selected candidates are removed from the available pool, while the target is still the empirical distribution of the same candidate pool. These results do not directly address the combination of an external target and without-replacement selection, where each selection changes the geometry of the remaining feasible pool. More importantly for mixture separation, representative selection alone does not address whether the unselected remainder approximates a second component distribution.

In this paper, we study mixture separation through greedy MMD selection without replacement. At each step, our algorithm greedily selects one sample from the remaining pool to minimize the MMD between the resulting selection and the empirical distribution of the reference sample, and

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

separates the mixture by stopping the greedy selection at an appropriate size. This formulation reveals two distinct challenges. First, without-replacement selection progressively depletes the candidate pool, so earlier selections can impair the ability of later selections to approximate the target. In the RKHS, we characterize this effect by tracking the distance from the target embedding to the shrinking convex hull of the remaining candidate embeddings. In particular, we derive a pathwise bound that explicitly isolates the resulting removal loss, which measures the deterioration in the best target approximation available from the remaining candidate pool due to earlier selections. Second, matching the reference distribution alone is insufficient for mixture separation: even the selected samples that closely approximate P may leave a poor approximation to Q if the greedy selection stops at an inappropriate size. An exact decomposition shows that, if the selected group is consistent with P, the remainder is consistent with Q if and only if the proportion of mixture observations selected is close to θ. We further establish convergence guarantees by controlling the removal loss under two structured regimes: an RKHS projection condition, which yields consistency under exact orthogonality and a quantified approximation bound otherwise; and a pairing condition when P has finite support, under which the exact greedy rule avoids prematurely exhausting target-support candidates. Experiments on controlled mixtures and real data show that similar target fit can yield substantially different remainder quality under different stopping sizes, while a structured experiment examines the predicted joint convergence rate.

Our contributions are summarized as follows:

- **Pathwise theory for greedy MMD selection without replacement.** For an arbitrary Hilbert-space target and a finite indexed candidate set, we derive a pathwise error bound that separates the best approximation available from the full candidate pool, the path-averaged removal loss caused by earlier greedy selections, and the one-step cost of restricting each update to a single candidate. We further give a projection-based bound on this path-averaged removal loss using a comparison subset that appears only in the analysis (Theorem 8). Sharp constructions show that the dependence of our bound on the path-averaged removal loss and selection size cannot be improved in the worst case.
- **A necessary-and-sufficient characterization of stopping for mixture separation.** We derive an exact identity linking the selected group, its remainder, and the selected fraction. Under the conditions on consistency of the selected group and distinct component embeddings, the remainder is consistent if and only if the selected fraction converges to the mixture proportion θ. This characterization is independent of the selection algorithm and isolates stopping accuracy as a requirement distinct from target matching.
- **Consistency and rates under structured mixture geometries.** We control the averaged removal loss under two complementary conditions: an RKHS projection condition and, when P has finite support, a pairing condition. These controls yield guarantees for both the MMD between the empirical distribution of the selected samples and P and that between the empirical distribution of the remaining samples and Q. For the projection-based result, we also quantify how departures from the condition affect the bound and show that this dependence is sharp.

2 Problem Setup

2.1 Mixture Model and Energy-Distance Objective

We first describe the statistical selection problem studied in the paper. Let $d \geq 1$, and let P and Q be probability distributions on $\mathbb{R}^d$ with finite first moments. Let $\theta \in (0,1)$, and set $\mathsf{M} := \theta P + (1-\theta)Q$. For $N_1 \geq 1$ and $N_2 \geq 2$, we observe independent data sets

$$D_1 := (y_1, \ldots, y_{N_1}) \sim P^{\otimes N_1}, \qquad D_2 := (z_1, \ldots, z_{N_2}) \sim \mathsf{M}^{\otimes N_2}.$$

Note that the observations do not reveal whether each candidate z_j was generated from P or Q. For every nonempty index set $A \subseteq [N_2]$,[1] define the candidate empirical measure by

$$\widehat{P}_A := \frac{1}{|A|} \sum_{j \in A} \delta_{z_j}.$$

The overall goal is to choose a nonempty index subset $S \subsetneq [N_2]$ such that

$$\widehat{P}_S \approx P, \qquad \widehat{P}_{[N_2]\setminus S} \approx Q, \qquad \frac{|S|}{N_2} \approx \theta. \tag{1}$$

[1] For a positive integer n, we use the standard notation $[n] := \{1, \ldots, n\}$.

Repeated observations at different indices remain distinct candidates. The objective is to approximate the component distributions, not to identify whether each candidate was generated from P or Q. In particular, when P and Q overlap, a candidate from Q may still improve S's approximation to P.

The three approximations in (1) are not equivalent. For example, $\widehat{P}_S \approx P$ does not by itself ensure either $|S|/N_2 \approx \theta$ or $\widehat{P}_{[N_2]\setminus S} \approx Q$. We therefore separately control how well $\widehat{P}_S$ approximates P and how close $|S|/N_2$ is to θ, measured by energy distance. Its Hilbert-space representation yields the empirical criterion used to choose S.

For probability distributions $\mathcal{D}_1$ and $\mathcal{D}_2$ on $\mathbb{R}^d$ with finite first moments, let $X, X' \overset{\text{iid}}{\sim} \mathcal{D}_1$ and $Y, Y' \overset{\text{iid}}{\sim} \mathcal{D}_2$, with the two pairs independent. We define the energy distance between $\mathcal{D}_1$ and $\mathcal{D}_2$ as

$$\mathcal{E}(\mathcal{D}_1, \mathcal{D}_2) := 2\,\mathbb{E}\,\|X - Y\|_2 - \mathbb{E}\,\|X - X'\|_2 - \mathbb{E}\,\|Y - Y'\|_2. \tag{2}$$

Then $\sqrt{\mathcal{E}(\cdot,\cdot)}$ is a metric on this class of probability distributions; see, e.g. (Székely & Rizzo, 2013). The energy kernel induced by the distance is

$$k_{\mathrm{E}}(x, x') := \frac{1}{2}\big(\|x\|_2 + \|x'\|_2 - \|x - x'\|_2\big),$$

with RKHS $\mathcal{H}_{\mathrm{E}}$ and canonical feature map $\Phi_{\mathrm{E}}(x) := k_{\mathrm{E}}(x, \cdot)$. Two basic identities are $k_{\mathrm{E}}(x, x) = \|x\|_2$ and $\|\Phi_{\mathrm{E}}(x) - \Phi_{\mathrm{E}}(x')\|^2_{\mathcal{H}_{\mathrm{E}}} = \|x - x'\|_2$. For a probability distribution $\mathcal{D}$ on $\mathbb{R}^d$ with finite first moment, define its kernel mean embedding by

$$\mu_{\mathcal{D}} := \mathop{\mathbb{E}}_{X\sim\mathcal{D}}[\Phi_{\mathrm{E}}(X)] = \int_{\mathbb{R}^d} \Phi_{\mathrm{E}}(x)\,\mathcal{D}(dx).$$

Write

$$\mathrm{MMD}_{k_{\mathrm{E}}}(\mathcal{D}_1, \mathcal{D}_2) := \|\mu_{\mathcal{D}_1} - \mu_{\mathcal{D}_2}\|_{\mathcal{H}_{\mathrm{E}}}.$$

For the energy kernel, the standard identity (Sejdinovic et al., 2013) is

$$\mathcal{E}(\mathcal{D}_1, \mathcal{D}_2) = 2\,\|\mu_{\mathcal{D}_1} - \mu_{\mathcal{D}_2}\|^2_{\mathcal{H}_{\mathrm{E}}} = 2\,\mathrm{MMD}^2_{k_{\mathrm{E}}}(\mathcal{D}_1, \mathcal{D}_2). \tag{3}$$

The selected distribution target in (1) is P. Since P is unknown, the empirical objective uses the reference empirical measure

$$\widehat{P}_{N_1} := \frac{1}{N_1}\sum_{i=1}^{N_1} \delta_{y_i}$$

as an observable proxy. Define its mean embedding and the candidate feature points in $\mathcal{H}_{\mathrm{E}}$ by

$$\widehat{\mu}^P_{N_1} := \mu_{\widehat{P}_{N_1}} = \frac{1}{N_1}\sum_{i=1}^{N_1} \Phi_{\mathrm{E}}(y_i), \qquad x_j := \Phi_{\mathrm{E}}(z_j), \quad j \in [N_2]. \tag{4}$$

For every nonempty $A \subseteq [N_2]$, define its empirical mean embedding by

$$\widehat{\mu}(A) := \mu_{\widehat{P}_A} = \frac{1}{|A|}\sum_{j\in A} \Phi_{\mathrm{E}}(z_j).$$

The corresponding pairwise-distance formula is recorded in Appendix A. For each size $1 \le k \le N_2$, empirical energy-distance selection is an equal-weight subset approximation problem in $\mathcal{H}_{\mathrm{E}}$.

2.2 Greedy Approximation Without Replacement

Let $\mathcal{H}$ be a real Hilbert space, $x_1, \ldots, x_n \in \mathcal{H}$ be a finite indexed candidate set and $u \in \mathcal{H}$ be a target point. For every nonempty subset $S \subseteq [n]$, define

$$\mathrm{err}(S) := \left\| u - \frac{1}{|S|}\sum_{i\in S} x_i \right\|_{\mathcal{H}}. \tag{5}$$

The empirical energy-distance problem is recovered by taking $n = N_2$, $\mathcal{H} = \mathcal{H}_{\mathrm{E}}$, $u = \widehat{\mu}^P_{N_1}$, and the candidate points $x_1, \ldots, x_n$ from (4). Under this identification,

$$\mathrm{err}(S) = \big\|\widehat{\mu}^P_{N_1} - \widehat{\mu}(S)\big\|_{\mathcal{H}_{\mathrm{E}}} = \mathrm{MMD}_{k_{\mathrm{E}}}(\widehat{P}_S, \widehat{P}_{N_1}), \qquad \mathrm{err}^2(S) = \frac{1}{2}\mathcal{E}(\widehat{P}_S, \widehat{P}_{N_1}). \tag{6}$$

By the triangle inequality,

$$\|\widehat{\mu}(S) - \mu_P\|_{\mathcal{H}_E} \le \mathrm{err}(S) + \|\widehat{\mu}^P_{N_1} - \mu_P\|_{\mathcal{H}_E}.$$

The second term is independent of S, so controlling $\mathrm{err}(S)$ controls the selected-set error. Rather than solve a separate problem for each size, we seek one ordering whose prefixes have controlled error simultaneously.

Problem 1. Construct a nested sequence

$$\varnothing = S_0 \subset S_1 \subset \cdots \subset S_n = [n], \qquad |S_k| = k,$$

and control the prefix errors $\mathrm{err}(S_k)$ simultaneously for $1 \le k \le n$.

Problem 1 treats $u, x_1, \ldots, x_n$ as arbitrary fixed points and imposes no probability model. Selection is without replacement at the index level: if $x_i = x_j$ for $i \ne j$, both indices may be selected, but neither may be selected twice.

A natural additional assumption is $u \in \mathrm{conv}\{x_1, \ldots, x_n\}$. Under (4), however, the target and candidate points come from different samples, so this condition need not hold. Even when it does, convex-hull membership alone does not ensure improving equal-weight approximation without replacement; Proposition 12 gives a sharp example. We therefore track how selection changes the convex hull of the unselected candidates.

We address Problem 1 with the recursive ordering

$$i_{k+1} \in \operatorname*{arg\,min}_{i \in R_k} \mathrm{err}(S_k \cup \{i\}), \quad S_{k+1} := S_k \cup \{i_{k+1}\}, \tag{7}$$

where $R_k := [n] \setminus S_k$, and the initial values are $i_1 \in \arg\min_{i \in [n]} \mathrm{err}(\{i\})$ and $S_1 = \{i_1\}$.

At any step with more than one minimizing index, we choose the smallest one. For $u = n^{-1} \sum_{j=1}^n x_j$, this objective and update coincide with MMD-Critic (Kim et al., 2016, Equation (6) and Algorithm 1). Here, u is arbitrary and the remaining convex hull can move farther from it as candidates are removed.

2.3 Energy-Greedy Selection and Stopping

Under (4), the abstract rule in (7) adds the available candidate that minimizes the squared empirical energy distance. We call this the *energy-greedy* (EG) ordering. Its Euclidean form, pseudocode, and incremental implementation are given in Appendix A. Constructing the full EG path costs $O(N_1 N_2 + N_2^2)$ distance evaluations and $O(N_2)$ additional memory. The maintained sums give the PathMin score for every prefix, so no additional candidate-distance evaluations are required.

For $1 \le k \le N_2$, write $\alpha_k := k/N_2$ for the fraction of candidates in the kth prefix. The full ordering gives the path in Problem 1. For mixture separation, the final prefix must have size $k < N_2$ so that its remainder is nonempty. PathMin returns the admissible prefix with the smallest empirical error:

$$k^* \in \operatorname*{arg\,min}_{1 \le k < N_2} \mathrm{err}(S_k). \tag{8}$$

Ties are resolved in favor of the earliest prefix. PathMin returns S_{k^*} and R_{k^*}, but a small $\mathrm{err}(S_{k^*})$ need not imply a small $|k^*/N_2 - \theta|$. Other stopping rules may use a known value or an externally justified estimator of θ.

Section 3.1 bounds $\mathrm{err}(S_k)$ using convex hulls of the remaining candidates. Section 3.2 shows how $\mathrm{err}(S_{\widehat{k}})$ and $|\widehat{k}/N_2 - \theta|$ control both group errors. Section 3.3 verifies the required depletion control under two structural regimes.

3 Greedy Approximation and Mixture Separation

3.1 Pathwise Greedy Depletion

Consider the greedy construction for Problem 1. Let $S_k = \{i_1, \ldots, i_k\}$ be the first k greedy selections and let $R_k = [n] \setminus S_k$ for $1 \le k \le n$. Set $S_0 := \varnothing$ and $R_0 := [n]$. Write

$$\mathrm{err}_k := \mathrm{err}(S_k), \quad 1 \le k \le n, \qquad \mathcal{C}_k := \mathrm{conv}\{x_i : i \in R_k\}, \quad 0 \le k < n. \tag{9}$$

For a nonempty closed convex set $C \subseteq \mathcal{H}$, write Π_C for its metric projection. When C is a closed subspace, Π_C is the orthogonal projection.

Define

$$d_j := \operatorname{dist}(u, \mathcal{C}_j), \quad \epsilon_j := d_j - d_0, \quad C_{\mathrm{pool}} := \operatorname{diam}^2_{\mathcal{H}}(\mathcal{C}_0), \quad \kappa_j := \frac{\operatorname{diam}^2_{\mathcal{H}}(\mathcal{C}_j)}{2}\left(1 - \frac{1}{n-j}\right).$$

Since taking a convex hull does not change the diameter, d_0 is the best error attainable from the full candidate set, and ϵ_j is the additional *removal loss* after j selections. To analyse one step, one first allows the next choice to be any convex combination of the remaining candidates and then accounts for the restriction to one candidate. As shown in Appendix B, this gives

$$(j+1)^2 \operatorname{err}_{j+1}^2 \le (j \operatorname{err}_j + d_j)^2 + \kappa_j, \qquad 1 \le j < n.$$

Here d_j measures the mismatch between u and the remaining convex hull and κ_j bounds the additional cost of selecting one candidate rather than an arbitrary convex combination. Iterating this yields the first upper bound below. For $1 \le k \le n$, define

$$\overline{d}_k := \frac{1}{k}\sum_{j=1}^{k-1} d_j, \qquad \overline{\epsilon}_k := \frac{1}{k}\sum_{j=1}^{k-1} \epsilon_j, \tag{10}$$

Theorem 1. *For the greedy sequence of Section 2.2, the following inequalities hold for every $2 \le k \le n$:*

$$d_0 \le \operatorname{err}_k \le \overline{d}_k + \frac{\sqrt{\operatorname{err}_1^2 + \sum_{j=1}^{k-1} \kappa_j}}{k} \le d_0 + \overline{\epsilon}_k + \sqrt{\frac{C_{\mathrm{pool}}}{2k}}. \tag{11}$$

The proof of Theorem 1 can be found in Appendix B. The first upper bound retains the stepwise values of d_j and κ_j along the greedy path. The second upper bound keeps the removal loss $\overline{\epsilon}_k$ but uses $\operatorname{diam}^2_{\mathcal{H}}(\mathcal{C}_j) \le C_{\mathrm{pool}}$ to yield a simpler but potentially looser bound. If $d_j = d_0$ for all $j < k$, then $\overline{\epsilon}_k = 0$; when also $d_0 = 0$, one obtains $\operatorname{err}_k^2 \le C_{\mathrm{pool}}/(2k)$. This $O(1/k)$ order is tight; Proposition 13 gives a matching construction.

The bound of Theorem 1 is useful when $\overline{\epsilon}_k$ is small, meaning that earlier selections have not substantially worsened the best approximation available from the remaining candidates. Section 3.3 gives two sets of structural assumptions under which $\overline{d}_k$, and hence $\overline{\epsilon}_k$, can be bounded.

3.2 From Stopping to Separation

Recall that the selection procedure returns a subset S and its remainder is $R := [N_2] \setminus S$. Independently of how S is constructed, it is expected that $\widehat{\mu}(S)$ approximates u and $|S|/N_2$ approximates θ. Let $\alpha := |S|/N_2$ and $\bar{\mu}_2 := (1/N_2)\sum_{j=1}^{N_2} \Phi_{\mathrm{E}}(z_j)$. It follows from $\mu_{\mathsf{M}} = \theta\mu_P + (1-\theta)\mu_Q$ that

$$(1-\alpha)(\widehat{\mu}(R) - \mu_Q) = (\bar{\mu}_2 - \mu_{\mathsf{M}}) + \alpha(\mu_P - \widehat{\mu}(S)) + (\theta - \alpha)(\mu_P - \mu_Q). \tag{12}$$

Thus the remainder error $\|\widehat{\mu}(R) - \mu_Q\|_{\mathcal{H}_{\mathrm{E}}}$ has three sources: mixture-sample error, selected-set error, and selected-fraction error. In particular, even if $\operatorname{err}(S)$ is small, R need not approximate Q unless $\bar{\mu}_2$ is close to μ_{M} and α is close to θ.

The following result makes these requirements precise for any possibly data-dependent selected set.

Theorem 2. *Under the mixture sampling model of Section 2.1, let $S \subset [N_2]$ be any possibly data-dependent nonempty proper selected set. Suppose that $N_1, N_2 \to \infty$ while P, Q, and $\theta \in (0,1)$ remain fixed. If, for deterministic sequences $a_{N_1,N_2}, r_{N_1,N_2} > 0$ with $r_{N_1,N_2} \to 0$,*[2]

$$\operatorname{err}(S) = O_{\mathbb{P}}(a_{N_1,N_2}), \qquad |\alpha - \theta| = O_{\mathbb{P}}(r_{N_1,N_2}), \tag{13}$$

then

$$\|\widehat{\mu}(S) - \mu_P\|_{\mathcal{H}_{\mathrm{E}}} = O_{\mathbb{P}}\left(a_{N_1,N_2} + N_1^{-1/2}\right), \tag{14}$$

$$\|\widehat{\mu}(R) - \mu_Q\|_{\mathcal{H}_{\mathrm{E}}} = O_{\mathbb{P}}\left(a_{N_1,N_2} + r_{N_1,N_2} + N_1^{-1/2} + N_2^{-1/2}\right). \tag{15}$$

[2]For a random quantity $f(x)$ and deterministic $g(x) > 0$, $f(x) = O_{\mathbb{P}}(g(x))$ means that, for every $\delta > 0$, there exists a constant $C_\delta < \infty$, independent of x, such that $\mathbb{P}\{|f(x)| > C_\delta g(x)\} \le \delta$ for all sufficiently large x in the asymptotic regime under consideration.

Theorem 2 establishes sufficiency. To see that $\alpha \xrightarrow{\mathbb{P}} \theta$ is also necessary for R to consistently approximate Q, suppose that $\mu_P \neq \mu_Q$ and $\|\widehat{\mu}(S) - \mu_P\|_{\mathcal{H}_\mathrm{E}} \xrightarrow{\mathbb{P}} 0$. Since $\bar{\mu}_2 \xrightarrow{\mathbb{P}} \mu_\mathsf{M}$, identity (12) shows that R consistently approximates Q, meaning $\|\widehat{\mu}(R) - \mu_Q\|_{\mathcal{H}_\mathrm{E}} \xrightarrow{\mathbb{P}} 0$, if and only if the selected fraction $\alpha = |S|/N_2$ converges in probability to θ.

Theorem 2 converts a bound on $\mathrm{err}(S)$ into group-error bounds, but it does not show that a greedy prefix has small $\mathrm{err}(S)$. To derive such a bound from Theorem 1, we require a bound on the candidate set's diameter that does not depend on N_2.

Assumption 1. There is a compact set $\mathcal{X} \subseteq \mathbb{R}^d$ such that $\mathrm{supp}(P) \cup \mathrm{supp}(Q) \subseteq \mathcal{X}$.

By the embedding of Φ_E, we have that $C_\mathrm{pool} \leq \sup_{x,x' \in \mathcal{X}} \|x - x'\|_2 < \infty$.

The preceding results combine into the following guarantee for a stopped greedy prefix.

Corollary 3. *Under the mixture sampling model of Section 2.1, suppose that Assumption 1 holds and the greedy rule* (7) *is used. Let $N_1, N_2 \to \infty$ while P, Q, and $\theta \in (0,1)$ remain fixed. Let $1 \leq \widehat{k} < N_2$ be possibly data-dependent. If, for deterministic sequences $e_{N_1,N_2}, r_{N_1,N_2} > 0$ with $r_{N_1,N_2} \to 0$,*

$$\overline{\epsilon}_{\widehat{k}} = O_\mathbb{P}(e_{N_1,N_2}), \qquad |\alpha_{\widehat{k}} - \theta| = O_\mathbb{P}(r_{N_1,N_2}), \tag{16}$$

then

$$\mathrm{err}(S_{\widehat{k}}) = O_\mathbb{P}\Big(e_{N_1,N_2} + N_1^{-1/2} + N_2^{-1/2}\Big), \tag{17}$$

$$\|\widehat{\mu}(S_{\widehat{k}}) - \mu_P\|_{\mathcal{H}_\mathrm{E}} = O_\mathbb{P}\Big(e_{N_1,N_2} + N_1^{-1/2} + N_2^{-1/2}\Big), \tag{18}$$

$$\|\widehat{\mu}(R_{\widehat{k}}) - \mu_Q\|_{\mathcal{H}_\mathrm{E}} = O_\mathbb{P}\Big(e_{N_1,N_2} + r_{N_1,N_2} + N_1^{-1/2} + N_2^{-1/2}\Big). \tag{19}$$

Appendix C proves Theorem 2 and Corollary 3.

3.3 Conditions Controlling $\overline{d}_{\widehat{k}}$

It remains to bound $\overline{\epsilon}_{\widehat{k}}$. Under Assumption 1, we give two settings that control $\overline{d}_{\widehat{k}}$ and hence $\overline{\epsilon}_{\widehat{k}}$. Throughout this subsection, write $\mathcal{H} := \mathcal{H}_\mathrm{E}$ and $\mathrm{diam}_P := \mathrm{diam}_\mathcal{H}(\Phi_\mathrm{E}(\mathrm{supp}(P)))$. Define $\omega : [0,\infty) \to [0,\infty)$ as $\omega(t) = t\log(e/t)$ for $0 < t \leq 1$, $\omega(0) = 0$ and $\omega(t) = 1$ for $t > 1$.

Assumption 2. Let $\mathcal{L}_P = \overline{\mathrm{span}}\{\Phi_\mathrm{E}(x) - \mu_P : x \in \mathrm{supp}(P)\}$. Suppose that, for some $\zeta \geq 0$,

$$\operatorname*{ess\,sup}_{z \sim Q} \|\Pi_{\mathcal{L}_P}\{\Phi_\mathrm{E}(z) - \mu_P\}\|_\mathcal{H} \leq \zeta. \tag{20}$$

Assumption 3. The distribution P has finite support with at least two points. Define

$$g_P := \min_{x \in \mathrm{supp}(P)} \mathrm{dist}_\mathcal{H}(\Phi_\mathrm{E}(x), \mathrm{aff}\{\Phi_\mathrm{E}(x') : x' \in \mathrm{supp}(P),\ x' \neq x\}) > 0,$$

$$p_\mathrm{min} := \min_{x \in \mathrm{supp}(P)} P(\{x\}) > 0, \qquad R_P := \mathrm{diam}_P + \frac{\mathrm{diam}_P^2}{p_\mathrm{min} \cdot g_P}.$$

There is a map $\pi : \mathrm{supp}(Q) \to \mathrm{supp}(P)$ such that, for Q-almost every z, the displacement $b_z := \Phi_\mathrm{E}(z) - \Phi_\mathrm{E}(\pi(z))$ satisfies

$$b_z = 0 \quad \text{or} \quad \|b_z\|_\mathcal{H}^2 > 2R_P\|\Pi_{\mathcal{L}_P} b_z\|_\mathcal{H}. \tag{21}$$

In particular, Assumption 3 holds when b_z is orthogonal to $\mathcal{L}_P$.

Theorem 4. *Under the mixture sampling model of Section 2.1, suppose Assumption 1 holds and the exact greedy rule* (7) *is used. Let $N_1, N_2 \to \infty$ while P, Q, and $\theta \in (0,1)$ remain fixed, and let $r_{N_1,N_2} > 0$ be deterministic with $r_{N_1,N_2} \to 0$. Suppose a possibly data-dependent $\widehat{k} \in \{1, \ldots, N_2 - 1\}$ satisfies $|\alpha_{\widehat{k}} - \theta| = O_\mathbb{P}(r_{N_1,N_2})$. Assume either*

(a) Assumption 2 holds with fixed $\zeta \geq 0$, and $\rho_{N_1,N_2} := \omega(N_1^{-1/2} + N_2^{-1/2} + \zeta)$;
(b) Assumption 3 holds, and $\rho_{N_1,N_2} := N_1^{-1/2} + N_2^{-1/2}$.

Then, in either case,

$$\overline{d}_{\widehat{k}} + \mathrm{err}(S_{\widehat{k}}) + \|\widehat{\mu}(S_{\widehat{k}}) - \mu_P\|_{\mathcal{H}} + \|\widehat{\mu}(R_{\widehat{k}}) - \mu_Q\|_{\mathcal{H}} = O_{\mathbb{P}}(r_{N_1,N_2} + \rho_{N_1,N_2})\,. \tag{22}$$

For $t \in \mathbb{R}$, write $\mathrm{round}(t) := \lfloor t + \frac{1}{2} \rfloor$. In case (a), when $\zeta = 0$, θ is known, $\widehat{k} = \mathrm{round}(\theta N_2)$, and $N_1 \asymp N_2$, each term on the left-hand side of (22) is $O_{\mathbb{P}}((\log N_2)/\sqrt{N_2})$. If $0 < \zeta \leq 1$ is fixed, the bound instead includes a term of order $\zeta \log(e/\zeta)$. In case (b), the same choice of $\widehat{k}$ gives each term the sharper rate $O_{\mathbb{P}}(N_1^{-1/2} + N_2^{-1/2})$. This improvement follows because Assumption 3 ensures that $d_j = 0$ for all but fewer than $L_{\widehat{k}}$ indices $j < \widehat{k}$, as shown in Appendix D.3. Finite support alone does not ensure this property. Appendices D.2 and D.3 give the respective proofs and examples.

4 Experiments

The experiments examine the two requirements in Section 3.2: the EG ordering should contain prefixes that fit the target, and the stopping point should satisfy $\alpha_k \approx \theta$. Motivated by (12), we separately study the effect of varying the stopping point k along a fixed EG order and compare selection methods for a common k. Figure 2 illustrates the rate behaviour predicted by Theorem 4.

Let $\widehat{P}_{\mathrm{eval}}$ and $\widehat{Q}_{\mathrm{eval}}$ denote the empirical distributions of the evaluation samples. For S_k and R_k, we report

$$e_P(k) := \mathrm{MMD}_{k_{\mathrm{E}}}(\widehat{P}_{S_k}, \widehat{P}_{\mathrm{eval}}), \qquad e_Q(k) := \mathrm{MMD}_{k_{\mathrm{E}}}(\widehat{Q}_{R_k}, \widehat{Q}_{\mathrm{eval}}),$$

which estimate the left sides of (18) and (19). We use independent evaluation samples in the Gaussian experiments and held-out component samples in the real-data experiments. Evaluation samples are excluded from preprocessing, selection and stopping.

Simulation. We consider $P = \mathcal{N}(m_P, I_2)$, $Q = \mathcal{N}(0, I_2)$ and $\theta = 0.3$ with $m_P \in \{(2,2), (0.5, 0.5)\}$. The corresponding Euclidean mean gaps, approximately 2.83 and 0.71, produce settings with relatively strong and weak separation. We set $N_1 = N_2 = 3000$ and construct the mixture sample with exactly θN_2 observations from P and $(1-\theta)N_2$ from Q. We evaluate $e_P(k)$ and $e_Q(k)$ with $|\widehat{P}_{\mathrm{eval}}| = |\widehat{Q}_{\mathrm{eval}}| = 3000$. Results are averaged over 50 repetitions.

Figure 1(a)–(b) shows how $e_P(k)$ and $e_Q(k)$ vary with α_k. In the more strongly separated setting (Panel (a)), $e_P(k)$ remains small up to approximately $\alpha_k = \theta$ and increases thereafter. The convex-hull removal loss $\bar{\epsilon}_k$ rises over the same region, illustrating the mechanism captured by (11). In the more overlapping setting (Panel 1(b)), $e_P(k)$ changes little due to substantial overlap, while $e_Q(k)$ still attains a minimum near θ, showing that effective distribution separation remains possible. Panels (c) and (d) further examine the remainder error $e_Q(k)$ under strong and weak separability, respectively, and validate the upper bound in (31). Specifically,

$$U_k = T_{\mathrm{pool}} + T_{\mathrm{prefix}} + T_{\mathrm{cal}} = \frac{1}{1-\alpha_k}\|\bar{\mu}_2 - \mu_M\| + \frac{\alpha_k}{1-\alpha_k} e_P(k) + \frac{|\theta - \alpha_k|}{1-\alpha_k}\|\mu_P - \mu_Q\|.$$

The shape of $e_Q(k)$ closely matches that of the upper bound in (31), which we denote by $U_k^{(31)}$, with the minimum error attained when $\alpha_k = \theta$, validating the separation analysis in Section 3.2.

Real data. We use Page Blocks, Shuttle, and Wine Quality from UCI (Malerba, 1994; UCI Machine Learning Repository, 1992; Cortez et al., 2009). We use component labels to construct samples representing P and Q, but hide these labels from the selection and stopping procedures. Each mixture sample contains N_2 observations with component proportions fixed at θ and $1-\theta$, where $\theta \in \{0.1, 0.3, 0.5\}$. We reserve 35% of the observations from each component for evaluating energy MMD. We report means and standard deviations over 50 paired splits. Appendix F.3 provides further details of the experimental setup and results for Page Blocks and Wine Quality.

Selection methods and stopping rules. EG denotes the energy-greedy ordering in Algorithm 1. We compare three stopping rules along this order. PathMin selects the prefix with the smallest empirical energy discrepancy, resolving ties in favour of the earliest stopping time, as in (8). KM1 and KM2 instead use the mixture-proportion estimators of Ramaswamy et al. (2016) to determine the selected size. We also compare EG with energy-kernel adaptations of Generalized Kernel Herding without Replacement (GKHR) from Shao et al. (2024) and kernel herding without replacement (ED-Herding, abbreviated as EDH). Appendices F.1 and F.2 give the estimator and selection details.

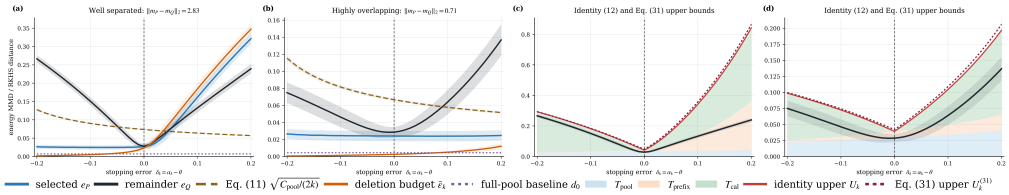


Figure 1: Changing the stopping point alters the selected and remaining distributions.

Table 1: Shuttle results with $N_1 = 4330$, different N_2 sizes, and true mixtures θs. Entries are mean $\pm$ standard deviation across 50 splits. GKHR/KM2 denotes GKHR stopped at the number of points specified by KM2. Bold marks the α_k closest to θ and the smallest $e_P(k)$ and $e_Q(k)$; ties in the displayed means are all bold.

		α_k				$e_P(k)$					$e_Q(k)$				
θ	N_2	EG/PathMin	EDH/PathMin	KM1	KM2	EG/PathMin	EDH/PathMin	EG/KM1	EG/KM2	GKHR/KM2	EG/PathMin	EDH/PathMin	EG/KM1	EG/KM2	GKHR/KM2
0.1	510	0.079 ± 0.009	$\mathbf{0.104 \pm 0.012}$	0.114 ± 0.010	$\mathbf{0.104 \pm 0.011}$	$\mathbf{0.0822 \pm 0.0085}$	0.1415 ± 0.0288	0.1098 ± 0.0201	0.0947 ± 0.0156	0.1513 ± 0.0278	0.0841 ± 0.0220	0.0814 ± 0.0217	0.0826 ± 0.0191	0.0811 ± 0.0194	$\mathbf{0.0803 \pm 0.0194}$
	1030	0.074 ± 0.007	$\mathbf{0.098 \pm 0.005}$	0.105 ± 0.009	0.103 ± 0.011	$\mathbf{0.0547 \pm 0.0051}$	0.0913 ± 0.0121	0.0702 ± 0.0117	0.0682 ± 0.0127	0.1029 ± 0.0115	0.0734 ± 0.0231	$\mathbf{0.0655 \pm 0.0210}$	0.0656 ± 0.0205	0.0657 ± 0.0203	0.0656 ± 0.0204
	1540	0.073 ± 0.006	0.088 ± 0.017	0.098 ± 0.011	$\mathbf{0.101 \pm 0.011}$	$\mathbf{0.0441 \pm 0.0044}$	0.0667 ± 0.0094	0.0532 ± 0.0087	0.0565 ± 0.0117	0.0841 ± 0.0069	0.0688 ± 0.0209	0.0625 ± 0.0196	0.0599 ± 0.0170	$\mathbf{0.0596 \pm 0.0163}$	0.0598 ± 0.0165
0.3	510	0.215 ± 0.018	0.259 ± 0.023	0.344 ± 0.013	$\mathbf{0.303 \pm 0.015}$	$\mathbf{0.0447 \pm 0.0046}$	0.0623 ± 0.0104	0.0978 ± 0.0142	0.0576 ± 0.0077	0.0710 ± 0.0088	0.1483 ± 0.0419	0.1093 ± 0.0367	0.1037 ± 0.0286	0.0915 ± 0.0235	$\mathbf{0.0909 \pm 0.0237}$
	1030	0.202 ± 0.017	0.251 ± 0.015	0.314 ± 0.009	$\mathbf{0.303 \pm 0.009}$	$\mathbf{0.0338 \pm 0.0045}$	0.0422 ± 0.0044	0.0508 ± 0.0081	0.0434 ± 0.0062	0.0515 ± 0.0051	0.1547 ± 0.0355	0.1037 ± 0.0313	0.0758 ± 0.0206	$\mathbf{0.0748 \pm 0.0208}$	0.0752 ± 0.0212
	1540	0.202 ± 0.014	0.248 ± 0.013	0.308 ± 0.008	$\mathbf{0.301 \pm 0.009}$	$\mathbf{0.0299 \pm 0.0044}$	0.0351 ± 0.0036	0.0412 ± 0.0064	0.0377 ± 0.0059	0.0445 ± 0.0049	0.1527 ± 0.0306	0.1024 ± 0.0275	0.0646 ± 0.0180	$\mathbf{0.0644 \pm 0.0183}$	0.0652 ± 0.0184
0.5	510	0.340 ± 0.027	0.396 ± 0.035	0.651 ± 0.043	$\mathbf{0.501 \pm 0.019}$	$\mathbf{0.0363 \pm 0.0046}$	0.0436 ± 0.0060	0.1725 ± 0.0368	0.0472 ± 0.0066	0.0527 ± 0.0066	0.2836 ± 0.0492	0.2134 ± 0.0586	0.3441 ± 0.1217	$\mathbf{0.1095 \pm 0.0272}$	0.1096 ± 0.0271
	1030	0.338 ± 0.022	0.394 ± 0.033	0.540 ± 0.011	$\mathbf{0.499 \pm 0.012}$	$\mathbf{0.0293 \pm 0.0046}$	0.0327 ± 0.0042	0.0604 ± 0.0092	0.0364 ± 0.0054	0.0403 ± 0.0051	0.2838 ± 0.0424	0.2100 ± 0.0541	0.1060 ± 0.0362	$\mathbf{0.0874 \pm 0.0240}$	0.0888 ± 0.0241
	1540	0.332 ± 0.022	0.387 ± 0.028	0.523 ± 0.008	$\mathbf{0.502 \pm 0.009}$	$\mathbf{0.0271 \pm 0.0049}$	0.0288 ± 0.0045	0.0420 ± 0.0058	0.0326 ± 0.0049	0.0356 ± 0.0046	0.2915 ± 0.0369	0.2204 ± 0.0432	0.0826 ± 0.0223	$\mathbf{0.0761 \pm 0.0208}$	0.0774 ± 0.0207

Table 1 illustrates the effects of selection and stopping. EG/PathMin achieves the smallest mean $e_P(k)$ in every displayed setting, but tends to stop too early, with α_k below θ, particularly when $\theta \in \{0.3, 0.5\}$. This is unsurprising; as a greedy method, EG/PathMin is designed to minimize the energy-distance discrepancy along the selection path, and therefore tends to stop at an early empirical minimum rather than provide an accurate estimate of θ. However, the $e_Q(k)$ results indicate that its remainder approximation can be poor. In contrast, EG/KM2 performs best in $e_Q(k)$, yielding the closest approximation of the remainder to Q, while its $e_P(k)$ remains close to that of EG/PathMin. Moreover, the baseline GKHR/KM2 is generally inferior to EG/KM2 in both $e_P(k)$ and $e_Q(k)$. These results empirically support our central claim: combining energy-distance greedy selection with an accurate stopping size k can effectively separate heterogeneous mixture data.

Next, we examine Theorem 4 using the continuous energy-kernel construction in Appendix D.2. We set $N_1 = N_2 = N$ and stop at $k = \theta N$, so that the stopping-fraction error r_{N_1,N_2} is zero. Appendix F.4 gives the constructions of the target distribution P, the exactly orthogonal contaminant Q_0, and the perturbed contaminant Q_{ϕ_N}. Here, $\phi_N = 2\arcsin(\zeta_N^2/2)$ and $\zeta_N \in \{0, N^{-1/2}, N^{-1/4}\}$, such that (20) holds with $\zeta = \zeta_N$. The case $\zeta_N = 0$ corresponds to $\phi_N = 0$ and hence to Q_0. Thus the rate $\omega(2N^{-1/2} + \zeta_N)$ has order $N^{-1/2}\log N$ when $\zeta_N \in \{0, N^{-1/2}\}$ and $N^{-1/4}\log N$ when $\zeta_N = N^{-1/4}$. Figure 2(a) shows decreasing errors under exact orthogonality, compatible with the theorem's $O_{\mathbb{P}}(N^{-1/2}\log N)$ upper bound. Figure 2(b) shows the ratio of the empirical 90% quantile of T_N to the theoretical rate $\omega(2N^{-1/2} + \zeta_N)$. For $\zeta_N = N^{-1/4}$, this ratio decreases over the sample sizes considered. The unnormalized errors remain comparable to those under exact orthogonality, suggesting a decay pattern closer to $N^{-1/2}\log N$ than to $N^{-1/4}\log N$ over the tested range. These observations suggest that our bound is conservative for this construction.

5 Related Work

Greedy MMD selection. The closest work differs in two respects: whether the target is the empirical mean of the candidate pool or an external distribution, and whether selected points may be chosen again. Table 9 in Appendix G.1 summarizes the comparison. We study selection without replacement for an external target and track the convex hull of the remaining candidates.

When the target is the mean of the candidate pool, $u = n^{-1}\sum_j x_j$, the squared objective $\mathrm{err}^2(S)$ and greedy prototype sequence coincide with those of MMD-Critic, up to tie breaking. Our contribution is therefore in the analysis of this rule. The exact one-step rule of Teymur et al. (2021) uses the same update over a fixed finite candidate set but allows candidates to be selected more than once. Their bounds show that squared MMD is at most the minimum squared error achievable by a convex combination of the fixed candidates, plus an $O((1+\log k)/k)$ term. Pronzato (2023) analyses closely

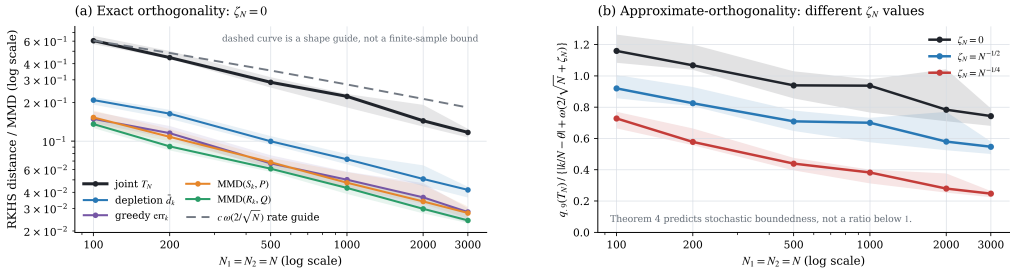


Figure 2: Experiments illustrating Theorem 4. In panel (a), T_N is the left-hand side of Equation (22).

related greedy-MMD and kernel-herding algorithms on a fixed domain. By contrast, Shao et al. (2024) remove selected indices but use the full empirical sample as the target; at their parameter value $\alpha_{\mathrm{H}} = 1$, their selection criterion produces the same selection sequence when the target is the candidate mean. Their $O(\log k/k)$ squared-error result assumes that sufficiently many favorable candidates remain after every selection step and is intended for settings where only a small subset is selected.

The distinction developed here is between the target's initial distance from the convex hull of the entire candidate set, $d_0 = \mathrm{dist}(u, \mathrm{conv}\{x_i : i \in [n]\})$, and its distance from the convex hull of the remaining candidates after k selections, $d_k = \mathrm{dist}(u, \mathcal{C}_k)$; their difference $\epsilon_k = d_k - d_0$ isolates the additional loss caused by earlier selections. The main sequence bound permits an arbitrary external target and quantifies how well the remaining candidates can still represent it, rather than requiring $d_k = 0$. The first bound retains the realized d_j and κ_j along the path; its global consequence depends only on d_0, $\bar{\epsilon}_k$, and C_{pool}. The sharpness results show that $d_0 = 0$ alone does not guarantee a convergence rate, that the coefficient of $\bar{\epsilon}_k$ cannot be reduced, and that the $O(1/k)$ squared-error order when $\bar{\epsilon}_k = 0$ cannot be improved in the worst case. We also give explicit projection and finite-support conditions under which the removal loss remains small, with convergence rates and energy-kernel examples in Section 3.3 and its appendices.

Kernel herding. Kernel herding greedily matches a target mean embedding and is a conditional-gradient method on a fixed convex domain (Chen et al., 2010; Bach et al., 2012). The standard search domain does not shrink and a selected point may be reused, so its convergence guarantees do not directly describe the first k distinct selections. Support points minimize energy distance to construct representative point sets (Mak & Joseph, 2018), and kernel thinning reduces an empirical distribution to a smaller sample while controlling integration error for functions in the reproducing kernel Hilbert space (Dwivedi & Mackey, 2024). These methods motivate representative subset selection but do not address when the remainder after a chosen stopping point represents a second mixture component.

Sparse convex approximation. Approximate Carathéodory and Maurey-type results approximate a point using a convex combination of a small number of points, with error bounds that do not depend on the dimension (Barman, 2018; Mirrokni et al., 2017); their sampling arguments may repeat points, and general sparse formulations may use nonuniform weights. Frank–Wolfe analyses likewise optimize over a fixed convex set (Combettes & Pokutta, 2023). These results do not directly bound the error of a greedy procedure that adds one previously unselected candidate at each step and assigns equal weight to all selected candidates.

Mixture fractions and stopping. Mixture-proportion estimation recovers the weight of a known component in a mixture. Without irreducibility, the identifiable quantity is the largest component proportion consistent with the observed mixture rather than necessarily the true mixture weight (Blanchard et al., 2010; Scott, 2015). The kernel-mean estimator of Ramaswamy et al. (2016) motivates the practical KM1 and KM2 stopping baselines used in our experiments. We do not propose a new mixture-proportion estimator or establish guarantees for the number of candidates selected by these stopping rules. In positive–unlabeled learning, the same proportion is the probability that an example belongs to the positive class and is used to train a classifier (Scott, 2015). Here, it determines when to stop selecting candidates.

AI USE DISCLOSURE

During the preparation of this work, the authors used ChatGPT 5.6 Sol in drafting assistance, editing and proofreading, research exploration, data analysis and coding. All AI-generated contents have been reviewed, edited and revised by the authors. The authors take full responsibility for the content of this manuscript.

REFERENCES

Francis Bach, Simon Lacoste-Julien, and Guillaume Obozinski. On the equivalence between herding and conditional gradient algorithms. In *Proceedings of the 29th International Conference on Machine Learning*, 2012. URL https://arxiv.org/abs/1203.4523.

Siddharth Barman. Approximating nash equilibria and dense subgraphs via an approximate version of carathéodory's theorem. *SIAM Journal on Computing*, 47(3):960–981, 2018. doi: 10.1137/15M1050574.

Jessa Bekker and Jesse Davis. Estimating the class prior in positive and unlabeled data through decision tree induction. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

Jessa Bekker and Jesse Davis. Learning from Positive and Unlabeled Data: A Survey. *Machine Learning*, 109(4):719–760, 2020. doi: 10.1007/s10994-020-05877-5. URL https://link.springer.com/article/10.1007/s10994-020-05877-5.

Gilles Blanchard, Gyemin Lee, and Clayton Scott. Semi-Supervised Novelty Detection. *Journal of Machine Learning Research*, 11(99):2973–3009, 2010. URL https://www.jmlr.org/papers/v11/blanchard10a.html.

Yutian Chen, Max Welling, and Alex Smola. Super-samples from kernel herding. In *Proceedings of the Twenty-Sixth Conference on Uncertainty in Artificial Intelligence*. AUAI Press, 2010. URL https://event.cwi.nl/uai2010/papers/UAI2010_0238.pdf.

Cyrille W. Combettes and Sebastian Pokutta. Revisiting the approximate Carathéodory problem via the Frank–Wolfe algorithm. *Mathematical Programming*, 197:191–214, 2023. doi: 10.1007/s10107-021-01735-x.

Paulo Cortez, António Cerdeira, Fernando Almeida, Telmo Matos, and José Reis. Wine quality. UCI Machine Learning Repository, 2009. URL https://archive.ics.uci.edu/dataset/186/wine+quality.

Marthinus C. du Plessis, Gang Niu, and Masashi Sugiyama. Class-prior Estimation for Learning from Positive and Unlabeled Data. *Machine Learning*, 106(4):463–492, 2017. doi: 10.1007/s10994-016-5604-6. URL https://link.springer.com/article/10.1007/s10994-016-5604-6.

Raaz Dwivedi and Lester Mackey. Kernel thinning. *Journal of Machine Learning Research*, 25:1–77, 2024. URL https://www.jmlr.org/papers/volume25/21-1334/21-1334.pdf.

Charles Elkan and Keith Noto. Learning Classifiers from Only Positive and Unlabeled Data. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 213–220. ACM, 2008. doi: 10.1145/1401890.1401920. URL https://cseweb.ucsd.edu/~elkan/posonly.pdf.

Saurabh Garg, Yifan Wu, Alexander J. Smola, Sivaraman Balakrishnan, and Zachary Lipton. Mixture Proportion Estimation and PU Learning: A Modern Approach. In *Advances in Neural Information Processing Systems*, volume 34, pp. 8532–8544, 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/hash/47b4f1bfdf6d298682e610ad74b37dca-Abstract.html.

Dmitry Ivanov. DEDPUL: Difference-of-Estimated-Densities-based Positive-Unlabeled Learning. In *2020 19th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pp. 782–790. IEEE, 2020. doi: 10.1109/ICMLA51294.2020.00128. URL https://arxiv.org/abs/1902.06965.

Been Kim, Rajiv Khanna, and Oluwasanmi O. Koyejo. Examples are not enough, learn to criticize! criticism for interpretability. In *Advances in Neural Information Processing Systems*, volume 29, pp. 2280–2288. Curran Associates, Inc., 2016. URL https://proceedings.neurips.cc/paper/2016/hash/5680522b8e2bb01943234bce7bf84534-Abstract.html.

Ryuichi Kiryo, Gang Niu, Marthinus C. du Plessis, and Masashi Sugiyama. Positive-Unlabeled Learning with Non-Negative Risk Estimator. In *Advances in Neural Information Processing Systems*, volume 30, pp. 1675–1685, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/7cce53cf90577442771720a370c3c723-Abstract.html.

Simon Mak and V. Roshan Joseph. Support points. *The Annals of Statistics*, 46(6A):2562–2592, 2018. doi: 10.1214/17-AOS1629.

Donato Malerba. Page blocks classification. UCI Machine Learning Repository, 1994. URL https://archive.ics.uci.edu/dataset/78/page+blocks+classification.

Vahab Mirrokni, Renato Paes Leme, Adrian Vladu, and Sam Chiu-wai Wong. Tight bounds for approximate Carathéodory and beyond. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 2440–2448. PMLR, 2017. URL https://proceedings.mlr.press/v70/mirrokni17a.html.

Luc Pronzato. Performance analysis of greedy algorithms for minimising a maximum mean discrepancy. *Statistics and Computing*, 33(1):14, 2023. doi: 10.1007/s11222-022-10184-1.

Harish G. Ramaswamy, Clayton Scott, and Ambuj Tewari. Mixture proportion estimation via kernel embeddings of distributions. In *Proceedings of the 33rd International Conference on Machine Learning*, volume 48 of *JMLR: W&CP*, pp. 2052–2060. PMLR, 2016. URL https://proceedings.mlr.press/v48/ramaswamy16.html.

Clayton Scott. A rate of convergence for mixture proportion estimation, with application to learning from noisy labels. In *Proceedings of the 18th International Conference on Artificial Intelligence and Statistics*, volume 38 of *Proceedings of Machine Learning Research*, pp. 838–846. PMLR, 2015. URL https://proceedings.mlr.press/v38/scott15.html.

Dino Sejdinovic, Bharath Sriperumbudur, Arthur Gretton, and Kenji Fukumizu. Equivalence of distance-based and RKHS-based statistics in hypothesis testing. *The Annals of Statistics*, 41(5): 2263–2291, 2013. doi: 10.1214/13-AOS1140.

Qian Shao, Jiangrui Kang, Qiyuan Chen, Zepeng Li, Hongxia Xu, Yiwen Cao, Jiajuan Liang, and Jian Wu. Enhancing semi-supervised learning via representative and diverse sample selection. In *Advances in Neural Information Processing Systems*, volume 37, pp. 111199–111226, 2024. doi: 10.52202/079017-3531.

Gábor J. Székely and Maria L. Rizzo. Energy statistics: A class of statistics based on distances. *Journal of Statistical Planning and Inference*, 143(8):1249–1272, 2013. doi: 10.1016/j.jspi.2013.03.018.

Onur Teymur, Jackson Gorham, Marina Riabiz, and Chris J. Oates. Optimal quantisation of probability measures using maximum mean discrepancy. In *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pp. 1027–1035. PMLR, 2021. URL https://proceedings.mlr.press/v130/teymur21a.html.

UCI Machine Learning Repository. Statlog (shuttle). UCI Machine Learning Repository, 1992. URL https://archive.ics.uci.edu/dataset/148/statlog+shuttle.

Yilun Zhu, Aaron Fjeldsted, Darren Holland, George Landon, Azaree Lintereur, and Clayton Scott. Mixture Proportion Estimation Beyond Irreducibility. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 42962–42982. PMLR, 2023. URL https://proceedings.mlr.press/v202/zhu23c.html.

A Missing Details of Section 2.3

This section gives the Euclidean formulas and implementation details for the energy-greedy ordering in Section 2.3.

A.1 Pairwise Objective and Greedy Ordering

If $k := |S|$, the squared energy distance between the selected and reference empirical measures is

$$\mathcal{E}(\widehat{P}_S, \widehat{P}_{N_1}) = \frac{2}{kN_1} \sum_{j\in S} \sum_{i=1}^{N_1} \|z_j - y_i\|_2 - \frac{1}{k^2} \sum_{j,\ell\in S} \|z_j - z_\ell\|_2 - \frac{1}{N_1^2} \sum_{i=1}^{N_1} \sum_{\ell=1}^{N_1} \|y_i - y_\ell\|_2. \quad (23)$$

In particular, the first selected index has the Euclidean form

$$i_1 \in \arg\min_{j\in[N_2]} \mathrm{err}(\{j\}) = \arg\min_{j\in[N_2]} \frac{2}{N_1} \sum_{i=1}^{N_1} \|z_j - y_i\|_2. \quad (24)$$

The reference–reference term in (23) is constant in j. Multiplication by a positive constant and taking the square root do not change the minimizer, so these operations and the constant term are omitted in (24). The deterministic tie-breaking rule in Section 2.2 applies throughout.

Algorithm 1 Energy-Greedy Ordering Without Replacement

Require: Reference data $D_1 = (y_i)_{i=1}^{N_1}$, candidate data $D_2 = (z_j)_{j=1}^{N_2}$
Ensure: Ordered indices $(i_1, \ldots, i_{N_2})$; the prefixes $S_k := \{i_1, \ldots, i_k\}$ and complements $R_k := [N_2] \setminus S_k$ are implicit
1: Select i_1 using (24).
2: $S_1 \leftarrow \{i_1\}$ and $R_1 \leftarrow [N_2] \setminus S_1$.
3: **for** $k = 1, \ldots, N_2 - 1$ **do**
4: Select i_{k+1} using (7).
5: $S_{k+1} \leftarrow S_k \cup \{i_{k+1}\}$ and $R_{k+1} \leftarrow R_k \setminus \{i_{k+1}\}$.
6: **end for**
7: **return** $(i_1, \ldots, i_{N_2})$.

Only the current membership information is retained while constructing the order. Any selected set S_k and remainder R_k can subsequently be recovered from the first k indices, so returning all set pairs would add unnecessary quadratic storage.

A.2 Implementation of Algorithm

The greedy path can be computed without reevaluating $\mathrm{err}(S_k \cup \{j\})$ from scratch. Precompute

$$a_j := \frac{1}{N_1} \sum_{i=1}^{N_1} \|z_j - y_i\|_2, \qquad j \in [N_2],$$

and, for $j \in R_k$, maintain

$$b_{k,j} := \sum_{\ell\in S_k} \|z_j - z_\ell\|_2.$$

By (23), the next index minimizes

$$a_j - \frac{b_{k,j}}{k+1}, \qquad j \in R_k. \quad (25)$$

After selecting i_{k+1}, update

$$b_{k+1,j} := b_{k,j} + \|z_j - z_{i_{k+1}}\|_2, \qquad j \in R_{k+1}.$$

Precomputing $(a_j)_{j=1}^{N_2}$ takes $O(N_1N_2)$ distance evaluations. The scans and updates through the full path take $O(N_2^2)$ additional work and distance evaluations. Thus the total cost is $O(N_1N_2 + N_2^2)$ with $O(N_2)$ additional memory. Storing all candidate pairwise distances is unnecessary and would instead require $O(N_2^2)$ memory.

The PathMin scores can be maintained in parallel. Define

$$A_k := \sum_{j \in S_k} a_j, \qquad W_k := \sum_{j,\ell \in S_k} \|z_j - z_\ell\|_2.$$

Then

$$\mathrm{err}^2(S_k) = \frac{A_k}{k} - \frac{W_k}{2k^2} - \frac{1}{2N_1^2} \sum_{i=1}^{N_1} \sum_{\ell=1}^{N_1} \|y_i - y_\ell\|_2. \tag{26}$$

The last term is constant across prefixes, so PathMin only needs the score $A_k/k - W_k/(2k^2)$. The recursions

$$A_{k+1} := A_k + a_{i_{k+1}}, \qquad W_{k+1} := W_k + 2b_{k,i_{k+1}}$$

therefore compute every prefix score using quantities already available in the greedy update.

B Missing Proofs and Further Discussions for Section 3.1

For $1 \le k \le n$, put

$$v_k := \frac{1}{k} \sum_{t=1}^{k} x_{i_t}, \qquad \mathrm{err}_k = \|u - v_k\|_{\mathcal{H}}.$$

For $1 \le k < n$, the point that would make the next equal-weight average equal to u is

$$t_k := (k+1)u - kv_k.$$

For a hypothetical next choice $i \in R_k$, let $v_{k+1}(i) := (kv_k + x_i)/(k+1)$. Then $(k+1)(u - v_{k+1}(i)) = t_k - x_i$. Consequently, the exact greedy update satisfies the correction-point identity

$$(k+1)^2 \,\mathrm{err}_{k+1}^2 = \min_{i \in R_k} \|t_k - x_i\|_{\mathcal{H}}^2. \tag{27}$$

Define

$$\Delta_k := \mathrm{dist}(t_k, \mathcal{C}_k), \qquad p_k := \Pi_{\mathcal{C}_k}(u).$$

Lemma 5. *For every $1 \le k < n$,*

$$\mathrm{err}_{k+1}^2 \le \frac{\|k(u - v_k) + (u - p_k)\|_{\mathcal{H}}^2 + \kappa_k}{(k+1)^2} \le \frac{\left(k\,\mathrm{err}_k + d_k\right)^2 + \kappa_k}{(k+1)^2}. \tag{28}$$

Proof. Let $q_k = \Pi_{\mathcal{C}_k}(t_k)$ and write $q_k = \sum_{i \in R_k} \beta_i x_i$. By (27) and the weighted variance identity,

$$\begin{aligned}
(k+1)^2 \,\mathrm{err}_{k+1}^2 &\le \sum_{i \in R_k} \beta_i \|t_k - x_i\|_{\mathcal{H}}^2 \\
&= \Delta_k^2 + \frac{1}{2} \sum_{i,j \in R_k} \beta_i \beta_j \|x_i - x_j\|_{\mathcal{H}}^2 \\
&\le \Delta_k^2 + \frac{1}{2}\,\mathrm{diam}_{\mathcal{H}}^2(\mathcal{C}_k) \left(1 - \frac{1}{n-k}\right) \\
&= \Delta_k^2 + \kappa_k.
\end{aligned}$$

The last inequality uses $\sum_{i \in R_k} \beta_i^2 \ge 1/(n-k)$. Moreover,

$$\Delta_k \le \|t_k - p_k\|_{\mathcal{H}} = \|k(u - v_k) + (u - p_k)\|_{\mathcal{H}}.$$

Then the first inequality in (28) follows, and the second inequality follows from the triangle inequality. □

Proof of Theorem 1. Set $A_j := j\,\mathrm{err}_j$. Lemma 5 gives $A_{j+1} \le \sqrt{(A_j + d_j)^2 + \kappa_j}$. The inequality $\sqrt{(x+y)^2+z} \le \sqrt{x^2+z} + y$ for $x, y, z \ge 0$ yields by induction

$$A_k \le \sqrt{\mathrm{err}_1^2 + \sum_{j=1}^{k-1} \kappa_j} + \sum_{j=1}^{k-1} d_j.$$

Dividing by k proves the first upper bound in (11). Since $v_k \in \mathcal{C}_0$, one also has $\mathrm{err}_k \ge d_0$. To control the initial error, write $p_0 = \sum_{i=1}^n \beta_i x_i$ as a convex combination. First-point optimality and the weighted variance identity give

$$\mathrm{err}_1^2 \le \sum_{i=1}^n \beta_i \|u - x_i\|_{\mathcal{H}}^2 \le d_0^2 + \kappa_0.$$

Consequently,

$$\frac{\sqrt{\mathrm{err}_1^2 + \sum_{j=1}^{k-1} \kappa_j}}{k} \le \frac{d_0 + \sqrt{\sum_{j=0}^{k-1} \kappa_j}}{k} \le \frac{d_0}{k} + \sqrt{\frac{C_{\mathrm{pool}}}{2k}},$$

where the last step uses $\sum_{j=0}^{k-1} \kappa_j \le k C_{\mathrm{pool}}/2$. Finally use the identity in (10) to complete (11). □

Corollary 6. *Suppose $2 \le k \le n$ and $d_{k-1} = d_0$. Then*

$$0 \le \mathrm{err}_k - d_0 \le \frac{\sqrt{d_0^2 + \sum_{j=0}^{k-1} \kappa_j} - d_0}{k}. \tag{29}$$

In particular, if $d_0 = 0$, then

$$\mathrm{err}_k^2 \le \frac{\sum_{j=0}^{k-1} \kappa_j}{k^2} \le \frac{C_{\mathrm{pool}}}{2k}. \tag{30}$$

Proof. Monotonicity gives $d_j = d_0$ and $\epsilon_j = 0$ for every $j < k$. The first upper bound in Theorem 1 and $\mathrm{err}_1^2 \le d_0^2 + \kappa_0$ give (29). When $d_0 = 0$, square the result and use $\kappa_j \le C_{\mathrm{pool}}/2$. □

C Missing Proofs and Further Discussions for Section 3.2

This appendix gives supporting sampling bounds and proofs for the combination of Sections 3.1 and 3.2.

We first give general stopping certificates. Let $S \subset [N_2]$ be nonempty and proper, set $R := [N_2] \setminus S$ and $\alpha := |S|/N_2$, and use the notation surrounding the partition identity (12). Since $\|u - \widehat{\mu}(S)\|_{\mathcal{H}_{\mathrm{E}}} = \mathrm{err}(S)$, that identity implies

$$\begin{aligned}\|\widehat{\mu}(R) - \mu_Q\|_{\mathcal{H}_{\mathrm{E}}} \le \frac{\|\bar{\mu}_2 - \mu_{\mathsf{M}}\|_{\mathcal{H}_{\mathrm{E}}}}{1-\alpha} + \frac{\alpha(\mathrm{err}(S) + \|\mu_P - u\|_{\mathcal{H}_{\mathrm{E}}})}{1-\alpha} \\ + \frac{|\theta - \alpha|\,\|\mu_P - \mu_Q\|_{\mathcal{H}_{\mathrm{E}}}}{1-\alpha}.\end{aligned} \tag{31}$$

The three terms account for mixture-sample error, prefix approximation error, and stopping-fraction error, respectively. For a greedy prefix $S = S_k$, with $\alpha = \alpha_k$, we write

$$U_k := T_{\mathrm{pool}} + T_{\mathrm{prefix}} + T_{\mathrm{cal}},$$

with all three terms evaluated at $S = S_k$. Appendix E.1.1 shows that its three coefficients are separately attainable.

Proof of Theorem 2. Equation (14) follows from

$$\|\widehat{\mu}(S) - \mu_P\|_{\mathcal{H}_{\mathrm{E}}} \le \mathrm{err}(S) + \|u - \mu_P\|_{\mathcal{H}_{\mathrm{E}}}.$$

The second condition in (13) gives $\alpha \xrightarrow{\mathbb{P}} \theta < 1$, so $(1-\alpha)^{-1} = O_{\mathbb{P}}(1)$. The result follows from (31) and the rates $\|u - \mu_P\|_{\mathcal{H}_{\mathrm{E}}} = O_{\mathbb{P}}(N_1^{-1/2})$ and $\|\bar{\mu}_2 - \mu_{\mathsf{M}}\|_{\mathcal{H}_{\mathrm{E}}} = O_{\mathbb{P}}(N_2^{-1/2})$. □

Next we consider a finite-sample bound that holds uniformly over all possibly data-dependent selected sets S. Under Assumption 1, let

$$B_{\mathcal{X}} := \sup_{x\in\mathcal{X}} \|\Phi_{\mathrm{E}}(x)\|_{\mathcal{H}_{\mathrm{E}}} = \sqrt{\sup_{x\in\mathcal{X}} \|x\|_2},$$
$$C_{\mathcal{X}} := \sup_{x,x'\in\mathcal{X}} \|\Phi_{\mathrm{E}}(x) - \Phi_{\mathrm{E}}(x')\|^2_{\mathcal{H}_{\mathrm{E}}} = \sup_{x,x'\in\mathcal{X}} \|x - x'\|_2.$$

Both constants are finite.

Fix $\delta \in (0,1)$, let $\Delta_{P,Q} := \|\mu_P - \mu_Q\|_{\mathcal{H}_{\mathrm{E}}}$, and suppose that

$$\Pr\{\|u - \mu_P\|_{\mathcal{H}_{\mathrm{E}}} \le r_P(\delta) \text{ and } \|\bar{\mu}_2 - \mu_{\mathsf{M}}\|_{\mathcal{H}_{\mathrm{E}}} \le r_{\mathsf{M}}(\delta)\} \ge 1 - \delta. \tag{32}$$

On this event, simultaneously for every (possibly data-dependent) nonempty proper S,

$$\begin{aligned} \|\widehat{\mu}(S) - \mu_P\|_{\mathcal{H}_{\mathrm{E}}} &\le \mathrm{err}(S) + r_P(\delta), \\ \|\widehat{\mu}(R) - \mu_Q\|_{\mathcal{H}_{\mathrm{E}}} &\le \frac{r_{\mathsf{M}}(\delta) + \alpha(\mathrm{err}(S) + r_P(\delta)) + |\alpha - \theta|\Delta_{P,Q}}{1 - \alpha}. \end{aligned} \tag{33}$$

If the same event also gives $|\alpha - \theta| \le r_\theta < 1 - \theta$, then

$$\|\widehat{\mu}(R) - \mu_Q\|_{\mathcal{H}_{\mathrm{E}}} \le \frac{r_{\mathsf{M}}(\delta) + \alpha(\mathrm{err}(S) + r_P(\delta)) + r_\theta\Delta_{P,Q}}{1 - \theta - r_\theta}. \tag{34}$$

If an externally justified estimator satisfies $|\widetilde{\theta} - \theta| = O_{\mathbb{P}}(s_{N_1,N_2})$ with $s_{N_1,N_2} \to 0$, define

$$\widehat{k} := \min\Big\{N_2 - 1, \max\Big\{2, \mathrm{round}(N_2\widetilde{\theta})\Big\}\Big\}. \tag{35}$$

Because $\theta \in (0,1)$, for all sufficiently large N_2, $\theta \in [2/N_2, 1 - 1/N_2]$. Equation (35) truncates the rounded fraction to this interval, so

$$|\alpha_{\widehat{k}} - \theta| \le \left|\frac{\mathrm{round}(N_2\widetilde{\theta})}{N_2} - \theta\right| \le |\widetilde{\theta} - \theta| + \frac{1}{2N_2}.$$

No independence from the selected sequence is required. Under unrestricted P, Q, however, θ need not be identifiable, and we supply no estimator.

For completeness, we shall make the radii $r_P(\delta)$ and $r_{\mathsf{M}}(\delta)$ explicit by a standard bounded difference argument. Let $X_1, \dots, X_N$ be i.i.d. random variables, ϕ be a function taking values in a Hilbert space $\mathcal{H}$, and suppose $\|\phi(X_i)\|_{\mathcal{H}} \le B$ almost surely. Define

$$\mu := \mathbb{E}\,\phi(X_1), \qquad \widehat{\mu}_N := \frac{1}{N}\sum_{i=1}^{N}\phi(X_i), \qquad q_N(\delta; B) := \frac{B(1 + \sqrt{2\log(1/\delta)})}{\sqrt{N}}.$$

Then, for every $\delta \in (0,1)$,

$$\Pr(\|\widehat{\mu}_N - \mu\|_{\mathcal{H}} \le q_N(\delta; B)) \ge 1 - \delta. \tag{36}$$

Indeed, the Hilbert-space second-moment identity and Jensen's inequality give $\mathbb{E}\,\|\widehat{\mu}_N - \mu\|_{\mathcal{H}} \le B/\sqrt{N}$. Observe that changing the value of one X_i changes $\|\widehat{\mu}_N - \mu\|_{\mathcal{H}}$ by at most $2B/N$, then (36) follows from McDiarmid's inequality.

For the energy feature map Φ_{E} under Assumption 1, applying (36) to the reference and candidate samples and then taking a union bound shows that (32) holds with

$$r_P(\delta) = q_{N_1}(\delta/2; B_{\mathcal{X}}), \qquad r_{\mathsf{M}}(\delta) = q_{N_2}(\delta/2; B_{\mathcal{X}}).$$

Lemma 7. *Assume the mixture sampling model of Section 2.1 and suppose that $\theta \in (0,1)$ is fixed. Then*

$$\mathbb{E}\,d_0^2 \le \frac{\sigma_P^2}{N_1} + \frac{2\sigma_P^2}{(N_2+1)\theta} + \left(\sigma_Q^2 + \Delta_{P,Q}^2\right)(1-\theta)^{N_2}, \tag{37}$$

where

$$\sigma_P^2 := \mathop{\mathbb{E}}_{P}\|\Phi_{\mathrm{E}}(X) - \mu_P\|^2_{\mathcal{H}_{\mathrm{E}}}, \qquad \sigma_Q^2 := \mathop{\mathbb{E}}_{Q}\|\Phi_{\mathrm{E}}(X) - \mu_Q\|^2_{\mathcal{H}_{\mathrm{E}}}.$$

Consequently, $d_0 = O_{\mathbb{P}}(N_1^{-1/2} + N_2^{-1/2})$ as $N_1, N_2 \to \infty$.

Proof. Let $I_P := \{j : z_j \text{ was generated from } P\}$ and $M := |I_P|$. When $M \geq 1$, let

$$\bar{\mu}_{P,M} := \frac{1}{M} \sum_{j \in I_P} \Phi_{\mathrm{E}}(z_j).$$

When $M \geq 1$, the point $\bar{\mu}_{P,M}$ is a convex combination of candidate feature vectors and hence belongs to $\mathcal{C}_0$. Therefore

$$d_0 \leq \|u - \bar{\mu}_{P,M}\|_{\mathcal{H}_{\mathrm{E}}}.$$

Conditional on $M = m \geq 1$, the P-origin candidate points are iid from P and independent of the reference sample. The Hilbert-space second-moment identity therefore gives

$$\mathbb{E}\big[\|u - \bar{\mu}_{P,M}\|^2_{\mathcal{H}_{\mathrm{E}}} \,\big|\, M = m\big] = \sigma_P^2 \left(\frac{1}{N_1} + \frac{1}{m}\right).$$

Since $M \sim \mathrm{Binomial}(N_2, \theta)$,

$$\mathbb{E}\left[\frac{\mathbf{1}\{M > 0\}}{M}\right] \leq 2\,\mathbb{E}\left[\frac{1}{M+1}\right] = \frac{2\{1 - (1-\theta)^{N_2+1}\}}{(N_2+1)\theta} \leq \frac{2}{(N_2+1)\theta}.$$

When $M = 0$, all candidates come from Q, so $d_0 \leq \|u - \Phi_{\mathrm{E}}(z_1)\|_{\mathcal{H}_{\mathrm{E}}}$. Conditioned on this event,

$$\mathbb{E}\,\|u - \Phi_{\mathrm{E}}(z_1)\|^2_{\mathcal{H}_{\mathrm{E}}} = \frac{\sigma_P^2}{N_1} + \sigma_Q^2 + \Delta^2_{P,Q}.$$

Combining the two events proves (37). The σ_P^2/N_1 contributions combine because the two events partition the sample space. Markov's inequality gives the stated stochastic rate because θ is fixed. □

Proof of Corollary 3. Since $\alpha_{\widehat{k}} \xrightarrow{\mathbb{P}} \theta > 0$, $\Pr(\widehat{k} \geq 2) \to 1$ and $\widehat{k} \asymp N_2$ in probability. Conditioned on the event $\{\widehat{k} \geq 2\}$, it follows from Lemma 7, Theorem 1 and $C_{\mathrm{pool}} \leq C_{\mathcal{X}}$ that

$$\mathrm{err}(S_{\widehat{k}}) = O_{\mathbb{P}}\Big(e_{N_1,N_2} + N_1^{-1/2} + N_2^{-1/2}\Big).$$

Since $\Pr(\widehat{k} \geq 2) \to 1$, this rate also holds unconditionally. Apply Theorem 2 with $S = S_{\widehat{k}}$ and $a_{N_1,N_2} = e_{N_1,N_2} + N_1^{-1/2} + N_2^{-1/2}$. □

D Missing Proofs and Further Discussions for Section 3.3

This appendix proves Theorem 4. Appendix D.1 bounds $\overline{d}_k$ for a general set I. Appendix D.2 uses that result to prove case (a), and Appendix D.3 proves case (b). The latter two appendices also give examples satisfying their respective assumptions.

D.1 A General Bound for $\overline{d}_k$

Fix a nonempty comparison set $I \subseteq [n]$, put $M := |I|$, and define

$$\bar{x}_I := \frac{1}{M} \sum_{i \in I} x_i, \qquad \mathrm{diam}_I := \max_{i,r \in I} \|x_i - x_r\|_{\mathcal{H}}, \qquad B_u := \max_{i \in [n]} \|x_i - u\|_{\mathcal{H}}.$$

Let $\mathcal{L} \subseteq \mathcal{H}$ be a closed subspace and let $\beta \geq 0$ satisfy

$$x_i - u \in \mathcal{L} \quad (i \in I), \qquad \|\Pi_{\mathcal{L}}(x_i - u)\|_{\mathcal{H}} \leq \beta \quad (i \notin I). \tag{38}$$

For $H_m := \sum_{r=1}^{m} r^{-1}$ and $H_0 := 0$, define

$$\Lambda_{M,k} := H_M - H_{(M-k)_+}, \tag{39}$$

and

$$G_{I,k} := M\|\bar{x}_I - u\|_{\mathcal{H}} + (k-1)\beta.$$

Theorem 8. *Consider the greedy sequence of Section 2.2. Under* (38)*, the following bound holds simultaneously for every* $2 \le k \le n$*:*

$$\overline{d}_k \le \frac{1}{k}\left\{2G_{I,k}\Lambda_{M,k} + \operatorname{diam}_I \sqrt{\frac{\min\{M,k\}\Lambda_{M,k}}{2}} + B_u(k-M)_+\right\}. \tag{40}$$

The bound remains valid at a possibly data-dependent index because it holds simultaneously along the realized sequence.

For $k \le M$, one has $\min\{M,k\} = k$, the final term vanishes, and $\Lambda_{M,k} \le k/(M-k+1)$. Thus

$$\overline{d}_k \le \frac{2\{M\|\bar{x}_I - u\|_{\mathcal{H}} + (k-1)\beta\}}{M-k+1} + \frac{\operatorname{diam}_I}{\sqrt{2(M-k+1)}}.$$

In particular, if $k \le (1-\eta)M$ for fixed $\eta \in (0,1)$, then

$$\overline{d}_k \le \frac{2}{\eta}\big(\|\bar{x}_I - u\|_{\mathcal{H}} + \beta\big) + \frac{\operatorname{diam}_I}{\sqrt{2\eta M}}.$$

At $k = M$, $\Lambda_{M,M} = H_M \le 1 + \log M$; the theorem does not assert that this logarithmic factor is necessary.

Proof of Theorem 8. Let

$$A_j := \sum_{i \in S_j}(x_i - u) = j(v_j - u), \qquad A_0 := 0.$$

Decompose $A_j = A_{j,\mathcal{L}} + A_{j,\perp}$ according to $\mathcal{H} = \mathcal{L} \oplus \mathcal{L}^{\perp}$. Let

$$I_j := I \setminus S_j, \qquad s_j := |I_j|, \qquad O_j := S_j \setminus I.$$

For $j < M$, the set I_j contains at least $M - j$ points. Hence $s_j \ge M - j \ge 1$. Define

$$Z_j := M(\bar{x}_I - u) + \sum_{i \in O_j} \Pi_{\mathcal{L}}(x_i - u).$$

Condition (38) implies, for every $j < k$,

$$\|Z_j\|_{\mathcal{H}} \le M\|\bar{x}_I - u\|_{\mathcal{H}} + j\beta \le G_{I,k}.$$

Let $y_j := s_j^{-1}\sum_{i \in I_j} x_i$. Subtracting the terms indexed by $I \cap S_j$ from the sum over I gives

$$s_j(y_j - u) = Z_j - A_{j,\mathcal{L}}. \tag{41}$$

In particular, $y_j \in \mathcal{C}_j$, and therefore

$$d_j \le \|y_j - u\|_{\mathcal{H}} \le \frac{G_{I,k} + \|A_{j,\mathcal{L}}\|_{\mathcal{H}}}{s_j}. \tag{42}$$

Since a minimum is no greater than the average over I_j,

$$\|A_{j+1}\|_{\mathcal{H}}^2 = \min_{i \in R_j}\|A_j + (x_i - u)\|_{\mathcal{H}}^2 \le \frac{1}{s_j}\sum_{i \in I_j}\|A_j + (x_i - u)\|_{\mathcal{H}}^2 = \|A_j + (y_j - u)\|_{\mathcal{H}}^2 + V_j,$$

where

$$V_j := \frac{1}{s_j}\sum_{i \in I_j}\|x_i - y_j\|_{\mathcal{H}}^2 = \frac{1}{2s_j^2}\sum_{i,r \in I_j}\|x_i - x_r\|_{\mathcal{H}}^2 \le \frac{\operatorname{diam}_I^2}{2}.$$

By (41), the orthogonal decomposition of A_j and convexity of $\|\cdot\|_{\mathcal{H}}^2$,

$$\begin{aligned}\|A_j + (y_j - u)\|_{\mathcal{H}}^2 &= \|A_{j,\perp}\|_{\mathcal{H}}^2 + \left\|\left(1 - \frac{1}{s_j}\right)A_{j,\mathcal{L}} + \frac{1}{s_j}Z_j\right\|_{\mathcal{H}}^2 \\ &\le \|A_j\|_{\mathcal{H}}^2 - \frac{\|A_{j,\mathcal{L}}\|_{\mathcal{H}}^2}{s_j} + \frac{\|Z_j\|_{\mathcal{H}}^2}{s_j}.\end{aligned}$$

Thus, for $j < M$ and $j < k$,

$$\|A_{j+1}\|_{\mathcal{H}}^2 \le \|A_j\|_{\mathcal{H}}^2 - \frac{\|A_{j,\mathcal{L}}\|_{\mathcal{H}}^2}{s_j} + \frac{\|Z_j\|_{\mathcal{H}}^2}{s_j} + \frac{\operatorname{diam}_I^2}{2}. \tag{43}$$

Put $J := \min\{k-1, M-1\}$. Since $s_j \ge M - j$,

$$\sum_{j=0}^{J} \frac{1}{s_j} \le \sum_{j=0}^{J} \frac{1}{M-j} = \Lambda_{M,k}. \tag{44}$$

Summing (43) from 0 through J and discarding the final nonnegative squared norm gives

$$\sum_{j=0}^{J} \frac{\|A_{j,\mathcal{L}}\|_{\mathcal{H}}^2}{s_j} \le \sum_{j=0}^{J} \frac{\|Z_j\|_{\mathcal{H}}^2}{s_j} + \frac{\min\{M,k\}\operatorname{diam}_I^2}{2} \le G_{I,k}^2 \Lambda_{M,k} + \frac{\min\{M,k\}\operatorname{diam}_I^2}{2}. \tag{45}$$

Apply (42) for $1 \le j \le J$ and use Cauchy–Schwarz with (44)–(45):

$$\begin{aligned}
\sum_{j=1}^{J} d_j &\le \sum_{j=1}^{J} \frac{\|Z_j\|_{\mathcal{H}}}{s_j} + \left(\sum_{j=1}^{J} \frac{\|A_{j,\mathcal{L}}\|_{\mathcal{H}}^2}{s_j}\right)^{1/2} \left(\sum_{j=1}^{J} \frac{1}{s_j}\right)^{1/2} \\
&\le \sum_{j=1}^{J} \frac{\|Z_j\|_{\mathcal{H}}}{s_j} + \sqrt{\Lambda_{M,k}\left(\sum_{j=0}^{J} \frac{\|Z_j\|_{\mathcal{H}}^2}{s_j} + \frac{1}{2}\min\{M,k\}\operatorname{diam}_I^2\right)} \\
&\le G_{I,k}\Lambda_{M,k} + \sqrt{\Lambda_{M,k}\left(G_{I,k}^2\Lambda_{M,k} + \frac{1}{2}\min\{M,k\}\operatorname{diam}_I^2\right)}.
\end{aligned}$$

Apply $\sqrt{x^2+y^2} \le x + y$ with $x = G_{I,k}\Lambda_{M,k}$ and $y = \operatorname{diam}_I \sqrt{\min\{M,k\}\Lambda_{M,k}/2}$ to obtain

$$\sum_{j=1}^{J} d_j \le 2G_{I,k}\Lambda_{M,k} + \operatorname{diam}_I \sqrt{\frac{\min\{M,k\}\Lambda_{M,k}}{2}}.$$

If $k > M$, then for each $j = M, \dots, k-1$, any $i \in R_j$ gives $d_j \le \|u - x_i\|_{\mathcal{H}} \le B_u$. Adding these terms and dividing by k proves (40). □

D.2 Approximate Orthogonality

In this and the following subsection, we return to the energy-kernel setting. We use $\mathcal{H}$, $\mathcal{L}_P$, and diam_P as defined in Section 3.3. Also let

$$\operatorname{diam}_{P\cup Q} := \operatorname{diam}_{\mathcal{H}}(\Phi_{\mathrm{E}}(\operatorname{supp}(P) \cup \operatorname{supp}(Q))).$$

Under Assumption 1, both diam_P and $\operatorname{diam}_{P\cup Q}$ are finite. We use I_P, M and $\bar{\mu}_{P,M}$ as defined in the proof of Lemma 7. Also define

$$G_{P,k} := M\|\bar{\mu}_{P,M} - u\|_{\mathcal{H}} + (k-1)\big(\|u - \mu_P\|_{\mathcal{H}} + \zeta\big).$$

Corollary 9. *Under the mixture sampling model of Section 2.1, suppose Assumptions 1 and 2 hold. Almost surely on the event $\{M \ge 1\}$, the exact greedy sequence* (7) *satisfies, simultaneously for every $2 \le k \le N_2$,*

$$\bar{d}_k \le \frac{1}{k}\left\{2G_{P,k}\Lambda_{M,k} + \operatorname{diam}_P \sqrt{\frac{\min\{M,k\}\Lambda_{M,k}}{2}} + \operatorname{diam}_{P\cup Q}(k-M)_+\right\}. \tag{46}$$

Hence the bound remains valid at any data-dependent $\widehat{k} \in \{2, \dots, N_2\}$.

Proof. Apply Theorem 8 with $I = I_P$, subspace $\mathcal{L} = \mathcal{L}_P$, and $\beta = \|u - \mu_P\|_{\mathcal{H}} + \zeta$. For $i \in I_P$, both $x_i - \mu_P$ and $u - \mu_P$ belong to $\mathcal{L}_P$, so $x_i - u \in \mathcal{L}_P$. For every candidate generated from Q, almost surely,

$$\|\Pi_{\mathcal{L}_P}(x_i - u)\|_{\mathcal{H}} \le \|\Pi_{\mathcal{L}_P}(x_i - \mu_P)\|_{\mathcal{H}} + \|u - \mu_P\|_{\mathcal{H}} \le \zeta + \|u - \mu_P\|_{\mathcal{H}} = \beta.$$

Moreover, $\bar{x}_I = \bar{\mu}_{P,M}$, $G_{I,k} = G_{P,k}$, $\operatorname{diam}_I \le \operatorname{diam}_P$, and $B_u \le \operatorname{diam}_{P\cup Q}$ because $u \in \operatorname{conv}\{\Phi_{\mathrm{E}}(x) : x \in \operatorname{supp}(P)\}$. The result follows. □

Proof of case (a) of Theorem 4. Since $\mathbb{P}(M = 0) = (1-\theta)^{N_2} \to 0$, the event on which $\bar{\mu}_{P,M}$ is undefined is asymptotically negligible. The fact that M is a binomial variable and the boundedness of $\Phi_{\mathrm{E}}(z_j)$ give

$$M = \theta N_2 + O_{\mathbb{P}}(\sqrt{N_2}), \qquad \|\bar{\mu}_{P,M} - \mu_P\|_{\mathcal{H}} = O_{\mathbb{P}}(N_2^{-1/2}), \qquad \|u - \mu_P\|_{\mathcal{H}} = O_{\mathbb{P}}(N_1^{-1/2}).$$

Let

$$\gamma := N_1^{-1/2} + N_2^{-1/2} + \zeta, \qquad \rho_{N_1,N_2} = \omega(\gamma).$$

Then

$$\eta := \|\bar{\mu}_{P,M} - u\|_{\mathcal{H}} + \|u - \mu_P\|_{\mathcal{H}} + \zeta = O_{\mathbb{P}}(\gamma).$$

It is enough to consider $\gamma \le 1/2$. Otherwise, $\mathrm{diam}_{P\cup Q} < \infty$ implies that every term on the left-hand side of (22) is $O_{\mathbb{P}}(1)$. This is also $O_{\mathbb{P}}(\omega(\gamma))$ because $\inf_{t>1/2}\omega(t) = \omega(1/2) > 0$. Let $\ell_\gamma = \lceil \gamma M \rceil$ and

$$q := \min\{\widehat{k}, M - \ell_\gamma\}.$$

With probability tending to one, $2 \le q \le M - \ell_\gamma$. Note that

$$\Lambda_{M,q} \le H_M - H_{\ell_\gamma} \le \log(M/\ell_\gamma) \le \log(1/\gamma), \qquad G_{P,q} \le M\eta.$$

Evaluating (46) at $k = q$ and using $q \le M$ therefore gives

$$\sum_{j=1}^{q-1} d_j \le 2M\eta\log(1/\gamma) + \mathrm{diam}_P\sqrt{\frac{M\log(1/\gamma)}{2}}. \tag{47}$$

For $q \le j < \widehat{k}$, $d_j \le \mathrm{diam}_{P\cup Q}$, and

$$0 \le \widehat{k} - q \le |\widehat{k} - M| + \ell_\gamma.$$

Now, $M = \theta N_2 + O_{\mathbb{P}}(\sqrt{N_2})$, $\widehat{k} = \theta N_2 + O_{\mathbb{P}}(N_2 r_{N_1,N_2})$, and $\widehat{k} \asymp N_2$. Combining this with (47) and dividing by $\widehat{k}$ yields

$$\overline{d}_{\widehat{k}} = O_{\mathbb{P}}(r_{N_1,N_2} + \rho_{N_1,N_2})\,.$$

Here $\sqrt{\log(1/\gamma)/N_2}$ is absorbed by $\omega(\gamma)$ because $\gamma \ge N_2^{-1/2}$. Since $\overline{\epsilon}_{\widehat{k}} \le \overline{d}_{\widehat{k}}$, the preceding bound gives

$$\overline{\epsilon}_{\widehat{k}} = O_{\mathbb{P}}(r_{N_1,N_2} + \rho_{N_1,N_2})\,.$$

Moreover, $N_1^{-1/2} + N_2^{-1/2} \le \gamma \le \rho_{N_1,N_2}$. Corollary 3 therefore gives

$$\mathrm{err}(S_{\widehat{k}}) + \|\widehat{\mu}(S_{\widehat{k}}) - \mu_P\|_{\mathcal{H}} + \|\widehat{\mu}(R_{\widehat{k}}) - \mu_Q\|_{\mathcal{H}} = O_{\mathbb{P}}(r_{N_1,N_2} + \rho_{N_1,N_2})\,.$$

Together with the bound for $\overline{d}_{\widehat{k}}$, this proves (22). □

Continuous energy-kernel example. The case $\zeta = 0$ is restrictive. If $\mathcal{L}_P = \mathcal{H}$, it forces $\Phi_{\mathrm{E}}(z) = \mu_P$ for Q-almost every z. The following example shows that it nevertheless permits both components to have no point masses and uncountable support when $\mathcal{L}_P^{\perp} \neq \{0\}$.

Let $\mathbb{R}^d = V \oplus W$ be an orthogonal decomposition with $\dim(V), \dim(W) \ge 2$, and let $a, b > 0$. For the energy kernel, take

$$P = \mathrm{Unif}\{x \in V : \|x\|_2 = a\}, \qquad Q = \mathrm{Unif}\{z \in W : \|z\|_2 = b\}.$$

Then P and Q have no point masses, have uncountable supports and distinct mean embeddings, and satisfy (20) with $\zeta = 0$. Consequently, under the mixture sampling model of Section 2.1, with fixed $\theta \in (0,1)$ and $N_1 \asymp N_2 \to \infty$, the exact energy-greedy sequence stopped at $k_\theta = \mathrm{round}(\theta N_2)$ satisfies

$$\overline{d}_{k_\theta} + \mathrm{err}(S_{k_\theta}) + \|\widehat{\mu}(S_{k_\theta}) - \mu_P\|_{\mathcal{H}} + \|\widehat{\mu}(R_{k_\theta}) - \mu_Q\|_{\mathcal{H}} = O_{\mathbb{P}}\left(\frac{\log N_2}{\sqrt{N_2}}\right).$$

To verify these claims, fix $z \in W$. For every x on the radius-a sphere in V, orthogonality gives

$$\|z - x\|_2 = \sqrt{\|z\|_2^2 + a^2}.$$

Consequently $k_{\mathrm{E}}(z, x)$ is constant in x, and hence

$$\langle \Phi_{\mathrm{E}}(z), \Phi_{\mathrm{E}}(x) - \mu_P \rangle_{\mathcal{H}} = k_{\mathrm{E}}(z, x) - \mathop{\mathbb{E}}_{X \sim P} k_{\mathrm{E}}(z, X) = 0. \tag{48}$$

Rotational invariance of P makes $x \mapsto \mathbb{E}_{X \sim P}\, k_{\mathrm{E}}(X, x)$ constant on the target sphere. Its constant value is its P-average, $\|\mu_P\|_{\mathcal{H}}^2$, so

$$\langle \mu_P, \Phi_{\mathrm{E}}(x) - \mu_P \rangle_{\mathcal{H}} = 0. \tag{49}$$

Subtracting (49) from (48) shows that $\Phi_{\mathrm{E}}(z) - \mu_P$ is orthogonal to every generator of $\mathcal{L}_P$, and therefore belongs to $\mathcal{L}_P^{\perp}$.

Both spherical surface laws have no point masses when their subspaces have dimension at least two. Their supports lie in orthogonal subspaces, so the laws are distinct. Since $\sqrt{\mathcal{E}(\cdot,\cdot)}$ is a metric on Euclidean laws with finite first moment (Székely & Rizzo, 2013), $\mathcal{E}(P, Q) > 0$. Thus (3) gives $\mu_P \neq \mu_Q$. The feature-distance identity also gives

$$\mathrm{diam}_P^2 = 2a, \qquad \mathrm{diam}_{P \cup Q}^2 = 2 \max\{a, b\}.$$

The cross-support squared feature distance is $\sqrt{a^2 + b^2} \leq a + b \leq 2 \max\{a, b\}$, while the larger sphere contains an antipodal pair attaining $2 \max\{a, b\}$. Thus the displayed joint diameter is exact. For known θ, the rate follows from case (a) of Theorem 4 and $|\alpha_{k_\theta} - \theta| \leq 1/(2N_2)$.

Perturbing the Q-component. The approximate-orthogonality condition is stable under small perturbations of Q when P is fixed. Let Q_0 satisfy Assumption 2 with $\zeta = 0$, and suppose there is a coupling of $Z \sim Q$ and $Z_0 \sim Q_0$ such that

$$\|\Phi_{\mathrm{E}}(Z) - \Phi_{\mathrm{E}}(Z_0)\|_{\mathcal{H}} \leq \delta \qquad \text{almost surely.}$$

Because $\Phi_{\mathrm{E}}(Z_0) - \mu_P$ is orthogonal to $\mathcal{L}_P$ almost surely, projection contraction gives

$$\|\Pi_{\mathcal{L}_P}\{\Phi_{\mathrm{E}}(Z) - \mu_P\}\|_{\mathcal{H}} \leq \|\Pi_{\mathcal{L}_P}\{\Phi_{\mathrm{E}}(Z) - \Phi_{\mathrm{E}}(Z_0)\}\|_{\mathcal{H}} \tag{50}$$

$$\leq \|\Phi_{\mathrm{E}}(Z) - \Phi_{\mathrm{E}}(Z_0)\|_{\mathcal{H}} \leq \delta \tag{51}$$

almost surely. Thus Q satisfies Assumption 2 with $\zeta = \delta$. For the energy feature map, $\|\Phi_{\mathrm{E}}(Z) - \Phi_{\mathrm{E}}(Z_0)\|_{\mathcal{H}}^2 = \|Z - Z_0\|_2$. In particular, $\|Z - Z_0\|_2 \leq \eta$ almost surely implies approximate orthogonality with $\zeta = \sqrt{\eta}$. This argument applies only when P is fixed. If P is also perturbed, both μ_P and $\mathcal{L}_P$ change, so a separate stability argument would be needed.

D.3 Finite Support with Pairing

We need an auxiliary lemma.

Lemma 10 (Affine Independence)**.** *If $\xi_1, \ldots, \xi_m$ are distinct, then $\Phi_{\mathrm{E}}(\xi_1), \ldots, \Phi_{\mathrm{E}}(\xi_m)$ are affinely independent.*

Proof. Suppose $\sum_s c_s = 0$ and $\sum_s c_s \Phi_{\mathrm{E}}(\xi_s) = 0$. If $c = (c_1, \ldots, c_m) \neq 0$, then $r := \sum_s (c_s)_+ = \sum_s (-c_s)_+ > 0$, and

$$P_+ := \frac{1}{r} \sum_s (c_s)_+ \delta_{\xi_s}, \qquad P_- := \frac{1}{r} \sum_s (-c_s)_+ \delta_{\xi_s}$$

are probability measures, and

$$\mu_{P_+} - \mu_{P_-} = \frac{1}{r} \sum_s c_s \Phi_{\mathrm{E}}(\xi_s) = 0.$$

Thus (3) gives $\mathcal{E}(P_+, P_-) = 0$. Since $\sqrt{\mathcal{E}(\cdot,\cdot)}$ is a metric, $P_+ = P_-$. Their supports are disjoint, however, so this is impossible. Therefore $c = 0$. □

Under Assumption 3, suppose that $\mathrm{supp}(P) = \{\xi_1, \ldots, \xi_m\}$ and let $p_s := P(\{\xi_s\})$ and $a_s := \Phi_{\mathrm{E}}(\xi_s)$. Thus $p_{\min} = \min_s p_s$. By Lemma 10, the vectors $a_1, \ldots, a_m$ are affinely independent. Define

$$g_s := \mathrm{dist}_{\mathcal{H}}(a_s, \mathrm{aff}\{a_t : t \neq s\}), \qquad g_P = \min_s g_s > 0.$$

Use the map π, displacement b_z, and constant R_P from Assumption 3. For the proof, write

$$\Delta_\pi(z) := \|b_z\|_{\mathcal{H}}^2 - 2R_P\|\Pi_{\mathcal{L}_P} b_z\|_{\mathcal{H}}.$$

Thus the condition (21) becomes "$b_z = 0$ or $\Delta_\pi(z) > 0$". A useful sufficient condition is that the same assignment satisfies

$$\Phi_{\mathrm{E}}(z) = \Phi_{\mathrm{E}}(\pi(z)) + w_z, \qquad w_z \in \mathcal{L}_P^{\perp}. \tag{52}$$

Under (52), $b_z = w_z$ and $\Pi_{\mathcal{L}_P} b_z = 0$. Hence $\Delta_\pi(z) = \|w_z\|_{\mathcal{H}}^2$, which is positive whenever $b_z \neq 0$.

Let

$$B_* := \frac{\mathrm{diam}_P^2}{p_{\min} g_P}, \qquad K_* := \frac{B_*}{g_P}.$$

Then $R_P = \mathrm{diam}_P + B_*$. For $\delta \in (0,1)$, define

$$\tau_1(\delta) := \sqrt{\frac{\log(4m/\delta)}{2N_1}}, \qquad \tau_2(\delta) := \sqrt{\frac{\log(4m/\delta)}{2N_2}}.$$

Lemma 11. *Under the mixture sampling model of Section 2.1, suppose Assumptions 1 and 3 hold and the exact greedy rule* (7) *is used. If $\tau_1(\delta) \le p_{\min}/2$, then, with probability at least $1-\delta$, simultaneously for every $2 \le k < N_2$ satisfying $\eta_k := |\alpha_k - \theta| \le \theta/2$,*

$$\overline{d}_k \le \frac{4\,\mathrm{diam}_{P\cup Q}}{\theta p_{\min}}\left(\eta_k + \tau_1(\delta) + \tau_2(\delta)\right) + \frac{4\,\mathrm{diam}_{P\cup Q}}{\theta N_2}\left(1 + \frac{K_*}{p_{\min}}\right). \tag{53}$$

Proof. For $s \in [m]$, let

$$\widehat{p}_s := \frac{1}{N_1}\sum_{i=1}^{N_1} \mathbf{1}_{\{y_i = \xi_s\}}, \qquad M_s := |\{j : z_j \text{ was generated from } P \text{ and equals } \xi_s\}|.$$

Then $u = \sum_s \widehat{p}_s a_s$ and $\mathbb{E}\, M_s = N_2 \theta p_s$. Here M_s counts only candidates generated from P, whereas selected features equal to a_s below may also have been generated from Q. Hoeffding's inequality and a union bound give, with probability at least $1-\delta$,

$$\max_s |\widehat{p}_s - p_s| \le \tau_1(\delta), \qquad \max_s \left|\frac{M_s}{N_2} - \theta p_s\right| \le \tau_2(\delta).$$

We also condition on the event of probability 1 that (21) holds for every sampled Q-candidate. The assumption $\tau_1(\delta) \le p_{\min}/2$ implies that $\widehat{p}_s \ge p_{\min}/2$ for every s.

Suppose that the first j selected features belong to $\{a_1, \dots, a_m\}$, and let $C_s(j)$ count those equal to a_s. Define

$$h_j := \sum_{s=1}^{m} \big(C_s(j) - j\widehat{p}_s\big) a_s, \qquad t_j := (j+1)u - \sum_{s=1}^{m} C_s(j) a_s = u - h_j.$$

The next greedy choice minimizes its distance from t_j.

Let $\rho_{\widehat{p}} := \min_s \widehat{p}_s g_s$. The simplex $\mathrm{conv}\{a_s - u : s \in [m]\}$ contains the closed $\mathcal{L}_P$-ball of radius $\rho_{\widehat{p}}$ centered at zero. Consequently, if a_s minimizes $\|t_j - a_s\|_{\mathcal{H}}$ among the support points, then selecting it gives

$$\|h_{j+1}\|_{\mathcal{H}}^2 \le \|h_j\|_{\mathcal{H}}^2 - 2\rho_{\widehat{p}}\|h_j\|_{\mathcal{H}} + \mathrm{diam}_P^2\,.$$

Because $\widehat{p}_s \ge p_{\min}/2$ and $g_s \ge g_P$ for every s, $\rho_{\widehat{p}} \ge p_{\min} g_P/2$. Moreover, $g_P \le \mathrm{diam}_P$ and $p_{\min} \le 1$, so $B_* \ge \mathrm{diam}_P$, while $2\rho_{\widehat{p}} B_* \ge \mathrm{diam}_P^2$. Convexity of the right-hand side as a function of $\|h_j\|_{\mathcal{H}}$ shows that these updates preserve $\|h_j\|_{\mathcal{H}} \le B_*$, starting from $h_0 = 0$.

For each s, let $n_s \in \mathcal{L}_P$ be the unit normal to the simplex facet opposite a_s, oriented so that $\langle a_s - a_t, n_s\rangle_{\mathcal{H}} = g_s$ for $t \neq s$. Since $\sum_s (C_s(j) - j\widehat{p}_s) = 0$,

$$\langle h_j, n_s\rangle_{\mathcal{H}} = g_s\big(C_s(j) - j\widehat{p}_s\big), \qquad |C_s(j) - j\widehat{p}_s| \le \frac{B_*}{g_s} \le K_*.$$

The same bound gives

$$\|t_j - a_s\|_{\mathcal{H}} \le B_* + \operatorname{diam}_P = R_P.$$

For a Q-generated candidate z, since $t_j - \Phi_{\mathrm{E}}(\pi(z)) \in \mathcal{L}_P$,

$$\begin{aligned}\|t_j - \Phi_{\mathrm{E}}(z)\|_{\mathcal{H}}^2 - \|t_j - \Phi_{\mathrm{E}}(\pi(z))\|_{\mathcal{H}}^2 &= \|b_z\|_{\mathcal{H}}^2 - 2\langle t_j - \Phi_{\mathrm{E}}(\pi(z)), \Pi_{\mathcal{L}_P} b_z\rangle_{\mathcal{H}} \\ &\ge \|b_z\|_{\mathcal{H}}^2 - 2R_P\|\Pi_{\mathcal{L}_P} b_z\|_{\mathcal{H}} \\ &= \Delta_\pi(z) \\ &\ge 0,\end{aligned}$$

with strict inequality unless $b_z = 0$. Thus, as long as one P-generated copy of every a_s remains, the greedy rule selects a feature in $\{a_1, \dots, a_m\}$. Starting from $h_0 = 0$, the preceding argument shows inductively that

$$\|h_j\|_{\mathcal{H}} \le B_*, \qquad |C_s(j) - j\widehat{p}_s| \le K_*.$$

Fix k and define

$$L_k := 2 + \max_{1 \le s \le m} \frac{(k\widehat{p}_s + K_* - M_s)_+}{\widehat{p}_s}.$$

It holds for every integer $0 \le j \le \lfloor k - L_k \rfloor$ and every s that

$$j\widehat{p}_s + K_* < M_s.$$

If this range is nonempty, then at $j = 0$ the preceding bounds hold, and the displayed inequality ensures that a P-generated copy of every a_s remains. If these properties hold after j selections, the preceding comparison selects another support feature and preserves the bounds. If $j + 1$ remains in the range, then $C_s(j+1) \le (j+1)\widehat{p}_s + K_* < M_s$, so a P-generated copy of every a_s still remains. Thus induction proves that the preceding bounds apply throughout this range. Since $u = \sum_s \widehat{p}_s a_s$, it follows that

$$u \in \operatorname{conv}\{x_i : i \in R_j\}, \qquad d_j = 0, \qquad 1 \le j \le \lfloor k - L_k \rfloor. \tag{54}$$

For every other $j < k$, choose any remaining candidate x_i. Then

$$d_j \le \|u - x_i\|_{\mathcal{H}} \le \sum_{s=1}^{m} \widehat{p}_s \|a_s - x_i\|_{\mathcal{H}} \le \operatorname{diam}_{P \cup Q}.$$

Thus fewer than L_k terms in $\overline{d}_k$ can be nonzero, including when $L_k \ge k$, and

$$\overline{d}_k \le \frac{\operatorname{diam}_{P\cup Q} L_k}{k}. \tag{55}$$

Finally, the concentration bounds imply

$$\begin{aligned}k\widehat{p}_s - M_s &= N_2(\alpha_k - \theta)\widehat{p}_s + N_2\theta(\widehat{p}_s - p_s) + N_2\left(\theta p_s - \frac{M_s}{N_2}\right) \\ &\le N_2\left(\eta_k + \tau_1(\delta) + \tau_2(\delta)\right).\end{aligned}$$

Because $\widehat{p}_s \ge p_{\min}/2$,

$$L_k \le 2 + \frac{2}{p_{\min}}\left(N_2(\eta_k + \tau_1(\delta) + \tau_2(\delta)) + K_*\right). \tag{56}$$

If $\eta_k \le \theta/2$, then $k \ge \theta N_2/2$. Combining (55) and (56), using $k \ge \theta N_2/2$, proves (53). The concentration event and the realized greedy sequence are common to every k, so the bound is simultaneous and remains valid at any data-dependent $\widehat{k}$ satisfying the same conditions. □

Proof of case (b) of Theorem 4. The stopping assumption gives $\widehat{k} \asymp N_2$, and Lemma 11 implies

$$\overline{d}_{\widehat{k}} = O_{\mathbb{P}}\left(r_{N_1,N_2} + N_1^{-1/2} + N_2^{-1/2}\right).$$

Moreover, $C_{\mathrm{pool}} \le \mathrm{diam}^2_{P\cup Q}$ and $\mathrm{err}_1^2 \le \mathrm{diam}^2_{P\cup Q}$. Therefore Theorem 1 gives

$$\mathrm{err}(S_{\widehat{k}}) \le \overline{d}_{\widehat{k}} + \frac{\sqrt{\mathrm{err}_1^2 + \sum_{j<\widehat{k}} \kappa_j}}{\widehat{k}} \le \overline{d}_{\widehat{k}} + \frac{\mathrm{diam}_{P\cup Q}}{\sqrt{\widehat{k}}},$$

so $\mathrm{err}(S_{\widehat{k}})$ has the same stochastic order. Finally apply Corollary 3, using $\overline{\epsilon}_{\widehat{k}} \le \overline{d}_{\widehat{k}}$, to obtain the selected- and remainder-set bounds in (22). For $k_\theta = \mathrm{round}(\theta N_2)$, $|\alpha_{k_\theta} - \theta| \le 1/(2N_2)$, so the conclusion for known θ follows by taking $r_{N_1,N_2} = N_2^{-1}$. □

Unlike Theorem 8, this proof does not sum $1/s_j$. Equation (54) gives $d_j = 0$ for $j \le \lfloor k - L_k \rfloor$, so fewer than L_k terms enter $\overline{d}_k$ in (55). This is why the resulting rate has no logarithmic factor. Finite support alone does not imply (54).

Four-point example. Let

$$\mathcal{X} = \{-b, -a, a, b\} \subset \mathbb{R}, \qquad 0 < a < b,$$

and

$$P = \lambda\delta_{-a} + (1-\lambda)\delta_a, \qquad Q = \chi\delta_{-b} + (1-\chi)\delta_b,$$

where $\lambda \in (0,1)$ and $\chi \in [0,1]$. For the energy kernel, $\Phi_{\mathrm{E}}(-a)$ and $\Phi_{\mathrm{E}}(a)$ are affinely independent, and Assumption 3 holds. In fact, the displacements satisfy (52). Consequently, case (b) of Theorem 4 applies under its sampling and asymptotic conditions. Any $\widehat{k}$ satisfying $|\alpha_{\widehat{k}} - \theta| = O_{\mathbb{P}}(r_{N_1,N_2})$ for a deterministic $r_{N_1,N_2} > 0$ with $r_{N_1,N_2} \to 0$ therefore satisfies (22) with $\rho_{N_1,N_2} = N_1^{-1/2} + N_2^{-1/2}$.

To verify the structural condition, recall that, for $x, y \in \mathbb{R}$, the energy kernel is

$$k_{\mathrm{E}}(x,y) = \frac{1}{2}\big(|x| + |y| - |x-y|\big).$$

Put $a_- := \Phi_{\mathrm{E}}(-a)$ and $a_+ := \Phi_{\mathrm{E}}(a)$. Direct evaluation gives

$$\langle a_-, a_+\rangle_{\mathcal{H}} = 0, \qquad \|a_-\|^2_{\mathcal{H}} = \|a_+\|^2_{\mathcal{H}} = a.$$

Define

$$w_- := \Phi_{\mathrm{E}}(-b) - a_-, \qquad w_+ := \Phi_{\mathrm{E}}(b) - a_+.$$

Take $\pi(-b) = -a$ and $\pi(b) = a$. Same-sign kernel values equal the smaller absolute coordinate, whereas opposite-sign values vanish. Hence

$$w_-, w_+ \perp \mathrm{span}\{a_-, a_+\}, \qquad \|w_-\|^2_{\mathcal{H}} = \|w_+\|^2_{\mathcal{H}} = b - a.$$

Because $a_- \ne a_+$, they are affinely independent. Since $\mathcal{L}_P \subseteq \mathrm{span}\{a_-, a_+\}$, the displayed orthogonality implies (52) and hence Assumption 3. Finally, $\|\Phi_{\mathrm{E}}(x) - \Phi_{\mathrm{E}}(x')\|^2_{\mathcal{H}} = |x - x'|$, so $\mathrm{diam}_{P\cup Q} \le \sqrt{2b}$, regardless of whether a Q endpoint has zero probability. The remaining hypotheses listed above are those of case (b) of Theorem 4, which gives the stated conclusion.

E Sharpness Results

E.1 Sharpness of the Main Bound

The upper bounds in (11) may be loose for a particular greedy ordering. The two examples below show, however, that neither the unit coefficient of $\overline{\epsilon}_k$ nor the $k^{-1/2}$ order of the square-root term can be improved uniformly.

Proposition 12. *Let $\mathcal{H} = \mathbb{R}$, $u = 0$, $x_1 = 0$, and $x_2 = \cdots = x_n = 1$, where $n \ge 3$. Then the greedy subset is optimal and, for every $1 \le k < n$,*

$$\mathrm{err}_k = \frac{k-1}{k}. \tag{57}$$

Moreover, $d_0 = 0$, while $d_j = \epsilon_j = 1$ and $\mathrm{diam}^2_{\mathcal{H}}(\mathcal{C}_j) = \kappa_j = 0$ for $1 \le j < n$. Hence $\overline{\epsilon}_k = (k-1)/k = \mathrm{err}_k$: the unit coefficient of $\overline{\epsilon}_k$ cannot be reduced even when every stagewise diameter term vanishes.

Proof. The first greedy choice is $x_1 = 0$ and every subsequent choice has value one. Hence

$$v_k = \frac{k-1}{k}, \qquad \mathrm{err}_k = \frac{k-1}{k}.$$

For $k < n$, a k-subset either contains x_1 and has average $(k-1)/k$, or excludes it and has average one. Thus the greedy subset is also optimal. Because $u = x_1 \in \mathcal{C}_0$, $d_0 = 0$. After the first choice, all remaining points equal one. Hence $\mathcal{C}_j = \{1\}$, $d_j = \epsilon_j = 1$, and $\mathrm{diam}^2_{\mathcal{H}}(\mathcal{C}_j) = \kappa_j = 0$ for $1 \le j < n$. These identities give the remaining claims. □

Because $\mathrm{err}_1 = 0$, this construction attains equality in the first upper bound of (11); it does not assert equality in the subsequent global relaxation.

Proposition 13. *Let $m \ge 2$, and let $y_1, \dots, y_m$ be the vertices of a regular simplex centered at $u = 0$ with $\|y_a - y_b\|^2_{\mathcal{H}} = C_{\mathrm{pool}}$ for $a \ne b$. Suppose that the candidate pool contains two indexed copies of every vertex. During its first m steps, the greedy ordering selects one copy of a previously unselected vertex. For $1 \le k \le m$,*

$$d_j = 0 \quad (1 \le j \le k), \qquad \mathrm{err}_k^2 = \frac{C_{\mathrm{pool}}(m-k)}{2mk}. \tag{58}$$

For $1 \le j < m$, $\mathrm{diam}^2_{\mathcal{H}}(\mathcal{C}_j) = C_{\mathrm{pool}}$ and $\kappa_j = \frac{1}{2}C_{\mathrm{pool}}\{1 - (2m-j)^{-1}\}$. Consequently,

$$\frac{\mathrm{err}_k^2}{C_{\mathrm{pool}}/(2k)} = 1 - \frac{k}{m}, \qquad \frac{\mathrm{err}_k^2}{\{\mathrm{err}_1^2 + \sum_{j=1}^{k-1}\kappa_j\}/k^2} = \frac{k(m-k)}{km - 1 - m\sum_{j=1}^{k-1}(2m-j)^{-1}}.$$

Both ratios tend to one along any sequence for which $k \to \infty$ and $k/m \to 0$.

Proof. The centered regular-simplex identities are

$$\|y_a\|^2_{\mathcal{H}} = \frac{C_{\mathrm{pool}}(m-1)}{2m}, \qquad \langle y_a, y_b\rangle_{\mathcal{H}} = -\frac{C_{\mathrm{pool}}}{2m}, \quad a \ne b.$$

Suppose the first $j < m$ choices have distinct locations and write their sum as T_j. For a vertex type not yet selected,

$$\langle T_j, y_a\rangle_{\mathcal{H}} = -\frac{jC_{\mathrm{pool}}}{2m},$$

whereas for a duplicate of a previously selected type,

$$\langle T_j, y_a\rangle_{\mathcal{H}} = \frac{C_{\mathrm{pool}}(m-j)}{2m}.$$

All vertices have the same norm, so the one-step greedy rule selects a previously unselected type. Induction proves the claimed order through step m. Because two indexed copies of each type were supplied, $\mathcal{C}_j$ contains every simplex vertex throughout these steps. Hence $0 \in \mathcal{C}_j$ and $d_j = 0$.

Since v_k averages k distinct simplex vertices,

$$\left\|\sum_{t=1}^{k} y_{i_t}\right\|^2_{\mathcal{H}} = k\frac{C_{\mathrm{pool}}(m-1)}{2m} + k(k-1)\left(-\frac{C_{\mathrm{pool}}}{2m}\right) = \frac{C_{\mathrm{pool}}k(m-k)}{2m}.$$

Dividing by k^2 proves (58), and comparison with $C_{\mathrm{pool}}/(2k)$ gives the first displayed ratio. Also, $\mathrm{err}_1^2 = C_{\mathrm{pool}}(m-1)/(2m)$. For $j < m$, the remaining pool still contains at least two distinct simplex vertices and has $2m - j$ indexed points. Hence $\mathrm{diam}^2_{\mathcal{H}}(\mathcal{C}_j) = C_{\mathrm{pool}}$ and $\kappa_j = \frac{1}{2}C_{\mathrm{pool}}\{1 - (2m-j)^{-1}\}$. Summing these terms gives the second ratio. □

Proposition 13 shows that the order C_{pool}/k and the constant $1/2$ in the global consequence of Corollary 6 cannot be improved uniformly. Since $d_0 = \bar{\epsilon}_k = 0$, the $k^{-1/2}$ order of the square-root term in the first upper bound of Theorem 1 is also separately unavoidable.

E.1.1 Sharpness of the Remainder Coefficients

The three coefficients in (31) can be attained separately by finite samples. For the first two coefficients, choose distinct $p, q \in \mathbb{R}^d$, let $P = \delta_p$ and $Q = \delta_q$, and take D_1 to consist only of copies of p, so that $u = \mu_P$. Take the proportions below to be rational and choose N_2 so that every count is integral. Let D_2 contain θN_2 copies of p and $(1-\theta)N_2$ copies of q. Then $\bar{\mu}_2 = \mu_{\mathsf{M}}$. For each selected set below, write $\alpha := |S|/N_2$ and $R := [N_2] \setminus S$. If $\alpha = \theta$ and S contains fractions $1-\eta$ and η of p and q, respectively, where

$$0 < \eta \le \min\{1, (1-\theta)/\theta\},$$

then

$$\|\widehat{\mu}(R) - \mu_Q\|_{\mathcal{H}_{\mathrm{E}}} = \frac{\alpha}{1-\alpha}\|\mu_P - \widehat{\mu}(S)\|_{\mathcal{H}_{\mathrm{E}}}.$$

If instead $0 < \alpha < \theta$ and S contains only copies of p, then

$$\|\widehat{\mu}(R) - \mu_Q\|_{\mathcal{H}_{\mathrm{E}}} = \frac{\theta-\alpha}{1-\alpha}\|\mu_P - \mu_Q\|_{\mathcal{H}_{\mathrm{E}}}.$$

For the mixture-sample coefficient, retain $P = \delta_p$ and the all-p reference sample, choose p, q_1, q_2 pairwise distinct, and let $Q = (\delta_{q_1} + \delta_{q_2})/2$. Take D_2 to contain θN_2 copies of p and $(1-\theta)N_2$ copies of q_1, and let S contain all copies of p. Then $\alpha = \theta$, $\widehat{\mu}(S) = \mu_P$, and

$$\|\widehat{\mu}(R) - \mu_Q\|_{\mathcal{H}_{\mathrm{E}}} = \frac{1}{1-\alpha}\|\bar{\mu}_2 - \mu_{\mathsf{M}}\|_{\mathcal{H}_{\mathrm{E}}} > 0.$$

The strict inequality follows because $q_1 \neq q_2$ and the energy kernel is characteristic.

These candidate pools have positive probability under their stated mixture models. Rational approximation gives the same uniform sharpness conclusion for arbitrary admissible proportions.

F Experimental Details and Results

F.1 Mixture Proportion Estimation

As a separate mixture-proportion estimation problem, Ramaswamy et al. (2016) estimate the proportion of a known component from samples of the mixture and the component distribution. We use their kernel-mean estimators KM1 and KM2 to obtain an estimate of the mixture proportion and hence a stopping size for the selection sequence. Let $\mathbf{K}$ denote the Gram matrix formed from the N_2 mixture observations and the N_1 reference observations. Specially, let $w_1, \ldots, w_{N_1+N_2} = (y_1, \ldots, y_{N_1}, z_1, \ldots, z_{N_2})$, $\mathbf{K} \in \mathbb{R}^{(N_1+N_2)\times(N_1+N_2)}$ and the element $K_{ij} = k_E(w_i, w_j)$. For $\lambda \ge 1$, define

$$\widehat{d}^2(\lambda) = \min_{\mathbf{p}\in\Delta_{N_1+N_2}} \{\mathbf{a}(\lambda) - \mathbf{p}\}^\top \mathbf{K}\{\mathbf{a}(\lambda) - \mathbf{p}\},$$

where $\Delta_{N_1+N_2} := \{\mathbf{p} \in \mathbb{R}^{N_1+N_2} : p_i \ge 0, \mathbf{1}^\top \mathbf{p} = 1\}$ and

$$\mathbf{a}(\lambda) = \begin{bmatrix} \lambda N_2^{-1}\mathbf{1}_{N_2} \\ (1-\lambda)N_1^{-1}\mathbf{1}_{N_1} \end{bmatrix}.$$

Let $\widehat{d}(\lambda)$ denote the nonnegative square root of $\widehat{d}^2(\lambda)$. The KM estimators search for the value of λ at which the slope of the convex function $\widehat{d}(\lambda)$ exceeds a prescribed threshold ν. KM1 uses

$$\nu_{\mathrm{KM1}} = \frac{1}{\sqrt{N_1 \wedge N_2}}.$$

KM2 instead uses a threshold interpolating between the initial and final slopes of $\widehat{d}$:

$$\nu_{\mathrm{KM2}} = 0.8\,\widehat{s}_{\mathrm{init}} + 0.2\,\widehat{s}_{\infty}.$$

Here $\widehat{s}_{\text{init}}$ denotes the estimated slope of $\widehat{d}(\lambda)$ near $\lambda = 1$, while the limiting slope is

$$\widehat{s}_{\infty} = \left\| \frac{1}{N_2} \sum_{j=1}^{N_2} \Phi_E(z_j) - \frac{1}{N_1} \sum_{i=1}^{N_1} \Phi_E(y_i) \right\|_{\mathcal{H}_E}.$$

For either threshold, the corresponding estimate $\widehat{\lambda}$ is obtained by the gradient-thresholding search of Ramaswamy et al. (2016) and its Algorithm 1. Denote the resulting estimates by $\widehat{\lambda}_{\text{KM1}}$ and $\widehat{\lambda}_{\text{KM2}}$. The associated mixture-proportion estimates are

$$\widehat{\theta}_r = 1 - \widehat{\lambda}_r^{-1}, \qquad r \in \{\text{KM1}, \text{KM2}\}.$$

Finally, the stopping sizes are

$$k^r = \text{clip}\left(\text{round}(N_2 \widehat{\theta}_r),\, 1,\, N_2 - 1\right), \qquad r \in \{\text{KM1}, \text{KM2}\}.$$

F.2 Selection Methods

GKHR selection rule adapts the Generalized Kernel Herding without Replacement from Shao et al. (2024), but we replace the Gaussian kernel as the energy-distance kernel. This adaptation can be stated directly. For $j \in [N_2]$, define

$$g_j := \frac{1}{N_1} \sum_{i=1}^{N_1} k_{\text{E}}(y_i, z_j).$$

Fix a requested selection size k_{H} and a parameter $0 \le \alpha_{\text{H}} \le 1$. Set $I_0 := \varnothing$, $J_0 := [N_2]$, and $h_{0,j} := 0$. For $1 \le t \le k_{\text{H}}$, choose

$$i_t^{\text{H}} \in \operatorname*{arg\,min}_{j \in J_{t-1}} \{h_{t-1,j} - \alpha_{\text{H}} g_j\}, \qquad I_t := I_{t-1} \cup \{i_t^{\text{H}}\}, \qquad J_t := J_{t-1} \setminus \{i_t^{\text{H}}\}, \tag{59}$$

and update, for $j \in J_t$,

$$h_{t,j} := \frac{t-1}{t} h_{t-1,j} + \frac{1}{t} k_{\text{E}}(z_{i_t^{\text{H}}}, z_j). \tag{60}$$

Let k_{true} denote the realized number of candidates generated from P, and let k^{KM2} be the count returned by KM2. The rows are defined as follows:

1. **EG/PathMin**: the exact greedy ordering from Algorithm 1, stopped by the PathMin rule in (8);
2. **ED-Herding/PathMin**: the ordering from (59)–(60) with $\alpha_{\text{H}} = 1$, run through all N_2 candidates and stopped by the same minimum-discrepancy rule;
3. **EG/KM1** and **EG/KM2**: the EG ordering cut at the corresponding estimate of the mixture proportion;
4. **EG/TrueCount**: the EG ordering cut at the true but hidden number of P-generated candidates, without using their identities;
5. **GKHR/TrueCount**: GKHR with $k_{\text{H}} = k_{\text{true}}$ and $\alpha_{\text{H}} = 1 - 1/\sqrt{k_{\text{H}}}$;
6. **GKHR/KM2**: GKHR with $k_{\text{H}} = k^{\text{KM2}}$ and $\alpha_{\text{H}} = 1 - 1/\sqrt{k_{\text{H}}}$;
7. **LabelOracle**: the diagnostic partition that puts every P-generated candidate in the selected set and every Q-generated candidate in the remainder.

F.3 REAL-DATA EXPERIMENTS

Datasets and components. We use three UCI datasets, (Malerba, 1994; UCI Machine Learning Repository, 1992; Cortez et al., 2009). The respond labels are hidden and used to construct distributions P and Q datasets. For example, in Shuttle dataset, P dataset is the class 1 data and Q is the other classes. Table 2 gives the detailed information. Labels are used to construct the mixtures, evaluation sets, and LabelOracle diagnostic. For Shuttle, we concatenate the official training and test files and form one fixed 10,000-observation subset by stratifying over all seven original classes. For Wine Quality, the quality score is excluded from the covariates.

Table 2: Dataset sizes and class definitions.

Dataset	Total	Dim.	# P	# Q	Components
Page Blocks	5,473	10	4,913	560	P: class 1; Q: classes 2–5
Shuttle	10,000	9	7,860	2,140	P: class 1; Q: classes 2–7
Wine Quality	6,497	11	4,898	1,599	P: white; Q: red

Data preparation and sampling. For each of 50 random repetitions, we first reserve 35% of each binary component as D_{eval} used to compute the evaluation MMD. The other 65% is divided to D_1 and D_2. D_1 and D_2 are sampled without replacement from the data not reserved for evaluation, with no P-observation appearing in both samples. Within a dataset and repetition, D_1 is fixed across all (θ, N_2) settings. The samples are constructed so that a larger N_2 contains the component observations used at smaller values of N_2; the displayed order of D_2 is then randomly permuted. We use $\theta \in \{0.1, 0.3, 0.5\}$. Dataset counts and sample sizes are reported in Tables 2 and 3.

Table 3: Reference and mixture sample sizes.

Dataset	N_1	Low N_2	Medium N_2	High N_2
Page Blocks	2,990	130	260	400
Shuttle	4,330	510	1,030	1,540
Wine Quality	2,650	350	710	1,060

Table 7 and 8 perform similarly to the Shuttle experiment results, seeing Table 1. They verify that under the accurate stopping size k, EG separates the mixture dataset D_2 as S and R well and get a overall balanced low MMD errors to distributions P and Q.

Particularly, we list Table 4, 5, and 6. These comparisons assess how closely the two subsets S and R recovered by the algorithm match the oracle partition of D_2 when the number of P-generated candidates is known (TrueCount, $k = k_{ture}$). For LabelOracle, S_{oracle} and R_{oracle} contain the candidates truly generated from P and Q, respectively, with $e_P(k) = \text{MMD}(S_{\text{oracle}}, P)$, and $e_Q(k) = \text{MMD}(R_{\text{oracle}}, Q)$. Notably, EG can achieve an $e_P(k)$ even smaller than LabelOracle, reflecting its greedy objective of constructing a subset that best approximates P, rather than recovering latent labels. More importantly, EG attains $e_Q(k)$ close to $\text{MMD}(R_{\text{oracle}}, Q)$, indicating that the remainder approximates Q nearly as well as the oracle partition. Thus, with the correct selection size, EG achieves mixture separation comparable to LabelOracle at the distributional level.

Table 4: Benchmarks using the true selection size or true labeled observations for Page Blocks. TrueCount uses the true but hidden number of P observations, whereas LabelOracle uses ture labeled observations. Energy MMD on the reserved evaluation data is reported as mean $\pm$ standard deviation across 50 splits. Bold marks the smallest $e_P(k)$ and $e_Q(k)$; ties in the displayed means are all bold.

		EG/TrueCount		GKHR/TrueCount		LabelOracle	
θ	N_2	$e_P(k)$	$e_Q(k)$	$e_P(k)$	$e_Q(k)$	$e_P(k)$	$e_Q(k)$
0.1	130	**0.1978** $\pm$ 0.0239	0.2347 ± 0.0523	0.5067 ± 0.0646	0.2336 ± 0.0538	0.2912 ± 0.0685	**0.2327** $\pm$ 0.0526
	260	**0.1341** $\pm$ 0.0161	0.1932 ± 0.0428	0.2844 ± 0.0309	0.1930 ± 0.0437	0.2073 ± 0.0451	**0.1922** $\pm$ 0.0419
	400	**0.1034** $\pm$ 0.0111	0.1806 ± 0.0407	0.2054 ± 0.0181	**0.1800** $\pm$ 0.0414	0.1677 ± 0.0368	0.1802 ± 0.0403
0.3	130	**0.1122** $\pm$ 0.0148	0.2639 ± 0.0614	0.1834 ± 0.0273	0.2605 ± 0.0607	0.1714 ± 0.0405	**0.2595** $\pm$ 0.0583
	260	**0.0769** $\pm$ 0.0076	0.2061 ± 0.0468	0.1121 ± 0.0106	0.2057 ± 0.0474	0.1175 ± 0.0266	**0.2024** $\pm$ 0.0437
	400	**0.0633** $\pm$ 0.0063	0.1880 ± 0.0441	0.0862 ± 0.0071	0.1877 ± 0.0443	0.0953 ± 0.0216	**0.1859** $\pm$ 0.0423
0.5	130	**0.0903** $\pm$ 0.0118	0.2949 ± 0.0701	0.1214 ± 0.0163	0.2922 ± 0.0687	0.1286 ± 0.0316	**0.2831** $\pm$ 0.0668
	260	**0.0640** $\pm$ 0.0079	0.2284 ± 0.0497	0.0800 ± 0.0082	0.2269 ± 0.0515	0.0921 ± 0.0207	**0.2227** $\pm$ 0.0480
	400	**0.0541** $\pm$ 0.0069	0.2026 ± 0.0437	0.0634 ± 0.0063	0.2023 ± 0.0439	0.0759 ± 0.0191	**0.1981** $\pm$ 0.0418

Table 5: Benchmarks using the true selection size or true labeled observations for Shuttle. TrueCount uses the true but hidden number of P observations, whereas LabelOracle uses ture labeled observations. Energy MMD on the reserved evaluation data is reported as mean $\pm$ standard deviation across 50 splits. Bold marks the smallest $e_P(k)$ and $e_Q(k)$; ties in the displayed means are all bold.

		EG/TrueCount		GKHR/TrueCount		LabelOracle	
θ	N_2	$e_P(k)$	$e_Q(k)$	$e_P(k)$	$e_Q(k)$	$e_P(k)$	$e_Q(k)$
0.1	510	**0.0912** $\pm$ 0.0122	0.0804 ± 0.0193	0.1516 ± 0.0286	**0.0796** $\pm$ 0.0195	0.1300 ± 0.0328	**0.0796** $\pm$ 0.0194
	1030	**0.0638** $\pm$ 0.0084	0.0655 ± 0.0204	0.1010 ± 0.0111	0.0655 ± 0.0207	0.0930 ± 0.0224	**0.0644** $\pm$ 0.0205
	1540	**0.0536** $\pm$ 0.0078	0.0596 ± 0.0171	0.0818 ± 0.0067	0.0599 ± 0.0173	0.0761 ± 0.0181	**0.0589** $\pm$ 0.0168
0.3	510	**0.0578** $\pm$ 0.0094	0.0897 ± 0.0235	0.0717 ± 0.0106	0.0892 ± 0.0237	0.0766 ± 0.0177	**0.0864** $\pm$ 0.0234
	1030	**0.0430** $\pm$ 0.0065	0.0734 ± 0.0205	0.0513 ± 0.0058	0.0739 ± 0.0207	0.0564 ± 0.0132	**0.0710** $\pm$ 0.0202
	1540	**0.0374** $\pm$ 0.0053	0.0634 ± 0.0186	0.0445 ± 0.0048	0.0643 ± 0.0188	0.0487 ± 0.0104	**0.0614** $\pm$ 0.0192
0.5	510	**0.0480** $\pm$ 0.0077	0.1036 ± 0.0222	0.0537 ± 0.0079	0.1037 ± 0.0222	0.0619 ± 0.0138	**0.0952** $\pm$ 0.0216
	1030	**0.0372** $\pm$ 0.0054	0.0834 ± 0.0222	0.0410 ± 0.0051	0.0846 ± 0.0221	0.0471 ± 0.0101	**0.0784** $\pm$ 0.0204
	1540	**0.0323** $\pm$ 0.0045	0.0743 ± 0.0213	0.0354 ± 0.0043	0.0758 ± 0.0211	0.0390 ± 0.0075	**0.0688** $\pm$ 0.0209

Table 6: Benchmarks using the true selection size or true labeled observations for Wine Quality. TrueCount uses the true but hidden number of P observations, whereas LabelOracle uses ture labeled observations. Energy MMD on the reserved evaluation data is reported as mean $\pm$ standard deviation across 50 splits. Bold marks the smallest $e_P(k)$ and $e_Q(k)$; ties in the displayed means are all bold.

		EG/TrueCount		GKHR/TrueCount		LabelOracle	
θ	N_2	$e_P(k)$	$e_Q(k)$	$e_P(k)$	$e_Q(k)$	$e_P(k)$	$e_Q(k)$
0.1	350	**0.1808** $\pm$ 0.0147	0.1029 ± 0.0164	0.2018 ± 0.0168	0.1034 ± 0.0162	0.2307 ± 0.0326	**0.1017** $\pm$ 0.0158
	710	**0.1245** $\pm$ 0.0101	0.0862 ± 0.0146	0.1353 ± 0.0095	0.0865 ± 0.0146	0.1631 ± 0.0227	**0.0856** $\pm$ 0.0146
	1060	**0.1039** $\pm$ 0.0102	0.0785 ± 0.0152	0.1163 ± 0.0089	0.0788 ± 0.0153	0.1400 ± 0.0232	**0.0782** $\pm$ 0.0152
0.3	350	**0.1123** $\pm$ 0.0120	0.1158 ± 0.0218	0.1170 ± 0.0114	0.1171 ± 0.0214	0.1404 ± 0.0232	**0.1119** $\pm$ 0.0212
	710	**0.0812** $\pm$ 0.0083	0.0943 ± 0.0168	0.0852 ± 0.0078	0.0952 ± 0.0166	0.1002 ± 0.0153	**0.0925** $\pm$ 0.0165
	1060	**0.0700** $\pm$ 0.0086	0.0841 ± 0.0150	0.0743 ± 0.0080	0.0848 ± 0.0147	0.0851 ± 0.0146	**0.0831** $\pm$ 0.0153
0.5	350	**0.0908** $\pm$ 0.0099	0.1395 ± 0.0291	0.0934 ± 0.0092	0.1413 ± 0.0286	0.1100 ± 0.0158	**0.1292** $\pm$ 0.0250
	710	**0.0688** $\pm$ 0.0078	0.1044 ± 0.0163	0.0710 ± 0.0076	0.1058 ± 0.0162	0.0807 ± 0.0127	**0.0987** $\pm$ 0.0146
	1060	**0.0596** $\pm$ 0.0069	0.0947 ± 0.0157	0.0619 ± 0.0066	0.0965 ± 0.0151	0.0679 ± 0.0101	**0.0905** $\pm$ 0.0160

Table 7: Page Blocks results with $N_1 = 2990$, different N_2 sizes, and true mixtures θs. Entries are mean $\pm$ standard deviation across 50 splits. GKHR/KM2 denotes GKHR stopped at the number of points specified by KM2. Bold marks the α_k closest to θ and the smallest $e_P(k)$ and $e_Q(k)$; ties in the displayed means are all bold.

		α_k				$e_P(k)$					$e_Q(k)$				
θ	N_2	EG/PathMin	EDH/PathMin	KM1	KM2	EG/PathMin	EDH/PathMin	EG/KM1	EG/KM2	GKHR/KM2	EG/PathMin	EDH/PathMin	EG/KM1	EG/KM2	GKHR/KM2
0.1	130	**0.126 ± 0.018**	0.236 ± 0.028	0.474 ± 0.113	0.180 ± 0.023	**0.1870 ± 0.0236**	0.3104 ± 0.0283	0.4062 ± 0.0683	0.2065 ± 0.0254	0.3366 ± 0.0430	**0.2398 ± 0.0541**	0.2833 ± 0.0589	0.5478 ± 0.1809	0.2604 ± 0.0592	0.2543 ± 0.0581
	260	**0.106 ± 0.015**	0.171 ± 0.016	0.323 ± 0.056	0.158 ± 0.015	**0.1306 ± 0.0166**	0.2181 ± 0.0216	0.2810 ± 0.0394	0.1520 ± 0.0207	0.2233 ± 0.0242	**0.1946 ± 0.0438**	0.2074 ± 0.0501	0.3147 ± 0.0595	0.2060 ± 0.0496	0.2040 ± 0.0498
	400	**0.097 ± 0.012**	0.145 ± 0.011	0.261 ± 0.034	0.149 ± 0.011	**0.1014 ± 0.0103**	0.1723 ± 0.0153	0.2211 ± 0.0271	0.1254 ± 0.0148	0.1755 ± 0.0146	**0.1804 ± 0.0408**	0.1865 ± 0.0439	0.2484 ± 0.0470	0.1890 ± 0.0440	0.1874 ± 0.0443
0.3	130	**0.264 ± 0.031**	0.356 ± 0.028	0.602 ± 0.155	0.356 ± 0.028	**0.1088 ± 0.0123**	0.1724 ± 0.0243	0.2937 ± 0.1045	0.1225 ± 0.0146	0.1741 ± 0.0240	**0.2587 ± 0.0587**	0.2822 ± 0.0679	0.6428 ± 0.2463	0.2906 ± 0.0738	0.2823 ± 0.0718
	260	0.254 ± 0.030	**0.315 ± 0.021**	0.572 ± 0.073	0.345 ± 0.019	**0.0733 ± 0.0060**	0.1079 ± 0.0110	0.2377 ± 0.0525	0.0859 ± 0.0081	0.1117 ± 0.0102	**0.2095 ± 0.0448**	0.2096 ± 0.0493	0.4729 ± 0.1229	0.2204 ± 0.0514	0.2178 ± 0.0520
	400	0.248 ± 0.024	**0.298 ± 0.018**	0.497 ± 0.045	0.341 ± 0.014	**0.0593 ± 0.0055**	0.0828 ± 0.0074	0.1741 ± 0.0311	0.0727 ± 0.0073	0.0892 ± 0.0074	0.1953 ± 0.0422	**0.1889 ± 0.0448**	0.3356 ± 0.0649	0.1980 ± 0.0495	0.1960 ± 0.0491
0.5	130	0.414 ± 0.052	**0.501 ± 0.039**	0.572 ± 0.099	0.528 ± 0.035	**0.0827 ± 0.0074**	0.1181 ± 0.0165	0.1171 ± 0.0579	0.0926 ± 0.0093	0.1205 ± 0.0151	**0.2968 ± 0.0740**	0.2985 ± 0.0746	0.3949 ± 0.1970	0.3174 ± 0.0790	0.3127 ± 0.0779
	260	0.394 ± 0.039	0.467 ± 0.028	0.697 ± 0.132	**0.531 ± 0.026**	**0.0582 ± 0.0051**	0.0769 ± 0.0076	0.1742 ± 0.0832	0.0683 ± 0.0061	0.0814 ± 0.0074	0.2588 ± 0.0587	**0.2317 ± 0.0523**	0.5735 ± 0.2802	0.2436 ± 0.0542	0.2398 ± 0.0564
	400	0.385 ± 0.033	0.444 ± 0.031	0.726 ± 0.083	**0.530 ± 0.021**	**0.0484 ± 0.0049**	0.0591 ± 0.0049	0.1794 ± 0.0586	0.0584 ± 0.0054	0.0660 ± 0.0054	0.2498 ± 0.0540	0.2191 ± 0.0518	0.5728 ± 0.1945	0.2165 ± 0.0483	**0.2159 ± 0.0485**

Table 8: Wine Quality results with $N_1 = 2650$, different N_2 sizes, and true mixtures θs. Entries are mean $\pm$ standard deviation across 50 splits. GKHR/KM2 denotes GKHR stopped at the number of points specified by KM2. Bold marks the α_k closest to θ and the smallest $e_P(k)$ and $e_Q(k)$; ties in the displayed means are all bold.

		α_k				$e_P(k)$					$e_Q(k)$				
θ	N_2	EG/PathMin	EDH/PathMin	KM1	KM2	EG/PathMin	EDH/PathMin	EG/KM1	EG/KM2	GKHR/KM2	EG/PathMin	EDH/PathMin	EG/KM1	EG/KM2	GKHR/KM2
0.1	350	0.094 ± 0.009	**0.099 ± 0.010**	0.097 ± 0.013	0.117 ± 0.011	**0.1768 ± 0.0128**	0.1967 ± 0.0173	0.1805 ± 0.0130	0.1938 ± 0.0150	0.2104 ± 0.0159	**0.1039 ± 0.0164**	0.1043 ± 0.0161	0.1045 ± 0.0167	0.1043 ± 0.0167	0.1044 ± 0.0163
	710	0.089 ± 0.007	0.090 ± 0.008	**0.094 ± 0.011**	0.112 ± 0.006	**0.1205 ± 0.0076**	0.1245 ± 0.0082	0.1225 ± 0.0079	0.1352 ± 0.0135	0.1425 ± 0.0126	0.0873 ± 0.0145	0.0878 ± 0.0147	0.0875 ± 0.0149	0.0870 ± 0.0145	**0.0869 ± 0.0146**
	1060	0.086 ± 0.006	0.087 ± 0.008	0.088 ± 0.009	**0.110 ± 0.004**	**0.0993 ± 0.0066**	0.1011 ± 0.0066	0.1003 ± 0.0065	0.1139 ± 0.0144	0.1230 ± 0.0134	0.0805 ± 0.0151	0.0807 ± 0.0150	0.0803 ± 0.0156	**0.0791 ± 0.0152**	**0.0791 ± 0.0153**
0.3	350	0.245 ± 0.019	0.250 ± 0.019	**0.314 ± 0.015**	0.315 ± 0.013	**0.1024 ± 0.0063**	0.1056 ± 0.0068	0.1180 ± 0.0086	0.1187 ± 0.0096	0.1228 ± 0.0099	0.1441 ± 0.0231	0.1421 ± 0.0237	0.1172 ± 0.0222	**0.1170 ± 0.0227**	0.1181 ± 0.0222
	710	0.243 ± 0.012	0.244 ± 0.014	**0.304 ± 0.011**	0.315 ± 0.010	**0.0733 ± 0.0055**	0.0742 ± 0.0052	0.0821 ± 0.0059	0.0882 ± 0.0073	0.0916 ± 0.0071	0.1279 ± 0.0181	0.1275 ± 0.0220	**0.0949 ± 0.0172**	0.0953 ± 0.0176	0.0957 ± 0.0175
	1060	0.238 ± 0.011	0.241 ± 0.011	**0.298 ± 0.012**	0.313 ± 0.008	**0.0630 ± 0.0057**	0.0632 ± 0.0055	0.0685 ± 0.0060	0.0763 ± 0.0079	0.0796 ± 0.0077	0.1248 ± 0.0175	0.1232 ± 0.0171	0.0857 ± 0.0154	**0.0849 ± 0.0152**	**0.0849 ± 0.0150**
0.5	350	0.404 ± 0.022	0.409 ± 0.022	0.535 ± 0.016	**0.511 ± 0.015**	**0.0802 ± 0.0058**	0.0814 ± 0.0054	0.1061 ± 0.0092	0.0940 ± 0.0078	0.0959 ± 0.0070	0.2329 ± 0.0338	0.2297 ± 0.0334	0.1431 ± 0.0301	**0.1398 ± 0.0296**	0.1413 ± 0.0291
	710	0.390 ± 0.017	0.395 ± 0.015	0.513 ± 0.014	**0.509 ± 0.012**	**0.0608 ± 0.0058**	0.0612 ± 0.0057	0.0729 ± 0.0068	0.0715 ± 0.0066	0.0731 ± 0.0064	0.2321 ± 0.0296	0.2261 ± 0.0280	0.1052 ± 0.0176	**0.1047 ± 0.0170**	0.1061 ± 0.0171
	1060	0.384 ± 0.013	0.387 ± 0.014	**0.507 ± 0.011**	0.510 ± 0.010	**0.0539 ± 0.0064**	0.0541 ± 0.0064	0.0614 ± 0.0066	0.0624 ± 0.0067	0.0642 ± 0.0068	0.2354 ± 0.0272	0.2312 ± 0.0297	0.0962 ± 0.0164	**0.0956 ± 0.0161**	0.0962 ± 0.0159

F.4 Orthogonal-Contamination Rate Experiment

This experiment evaluates the rate statement in the case (a) of Theorem 4 under $N_1 = N_2 = N$, rather than comparing alternative stopping heuristics. Specifically, it tests whether a fixed upper quantile of

$$T_N = \overline{d}_{k_\theta} + err_{k_\theta} + \mathrm{MMD}_{\mathrm{E}}(S_{k_\theta}, P) + \mathrm{MMD}_{\mathrm{E}}(R_{k_\theta}, Q_{\phi_N})$$

decreases at an order compatible with

$$|k_\theta/N - \theta| + \omega\Big(2N^{-1/2} + \zeta_N\Big), \qquad \omega(t) = t\log(e/t).$$

We use a known-fraction k_θ resulting in $|k_\theta/N - \theta| = 0$ so that stopping-size estimation does not obscure the greedy-sequence rate $\omega\big(2N^{-1/2} + \zeta_N\big)$.

Population model. All observations lie in $\mathbb{R}^4$. The target and exactly orthogonal contaminant distributions are

$$P = \mathrm{Law}\big(\cos U, \sin U, 0, 0\big), \qquad Q_0 = \mathrm{Law}\big(0, 0, \cos V, \sin V\big), \qquad U, V \overset{\text{iid}}{\sim} \mathrm{Unif}[0, 2\pi).$$

Thus both components are continuous and have no point masses, while their supports occupy orthogonal two-dimensional subspaces. Approximate orthogonality is generated by the coherently perturbed contaminant

$$Q_{\phi_N} = \mathrm{Law}\big(\sin\phi_N,\ 0,\ \cos\phi_N \cos V,\ \cos\phi_N \sin V\big).$$

Coupling a perturbed observation $q_{\phi_N}(V)$ with $q_0(V)$, and using $\|\Phi_{\mathrm{E}}(x) - \Phi_{\mathrm{E}}(y)\|^2_{\mathcal{H}_{\mathrm{E}}} = \|x - y\|_2$, Equation (50) gives

$$\|\Pi_{\mathcal{L}_P}\{\Phi_{\mathrm{E}}(q_{\phi_N}(V)) - \mu_P\}\|_{\mathcal{H}_{\mathrm{E}}} \le \sqrt{2\sin(\phi_N/2)}.$$

For each prescribed budget

$$\zeta_N \in \{0, N^{-1/2}, N^{-1/4}\},$$

we therefore set

$$\phi_N = 2\arcsin(\zeta_N^2/2),$$

so that the displayed certificate equals ζ_N.

Experiment setting. We use

$$N_1 = N_2 = N \in \{100, 200, 500, 1000, 2000, 3000\}, \qquad \theta = 0.3.$$

We perform 50 repetitions for every combination of N and ζ_N. For each repetition, the reference sample contains N iid observations from P. Each candidate independently comes from P with probability θ and from Q_{ϕ_N} otherwise. Consequently, the realized number of target-generated candidates is random. All sample sizes used here satisfy $k_\theta/N = \theta$, so the selected-fraction term is exactly zero.

Population energy-MMDs are evaluated analytically from the circle distributions, rather than from finite evaluation samples. The remaining-pool projections used to calculate $\overline{d}_{k_\theta}$ are solved with a certified squared-objective Frank–Wolfe gap of at most 10^{-6}; the upper endpoint of each projection certificate is used in T_N. In all repetitions, the largest final uncertainty in $\overline{d}_{k_\theta}$ was 1.05×10^{-4}.

Population energy-MMD evaluation. For $a \in \mathbb{R}$ and $b \geq 0$, define

$$\psi(a, b) := \frac{1}{2\pi} \int_0^{2\pi} \sqrt{2 - 2a - 2b\cos t}\, dt = \frac{2}{\pi}\sqrt{2(1-a+b)}\, F_\ell\left(\frac{2b}{1-a+b}\right),$$

where $F_\ell(\cdot)$ is the complete elliptic integral of the second kind. For a candidate $z = (z_1, z_2, z_3, z_4)$, let $r_P(z) = \sqrt{z_1^2 + z_2^2}, r_Q(z) = \sqrt{z_3^2 + z_4^2}$. Since all observations have unit norm, for any nonempty index set A,

$$\mathrm{MMD}_{\mathrm{E}}^2(\widehat{P}_A, P) = \frac{1}{|A|}\sum_{i\in A} \psi(0, r_P(z_i)) - \frac{1}{2|A|^2}\sum_{i,j\in A} \|z_i - z_j\|_2 - \frac{2}{\pi},$$

because $\mathbb{E}_{X,X'\sim P}\|X - X'\|_2 = \frac{4}{\pi}$.

Likewise, writing $s_N = \sin\phi_N$ and $c_N = \cos\phi_N$,

$$\mathrm{MMD}_{\mathrm{E}}^2(\widehat{P}_A, Q_{\phi_N}) = \frac{1}{|A|}\sum_{i\in A} \psi(s_N z_{i1},\, c_N r_Q(z_i)) - \frac{1}{2|A|^2}\sum_{i,j\in A} \|z_i - z_j\|_2 - \frac{2c_N}{\pi},$$

since $\mathbb{E}_{Z,Z'\sim Q_{\phi_N}}\|Z - Z'\|_2 = \frac{4c_N}{\pi}$.

We evaluate $\mathrm{MMD}_{\mathrm{E}}(S_{k_\theta}, P)$ and $\mathrm{MMD}_{\mathrm{E}}(R_{k_\theta}, Q_{\phi_N})$ by applying the two formulas with $A = S_{k_\theta}$ and $A = R_{k_\theta}$, respectively, and taking the nonnegative square root. These calculations use the observed points directly and require neither latent component labels nor finite evaluation samples.

Reported statistics. For every experimental cell, we report the empirical 90% quantile of T_N and its components. Shaded regions are 95% percentile-bootstrap intervals for the empirical quantile, based on 2000 bootstrap resamples of the 50 repetitions. Panel (a) additionally displays $c\,\omega(2/\sqrt{N})$, with c anchored at the first observed point; this curve illustrates only the predicted rate shape and is not a finite-sample confidence bound. Panel (b) reports

$$\frac{q_{.9}(T_N)}{|k_\theta/N - \theta| + \omega(2N^{-1/2} + \zeta_N)},$$

but due to the k_θ setting, $|k_\theta/N - \theta| = 0$ in practice.

G LIMITATIONS

Theoretical limits. The general greedy bounds depend on the sequence actually selected; they do not give an unconditional distribution-free rate, and the observed error err_k already provides a more precise numerical diagnostic. The bound for $\overline{d}_k$ in Theorem 8 applies to any distribution, but it is useful only when the proof can identify a set I for which $\bar{x}_I$ is close to u and condition (38) holds. It does not assert that every candidate set contains such an I. Taking $I = I_P$ in Corollary 9 and case (a) of Theorem 4 removes the requirement that the target have finite support but requires condition (20) for a fixed ζ. Exact orthogonality ($\zeta = 0$) gives consistency; for fixed $0 < \zeta \leq 1$, the

bound retains a nonvanishing term of order $\zeta \log(e/\zeta)$. The exact projection condition is strong: if $\mathcal{L}_P = \mathcal{H}_{\mathrm{E}}$, no nonzero orthogonal direction remains for Q. For the injective energy feature map, it also confines any probability that Q assigns within $\mathrm{supp}(P)$ to an exceptional set of at most one feature point, thereby excluding any shared component without point masses. It is not implied by ordinary support separation, distinct mean embeddings, or the target mean lying in the interior of the target feature convex hull. In the orthogonal-sphere example, neither component has point masses, but the construction is highly symmetric and the component supports are disjoint. The perturbation interpretation changes Q uniformly while holding P and $\mathcal{L}_P$ fixed; perturbing P as well would require separate control of $\mathcal{L}_P$. Case (b) of Theorem 4 instead permits Q-component projections relative to individual target support points. It requires Assumption 3 and the exact greedy rule (7). As the target feature points approach affine dependence, the admissible projected displacement shrinks. Because the support-point comparison is pointwise, it supplies no common numerical margin. An approximate-search or finite-precision extension would require both a uniform score advantage and a quantified score-error bound; neither is proved here. The result does not hold uniformly as the support size grows or the affine separation tends to zero. The two sets of assumptions cover different cases; neither covers the continuous overlapping mixtures used in the experiments. Outside them, the mixture assumptions do not control $\overline{\epsilon}_k$. Lemma 7 controls d_0, but not $\overline{\epsilon}_k$. The logarithm in Theorem 8 arises from $\sum_j 1/s_j \le \Lambda_{M,k}$. A sharper bound might use the realized values of s_j and $\|A_{j,\mathcal{L}}\|_{\mathcal{H}}$, but we neither prove such a bound nor show that the logarithmic factor is necessary. The condition $d_j = 0$ may be strong. Moreover, d_{k-1} may grow near the correct stopping point because the remaining points should then resemble Q rather than P.

Stopping rules. Theorem 2 assumes rates for $\mathrm{err}(S_{\widehat{k}})$ and $|\alpha_{\widehat{k}} - \theta|$; Theorem 1 bounds the former by $d_0 + \overline{\epsilon}_{\widehat{k}} + \sqrt{C_{\mathrm{pool}}/(2\widehat{k})}$, where d_0 has the same order as sampling noise. For known θ, the cut $\widehat{k} = k_\theta$ satisfies $|\alpha_{\widehat{k}} - \theta| \le (2N_2)^{-1}$. Thus neither case of Theorem 4 estimates an unknown θ. Equation (35) shows how the rate of any separately justified estimator would enter the guarantees, but θ is not identifiable without additional assumptions. The finite-difference KM1 rule (with its documented KM2 fallback) and the KM2 rule used in the experiments are heuristic baselines: this paper proves no rate for their $|\alpha_{\widehat{k}} - \theta|$ or for either group error in Theorem 2. The two cases of the structured theorem bound $\overline{d}_{\widehat{k}}$ only under their respective geometric assumptions. The examples in Appendix E.1.1 show that each coefficient in the remainder bound can be attained separately, but they do not provide a practical guarantee for $|\alpha_{\widehat{k}} - \theta|$. Beyond the structured settings with known θ, control of $\alpha_{\widehat{k}}$ and control of the greedy sequence remain separate requirements.

G.1 Related Work Supplemental

This section provides Table 9 in the related work section. It summaries the distinct research content among ours and others.

Work	Target	Candidate reuse	Relevant focus
MMD-Critic (Kim et al., 2016)	Matched empirical mean	Distinct prototypes	Prototype and criticism selection; the matched-target selection sequence used here
Teymur et al. (2021)	External	Fixed pool; repetitions allowed	Exact greedy MMD updates and finite-candidate approximation
Pronzato (2023)	External	Fixed domain	Greedy-MMD and kernel-herding performance bounds
Shao et al. (2024)	Matched empirical mean	Without replacement	Bound for small selected sets when enough favorable candidates remain at each step
This paper	External	Without replacement	Separates initial mismatch from removal loss and retains path-specific remaining-set geometry; gives a projection condition and a stopping identity for the remainder

Table 9: Closest greedy-MMD settings. "Matched" means that the target is the uniform mean of the candidate pool.