NONPARAMETRIC MULTIPLE CHANGE-POINT DETECTION FOR RANDOM OBJECTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Change-point analysis for non-Euclidean data has become increasingly important in modern applications, where observations often lie in complex metric spaces rather than vector spaces. We propose a distance profiles-based change-point estimation framework for sequences of objects taking values in general metric spaces, which combines distance profiles with a dynamic programming scheme to enable efficient multiple change-point detection. The proposed method is fully nonparametric, requires minimal tuning, and is applicable as long as pairwise distances between observations are available. We establish a non-asymptotic oracle inequality for our method under mild regularity conditions. Extensive simulation studies demonstrate favorable performance compared to existing approaches. We further illustrate the practical relevance of our method through applications to three representative non-Euclidean datasets, including compositional data, network-valued data, and financial correlation matrices.

1 INTRODUCTION

In traditional change-point detection, most methods assume that the data lie in a Euclidean space, such as real-valued sequences or vectors in $\mathbb{R}^d$. However, modern applications increasingly involve non-Euclidean data objects, including probability distributions (Matabuena et al., 2021), networks (Sporns, 2013; Nie et al., 2017), and compositional data (Gloor et al., 2016). Such data reside on non-Euclidean spaces and lack a natural linear structure, rendering classical vector-space methods inadequate for their characterization.

Considerable attention has been devoted to change-point analysis for non-Euclidean data. For instance, Peel & Clauset (2015); Jeong et al. (2016); Wang et al. (2021); Chen et al. (2023) investigated change-point detection in network data, Horváth et al. (2021) addressed change-point analysis for distributional data, and Prabuchandran et al. (2022) developed techniques for compositional data. While effective within their respective domains, these approaches are typically tailored to specific data structures and thus lack general applicability. The objective of this paper is to develop a general change-point detection method that does not depend on any particular non-Euclidean data structure.

In the change-point detection literature for Euclidean data, a central focus is nonparametric change-point detection, which aims to detect differences between entire distributions rather than being limited to changes in mean or variance. Existing nonparametric methods can be broadly categorized into kernel-based, graph-based, and distance-based approaches. Kernel-based methods (Harchaoui & Cappé, 2007; Harchaoui et al., 2008; Li et al., 2015; Garreau & Arlot, 2018; Arlot et al., 2019; Chang et al., 2019) provide versatile tools for detecting distributional differences. Graph-based methods (Wang et al., 2014; Chen & Zhang, 2015; Shi et al., 2017; Nie & Nicolae, 2021; Chen & Chu, 2023) leverage similarity graphs to capture complex dependencies in the data. Distance-based approaches (Matteson & James, 2014; Li, 2020; Chakraborty & Zhang, 2021) rely on pairwise distances between samples to detect changes in the underlying distribution.

In this paper, we develop a novel nonparametric multiple change-point detection method for random objects using the recent tool of distance profiles introduced by Dubey et al. (2024). Specifically, we first reformulate the nonparametric change-point detection problem as a regression problem (see (2)). We then propose a least squares segmentation criterion to estimate the locations of change points (see (3)), and develop a dynamic programming algorithm to compute these locations. Since

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.

our method depends on distance profiles, we call it the distance-profile-based change-point detection (DPDCP) procedure.

A recent related work is Dubey & Zheng (2026), which also uses distance profiles to identify distributional breaks. It is worth noting that our method and Dubey & Zheng (2026) differ significantly. Dubey & Zheng (2026) constructs a scan statistic based on distance profiles, localizes a single change point by testing for distributional differences, and then detects multiple change points using a binary segmentation strategy. In contrast, our method employs the proposed least squares criterion and globally optimizes multiple change-point locations via a dynamic programming algorithm. This global optimization avoids the error propagation inherent in sequential procedures and can be computed efficiently. Furthermore, the least squares formulation provides a principled framework for model selection, enabling data-driven determination of the number of change points.

2 METHODOLOGY

2.1 DISTANCE-PROFILE-BASED LEAST-SQUARES CRITERION

Our method depends on distance profiles (Dubey et al., 2024). For clarity, we first briefly review them. Let (Ω, d) be a separable metric space, where d is a metric on Ω, and let X be a random object taking values in Ω with distribution $\mathbb{P}$. For any $w \in \Omega$, the distance profile of X at w is defined as

$$F_{\mathbb{P}}(w,t) = \mathbb{P}\big(d(w,X) \le t\big), \;\; t \in [0,+\infty).$$

An appealing property of $F_{\mathbb{P}}(w,t)$ is that it uniquely characterizes the distribution of X. That is, under the assumption that (Ω, d) is of strong negative type, for any two probability measures $\mathbb{P}_1$ and $\mathbb{P}_2$, $\mathbb{P}_1 = \mathbb{P}_2$ if and only if $F_{\mathbb{P}_1}(w,t) = F_{\mathbb{P}_2}(w,t)$ for all $w \in \Omega$ and $t \in [0,+\infty)$. See Proposition 1 in Dubey et al. (2024). This property ensures that our distance-profile-based method can effectively detect changes in the underlying distribution of the random object X.

We next reformulate the nonparametric change-point detection problem as a novel regression problem. To this end, we recall the classical mean change-point model (Yao, 1988):

$$\boldsymbol{Z}_i = \mu_k^\star + \varepsilon_i, \; t_k^\star < i \le t_{k+1}^\star, \; k = 0, \ldots, K,$$

where $0 = t_0^\star < t_1^\star < \cdots < t_K^\star < t_{K+1}^\star = n$ are the true underlying change-point locations, and $\{\boldsymbol{Z}_i\}_{i=1}^n$ is a sequence of vector-valued data. Here, $\mu_k^\star = \mathbb{E}\big[\boldsymbol{Z}_i\big]$, for $t_k^\star < i \le t_{k+1}^\star$. To estimate $t_k^\star$s, a commonly used strategy is to minimize the following loss function with respect to $(t_1, \ldots, t_K)$ satisfying $0 = t_0 < t_1 < \cdots < t_K < t_{K+1} = n$:

$$\frac{1}{n} \sum_{k=0}^{K} \sum_{i=t_k+1}^{t_{k+1}} \|\boldsymbol{Z}_i - \bar{\boldsymbol{Z}}_{(t_k, t_{k+1}]}\|^2, \tag{1}$$

where $\bar{\boldsymbol{Z}}_{(t_k,t_{k+1}]} = (t_{k+1} - t_k)^{-1} \sum_{j=t_k+1}^{t_{k+1}} \boldsymbol{Z}_j$. See, Yao (1988) and Lavielle & Moulines (2000).

Motivated by the above classical mean change-point method, we propose the following change-point regression model based on distance profiles for the random object sequence $\{X_i\}_{i=1}^n$:

$$\mathbb{I}\{d(w, X_i) \le t\} = F_k^\star(w,t) + \varepsilon_i(w,t), \; t_k^\star < i \le t_{k+1}^\star, \; k = 0, \ldots, K, \tag{2}$$

for all $w \in \Omega$ and $t \in [0,+\infty)$. Here, $\mathbb{I}\{\cdot\}$ is the indicator function, and $F_k^\star(w,t) = \mathbb{E}\big[\mathbb{I}\{d(w,X_i) \le t\}\big]$, for $t_k^\star < i \le t_{k+1}^\star$. Since distance profiles uniquely identify probability measures, model (2) fully characterizes changes in the underlying distribution of $\{X_i\}_{i=1}^n$. Thus, model (2) provides a nonparametric regression formulation of the change-point detection problem. To the best of our knowledge, this framework is novel in the nonparametric change-point literature. In particular, no analogous nonparametric regression approach exists even for Euclidean data.

Note that the loss function in (1) can be viewed as the sum of the within-group variances. Similarly, for model (2), we define a corresponding within-group variation for any fixed (w,t) as

$$\frac{1}{n} \sum_{k=0}^{K} \sum_{i=t_k+1}^{t_{k+1}} \Big[\mathbb{I}\{d(w,X_i) \le t\} - \widehat{F}_{(t_k,t_{k+1}]}(w,t)\Big]^2,$$

where $\widehat{F}_{(t_k,t_{k+1}]}(w,t)$ is the empirical counterpart of $F_k^\star(w,t)$ based on the subsample $\{X_{t_k+1},\ldots,X_{t_{k+1}}\}$, defined by

$$\widehat{F}_{(t_k,t_{k+1}]}(w,t)=\frac{1}{t_{k+1}-t_k}\sum_{j=t_k+1}^{t_{k+1}}\mathbb{I}\{d(w,X_j)\le t\}.$$

Integrating (w,t) over $\Omega\times[0,+\infty)$, we obtain the following distance-profile-based least squares criterion as

$$L_n(T_K)=\frac{1}{n}\sum_{k=0}^{K}\sum_{i=t_k+1}^{t_{k+1}}\int_{\Omega\times[0,+\infty)}\Big[\mathbb{I}\{d(w,X_i)\le t\}-\widehat{F}_{(t_k,t_{k+1}]}(w,t)\Big]^2\,dH(w,t), \tag{3}$$

where $T_K=(t_1,\ldots,t_K)$ with $0=t_0<t_1<\cdots<t_K<t_{K+1}=n$, and $H(w,t)$ is a pre-specified weight function on $\Omega\times[0,+\infty)$.

Since the definition of $L_n(T_K)$ involves a double integral, selecting an appropriate weight function $H(w,t)$ is crucial for computational feasibility. In this article, we focus on the following two types of H, namely **(H1)** and **(H2)**:

(H1) We take $dH(w,t)$ as

$$dH(w,t)=d\mu(w)\,d\widehat{F}_w(t),$$

where $\mu(w)=n^{-1}\sum_{i=1}^n\delta_{X_i}(w)$ and $\widehat{F}_w(t)=n^{-1}\sum_{r=1}^n\mathbb{I}\{d(w,X_r)\le t\}$. Here, $\mu(w)$ is the empirical measure on $\{X_i\}_{i=1}^n$ and $\widehat{F}_w(t)$ is the empirical distance profile distribution.

(H2) We take $dH(w,t)$ as

$$dH(w,t)=d\mu(w)\,d\nu(t),$$

where $\mu(w)=n^{-1}\sum_{i=1}^n\delta_{X_i}(w)$ and $d\nu(t)=(t_{\max}-t_{\min})^{-1}\mathbb{I}_{[t_{\min},t_{\max}]}(t)dt$, with $t_{\min}=\min_{1\le i<j\le n}d(X_i,X_j)$, and $t_{\max}=\max_{1\le i<j\le n}d(X_i,X_j)$. Here, $\nu(t)$ is the uniform distribution on $[t_{\min},t_{\max}]$.

As shown in the fowllowing Theorems 1 and 2, the above choices of $H(w,t)$ yield a closed-form expression for $L_n(T_K)$.

Note that the number of change points K in (3) is unknown. To estimate both the locations and the number of change points, we minimize the following penalized optimization problem:

$$\mathcal{C}_n(T_K)=L_n(T_K)+\frac{1}{n}\Big[c_1\log\binom{n-1}{K}+c_2(K+1)\Big],$$

where c_1 and c_2 are tuning parameters. Here, $\binom{n-1}{K}$ counts the candidate segmentations with K change points, so its logarithm penalizes the size of the candidate set, while $K+1$ accounts for the number of fitted segments. Together with $L_n(T_K)$, these terms balance fit and complexity. This type of complexity penalty arises in Gaussian least squares model selection (Birgé & Massart, 2007, Section 1.3 and Theorem 1). We refer to the procedure that minimizes $C_n(T_K)$ to estimate the locations and the number of change points as the DPDCP procedure.

2.2 Implementation for DPDCP

In this section, we provide a dynamic programming algorithm (Auger & Lawrence, 1989) to optimize our objective functions $\mathcal{C}_n(T_K)$. The algorithm is listed in Algorithm 3, which is deferred to Appendix A.5 due to space limitations. To implement the dynamic programming in Algorithm 3, we need to compute the cost matrices based on **(H1)** and **(H2)**, which are listed in Algorithms 1–2.

Given the pairwise distance matrix $\mathbf{D}=(d(X_i,X_j))$, we need to compute the aggregated matrix M in Algorithm 1 and the matrix A in Algorithm 2, both of which require $\mathcal{O}(n^3)$ operations. Prefix sums and incremental endpoint updates reuse these matrices to evaluate all segment costs $C(a,b)$, $1\le a<b\le n$, in $\mathcal{O}(n^2)$ operations. Segment additivity permits global optimization by dynamic

programming: given $C(a, b)$ and $K_{\max}$ (the maximal number of change points), Algorithm 3 computes an optimal segmentation for each $K \in \{0, \ldots, K_{\max}\}$ in $\mathcal{O}(n^2(K_{\max}+1))$ operations, giving a total complexity of $\mathcal{O}(n^3 + n^2(K_{\max} + 1))$.

A natural method to evaluate each transition cost $C(s+1, t)$ from the aggregated matrix takes $\mathcal{O}(L^2)$ operations for $L = t - s$. This yields a complexity of $\mathcal{O}(n^4(K_{\max} + 1))$. Rebuilding M or A for each evaluation gives a total complexity of $\mathcal{O}(n^5(K_{\max} + 1))$. Thus, Algorithm 3 is faster than the natural method.

3 Theoretical Analysis

DPDCP uses the same observations to construct the distance profile kernel and select the segmentation. Even under independence of the original observations, the empirical weights depend on the fitted data. Our oracle inequality accounts for this dependence through conditional scalar exponential moments, retaining every empirical anchor and radius under **(H1)** and both random radius endpoints under **(H2)**. Appendix A.1.2 links concentration bounds for the original slope regression coefficients computed in Algorithm 3 to this risk guarantee; see Baudry et al. (2012, Section 4) for background on penalty calibration.

3.1 Hilbert Space Formulation and Least Squares Segmentation

For (Ω, d) and the empirical measure H of Section 2, let $\mathcal{T} = [0, +\infty)$ under **(H1)** and $\mathcal{T} = [t_{\min}, t_{\max}]$ under **(H2)**, and define $\mathcal{H} = L^2(\Omega \times \mathcal{T}, H)$. All norms and inner products use the realized H; all **(H2)** quantities are evaluated on $t_{\max} > t_{\min}$. Define $\Phi : \Omega \to \mathcal{H}$ by $\Phi(x)(w, t) = \mathbb{I}\{d(w, x) \le t\}$ for $(w, t) \in \Omega \times \mathcal{T}$. We retain $F_k^\star$ and $\varepsilon_i(w, t)$ from (2), writing $K^\star$ for the true number of changes as in Appendix A.2.1, and K for a candidate number.

Assumption 1 *The true boundaries in (2) are deterministic integer indices. The observations $X_1, \ldots, X_n$, $n \ge 2$, are independent Borel measurable random elements of the separable metric space (Ω, d), with identical distributions within each true segment $(t_k^\star, t_{k+1}^\star]$, $k = 0, \ldots, K^\star$.*

Remark 1 *Independence permits conditioning on the observations defining anchors and radii, leaving the remaining observations independent with segmentwise profile means $F_k^\star$. Candidate intervals may cross true changes, and $K^\star = 0$ gives the globally iid model. Separability and Borel measurability ensure joint measurability, while bounded indicators and unit mass of H give finite integrals on unbounded spaces, including distributions with atoms and tied distances. The **(H2)** event is defined on the whole sample space and evaluated on the positive span domain. Appendix A.1 develops symmetrization (van der Vaart & Wellner, 1996, Section 2.3) and bounded differences (McDiarmid, 1989) arguments.*

Set $Y_i = \Phi(X_i)$, and let $\mu_i^\star \in \mathcal{H}$ be represented by the population profile $F_k^\star$ for $t_k^\star < i \le t_{k+1}^\star$, defined at deterministic (w, t) by the expectation in (2). Write $Y = (Y_i)_{i=1}^n$, $\mu^\star = (\mu_i^\star)_{i=1}^n$, and $\varepsilon = (\varepsilon_i)_{i=1}^n = Y - \mu^\star$. For $u, v \in \mathcal{H}^n$, use $\langle u, v\rangle_{\mathcal{H}^n} = \sum_{i=1}^n \int u_i(w, t)v_i(w, t)\, H(dw, dt)$ and $\|u\|_{\mathcal{H}^n}^2 = \langle u, u\rangle_{\mathcal{H}^n}$, separately at each realization.

For T_K of Section 2 with integer boundaries, let $S_k = \{t_k + 1, \ldots, t_{k+1}\}$ and $n_k = t_{k+1} - t_k$. The subspace $\mathcal{F}_{T_K} \subset \mathcal{H}^n$ consists of vectors constant on each S_k, and its orthogonal projection satisfies $(\Pi_{T_K} u)_i = n_k^{-1} \sum_{j \in S_k} u_j$ for $i \in S_k$. Thus $\widehat{\mu}_{T_K} = \Pi_{T_K} Y$ is the segmentwise empirical profile $\widehat{F}_{(t_k, t_{k+1}]}$ of Section 2, and

$$\widehat{\mathcal{R}}_n(T_K) = \frac{1}{n}\|Y - \widehat{\mu}_{T_K}\|_{\mathcal{H}^n}^2 = L_n(T_K), \qquad \mathcal{R}(u) = \frac{1}{n}\|u - \mu^\star\|_{\mathcal{H}^n}^2. \tag{4}$$

The risk is the random integrated squared error at the method's empirical anchors and radii. Let $\mathcal{P}$ be any nonempty deterministic collection of candidate configurations, including the full collection. Algorithm 3 uses all configurations with $n_k \ge \ell_{\min}$ and $K \le K_{\max}$ after its feasibility update. All minima are over finite sets, with a fixed ordering resolving ties measurably.

3.2 Distance Profile Kernels and the Kernel Least Squares Form

We use the aggregated matrices M, A, segment cost matrices $C^{\mathrm{H1}}, C^{\mathrm{H2}}$, and scaling constants from Algorithms 1 and 2.

Theorem 1 *Under **(H1)**, for $x, y \in \Omega$,*

$$k_{\mathrm{H1}}(x,y) = \langle \Phi(x), \Phi(y) \rangle_{\mathcal{H}} = \frac{1}{n^2} \sum_{m=1}^{n} \min\{w_m(x), w_m(y)\},$$

where $w_m(x) = \sum_{r=1}^{n} \mathbb{I}\{d(X_m, x) \le d(X_m, X_r)\}$ extends the reverse rank count of Algorithm 1, with $w_m(X_i) = w_m(i)$. The kernel is symmetric, positive semidefinite, and takes values in $[0,1]$. For every configuration with finite algorithmic costs,

$$\begin{aligned} L_n(T_K) &= \frac{1}{n} \sum_{i=1}^{n} k_{\mathrm{H1}}(X_i, X_i) - \frac{1}{n} \sum_{k=0}^{K} \frac{1}{n_k} \sum_{i,j \in S_k} k_{\mathrm{H1}}(X_i, X_j) \\ &= \frac{1}{n^3} \sum_{k=0}^{K} C^{\mathrm{H1}}(t_k + 1, t_{k+1}). \end{aligned} \tag{5}$$

Remark 2 *The minimum of the reverse rank counts measures the intersection of nested radius sets, including tied radii. At the observations, $k_{\mathrm{H1}}(X_i, X_j) = M_{ij}/n^2$, so (5) connects the empirical rank geometry to the computed costs at scale n^{-3}.*

Theorem 2 *Under **(H2)**, for $x, y \in \Omega$ on $t_{\max} > t_{\min}$,*

$$k_{\mathrm{H2}}(x,y) = \langle \Phi(x), \Phi(y) \rangle_{\mathcal{H}} = \frac{\sum_{m=1}^{n} \big(t_{\max} - \max\{t_{\min}, d(X_m, x), d(X_m, y)\}\big)_+}{n(t_{\max} - t_{\min})},$$

where $z_+ = \max\{z, 0\}$ for $z \in \mathbb{R}$. The kernel is symmetric, positive semidefinite, and takes values in $[0,1]$. For every configuration with finite algorithmic costs,

$$\begin{aligned} L_n(T_K) &= \frac{1}{n} \sum_{i=1}^{n} k_{\mathrm{H2}}(X_i, X_i) - \frac{1}{n} \sum_{k=0}^{K} \frac{1}{n_k} \sum_{i,j \in S_k} k_{\mathrm{H2}}(X_i, X_j) \\ &= \kappa_{\mathrm{H2}} \left\{ \sum_{k=0}^{K} C^{\mathrm{H2}}(t_k + 1, t_{k+1}) - t_{\min}(n - K - 1) \right\}. \end{aligned} \tag{6}$$

Remark 3 *The aggregated matrix A retains the absolute distance differences, while (6) accounts for both random endpoints and zero diagonal distances. For fixed K, its minimizing configurations equal those of the algorithmic cost; the segment count correction determines the effective penalty for selecting K.*

3.3 Oracle Inequality for the Distance Profile Kernel Least Squares Criterion

Let $\mathrm{pen}(T_K)$ be the penalty attached to $\kappa \sum_{k=0}^{K} C(t_k + 1, t_{k+1})$, using C, κ from the cost algorithm and $\mathrm{pen}(K)$ in Algorithm 3. Define

$$\mathfrak{p}(T_K) = \begin{cases} \mathrm{pen}(T_K), & \text{under (H1)}, \\ \mathrm{pen}(T_K) - \kappa_{\mathrm{H2}} t_{\min}(K+1), & \text{under (H2)}. \end{cases} \tag{7}$$

Let $\widehat{T}_{\widehat{K}}$ minimize $L_n(T_K) + \mathfrak{p}(T_K)$ over $\mathcal{P}$. For Algorithm 3 and its feasible collection, this criterion equals the normalized penalized algorithmic cost under **(H1)** and differs from it by the common additive constant $\kappa_{\mathrm{H2}} t_{\min} n$ under **(H2)**. Thus the same tie rule selects identical change counts and locations, as proved through the recursion and backtracking in Appendix A.1. The penalty may be

any finite measurable function of the data. Define

$$\mathfrak{a} = \begin{cases} 5/8, & \text{under } \textbf{(H1)}, \\ 2, & \text{under } \textbf{(H2)}, \end{cases}$$

$$\mathfrak{c}(T_K) = \begin{cases} \dfrac{5}{8}\log(\sqrt{5}+e^{8/5})(K+1) + \dfrac{5}{8}\log\binom{n-1}{K}, & \text{under } \textbf{(H1)}, \\ 2\log\left(\dfrac{2}{\sqrt{3}}e^{10/3} + \sqrt[4]{\dfrac{4}{3}}\,e^{23/12}\right)(K+1) + 2\log\binom{n-1}{K}, & \text{under } \textbf{(H2)}. \end{cases} \tag{8}$$

Here $\log$ is natural and $\binom{n-1}{K}$ counts boundary placements. Appendix A.1 derives these coefficients from the full sample distance profile processes.

Theorem 3 *Under Assumption 1, fix either weighting scheme and $y > 0$. On an event of probability at least $1 - e^{-y}$, common to all finite measurable penalties and all collections in Section 3.1, the following holds, with **(H2)** evaluated on $t_{\max} > t_{\min}$. For each penalty, set $\mathfrak{b} = \max_{T_K \in \mathcal{P}}\{4\mathfrak{c}(T_K)/n - \mathfrak{p}(T_K)\}_+$. Then*

$$\mathcal{R}(\widehat{\mu}_{\widehat{T}_{\widehat{K}}}) \le \inf_{T_K \in \mathcal{P}} \left\{ 3\mathcal{R}(\widehat{\mu}_{T_K}) + \frac{3}{2}\mathfrak{p}(T_K) + \frac{6\mathfrak{c}(T_K)}{n} \right\} + \frac{3}{2}\mathfrak{b} + \frac{6\mathfrak{a}y}{n}. \tag{9}$$

In particular, whenever

$$\mathfrak{p}(T_K) \ge \frac{4\mathfrak{c}(T_K)}{n}, \qquad T_K \in \mathcal{P}, \tag{10}$$

the selected configuration satisfies

$$\mathcal{R}(\widehat{\mu}_{\widehat{T}_{\widehat{K}}}) \le 3 \inf_{T_K \in \mathcal{P}} \{\mathcal{R}(\widehat{\mu}_{T_K}) + \mathfrak{p}(T_K)\} + \frac{6\mathfrak{a}y}{n}. \tag{11}$$

Remark 4 *Theorem 3 connects independence of the original observations to risk control in the random empirical geometry shared by kernel construction, profile estimation, and segmentation. Conditioning on anchor and radius observations under **(H1)**, and using a conditional threshold supremum under **(H2)**, yields scalar exponential bounds; averaging over the empirical weights gives the common event for candidate pairs and data dependent penalties. The oracle factor is 3, with remainders $15y/(4n)$ and $12y/n$, respectively. Under **(H2)**, the event within its domain has probability at least $\mathbb{P}(t_{\max} > t_{\min}) - e^{-y}$, hence $1 - e^{-y}$ when the domain has probability one. Replacing y by $y + \log 2$ gives a common event of probability at least $1 - e^{-y}$, with each scheme evaluated on its domain.*

*The term $\mathfrak{b}$ covers the realized coefficients of Algorithm 3. Sufficient conditions for (10) are $\widehat{c}_1 \ge 5/2$, $\widehat{c}_2 \ge (5/2)\log(\sqrt{5}+e^{8/5})$ under **(H1)**, and*

$$\widehat{c}_1 \ge 8, \qquad \widehat{c}_2 \ge 8\log\left(\frac{2}{\sqrt{3}}e^{10/3} + \sqrt[4]{\frac{4}{3}}\,e^{23/12}\right) + \frac{t_{\min}}{n(t_{\max} - t_{\min})}$$

*under **(H2)**. The corresponding segment coefficients are approximately 4.9314 and 29.4454, obtained from the conditional moments and model sums.*

4 Experiments

This section compares our DPDCP procedure under the **(H1)** and **(H2)** formulations with the MCPD method of Dubey & Zheng (2026, Algorithm 1) and the KCP method of Arlot et al. (2019) across five representative non-Euclidean data-generating mechanisms, including density-valued data, network data, symmetric positive definite (SPD) matrix data, and spherical data with and without temporal dependence. For brevity, the spherical-data experiments are reported in the Appendix A.3. The detailed data-generation procedures, together with the settings of tuning parameters, are provided in Appendix A.2.

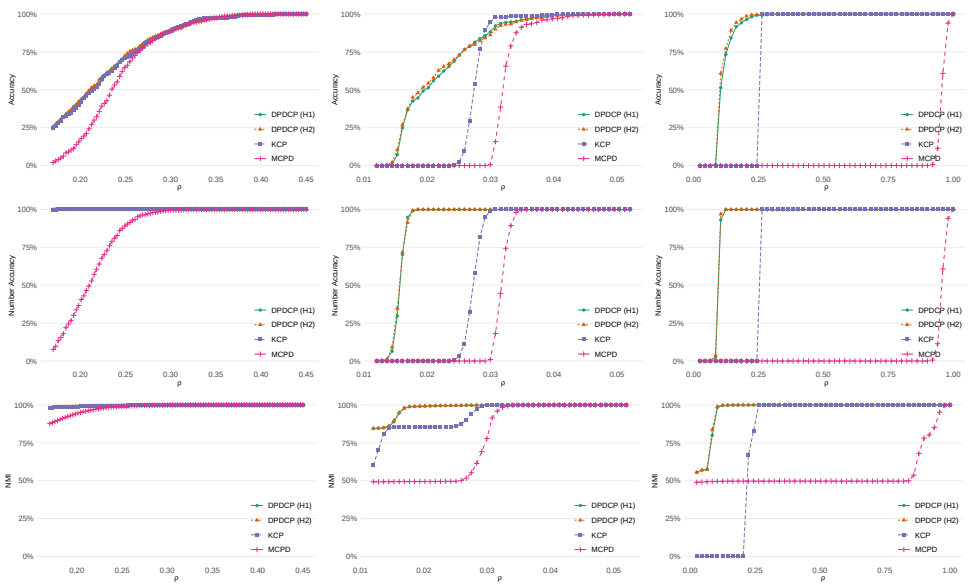

Figure 1: Percentage of correctly estimating both the number and locations of change-points (upper panels), accuracy of estimating the number of change points (middle panels), and NMI of change points detection (lower panels), as a function of ρ for different methods for the simulated density-valued, network and symmetric positive definite (SPD) matrix data (from left to right).

4.1 Simulation Results

Figure 1 compares the change-point estimation accuracy of the procedures across three data-generating mechanisms as ρ increases (with larger ρ corresponding to a higher signal to noise ratio), with upper panels reporting the proportion of replications in which both the number and locations of change points are correctly estimated, middle panels showing the proportion of replications in which the number of change points is correctly estimated, lower panels depicting the normalized mutual information (NMI) between the estimated and true segmentations. Across all four data-generating mechanisms, our DPDCP method, under both the **(H1)** and **(H2)** formulations, is more robust than KCP and MCPD in the low-SNR regime.

Table 1: Average runtime per run (in seconds) for different methods across the simulated density-valued data, network data, symmetric positive definite (SPD) matrix data, and spherical data without temporal dependence.

Scenario	DPDCP(**H1**)	DPDCP(**H2**)	KCP	MDCP
Density	0.0736	0.0540	0.0294	10.8820
Network	0.0692	0.0601	0.0279	9.9129
SPD	0.0893	0.0769	0.0375	10.5639
Spherical	0.0750	0.0647	0.0286	9.3232

Table 1 reports the average runtime of the competing procedures across the four scenarios, indicating that DPDCP achieves favorable computational performance. KCP achieves the shortest runtime, which is consistent with its dynamic programming complexity of $\mathcal{O}\big(n^2(K_{\max}+1)\big)$ when the kernel matrix is given. In practice, additional bandwidth selection strategies for KCP, such as searching over multiple candidate bandwidths or allowing observation-dependent bandwidths, can introduce substantial extra computational cost. By contrast, our DPDCP procedure does not require such a tuning stage. In addition, Section A.3 of the Appendix provides an additional simulation study based on a data generating mechanism with weak temporal dependence. The results show that DPDCP continues to perform well under this weak dependence structure. We also report the running

times for sample sizes $n = 400$, $n = 2000$, and $n = 5000$, which further illustrates the empirical computational behavior of the proposed method as the sample size increases.

5 Real Data Analysis

We demonstrate the practical effectiveness and robustness of the DPDCP procedure on three real-world datasets, which represent distinct non-Euclidean data, including network-valued observations (Section 5.1), SPD matrix sequences (Section 5.2), and compositional data (Appendix A.4.2).

5.1 MIT Reality Mining Proximity Networks

The MIT Reality Mining dataset is a longitudinal study of human interactions collected by the MIT Media Laboratory (Eagle & Pentland, 2006), available in the R package `GreedySBTM`[1]. The data consists of Bluetooth proximity networks where nodes represent individuals and an undirected edge indicates at least one physical encounter within a given time interval and contains 1392 networks observed at a 4-hour resolution from 2004-09-14 to 2005-05-05, each with $N = 96$ nodes. Aggregating the 4-hour networks within each day yields a sequence of $n = 232$ daily networks $\{G_t\}_{t=1}^{n}$.

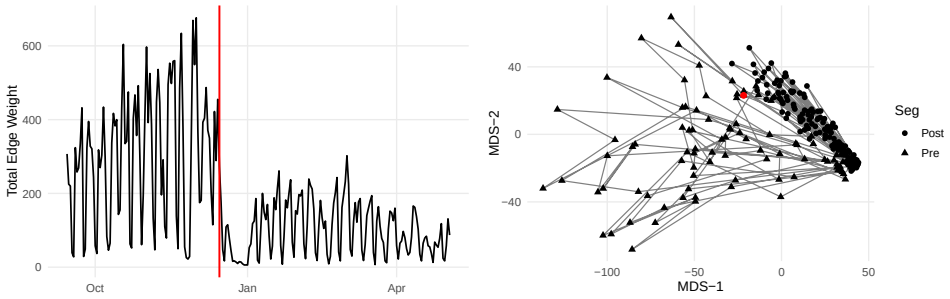

Figure 2: Temporal dynamics and geometric structure of the MIT Reality Mining daily proximity networks. The top panel depicts the network intensity over time, with the estimated change point at day 93 (2004-12-15) marked by a vertical line. The bottom panel shows an MDS embedding of networks, where each point represents one day and the change point is marked by the red dot.

Let $A_t \in \mathbb{R}^{N\times N}$ denote the adjacency matrix of G_t. To mitigate degree heterogeneity, each network is represented by its graph Laplacian $L_t = D_t - A_t$, where $D_t = \mathrm{diag}(A_t\mathbf{1})$, and dissimilarity between daily networks is measured by the Frobenius distance $d(G_s, G_t) = \|L_s - L_t\|_F$, yielding a pairwise distance matrix Δ used as input to the change-point procedure. The method identifies a pronounced structural change at day 93 (2004-12-15), which coincides with the final examination period at MIT and precedes the winter break, suggesting a plausible behavioral explanation for the observed shift in interaction patterns.

The top plot of Figure 2 shows the daily network intensity together with the detected change point. A pronounced shift around 2004-12-15 indicates a change in overall proximity activity, consistent with reduced on-campus interactions during the transition into the winter break. To further validate that the change is not driven solely by a marginal summary, the bottom plot displays a low-dimensional embedding of the daily networks obtained via classical multidimensional scaling based on the distance matrix Δ. The resulting trajectory exhibits a clear geometric turning behavior near the detected date, with networks before and after the change occupying distinct regions of the embedded space, indicating a genuine structural transition.

5.2 U.S. Stock Market Correlation Matrix

We next consider a financial application in which each observation is a matrix-valued object. Daily closing prices were obtained from the publicly accessible Stooq database.[2] Specifically, we collect

[1] `https://github.com/cran/GreedySBTM/blob/master/data/reality_mining_1392.RData`

[2] `https://stooq.com/`

daily closing prices for a fixed set of $p = 15$ U.S. equities and exchange-traded funds over the period from 2014-03-27 to 2026-01-22. Let $r_t \in \mathbb{R}^p$ denote the vector of daily log returns on trading day t. To summarize the evolving dependence structure of the market, we construct a sequence of rolling sample correlation matrices. Using a rolling window of length $w = 60$ trading days with consecutive windows advancing by step $= 10$, we yield a sequence $\{\widehat{R}_k\}_{k=1}^{292}$ of SPD matrices by computing the sample correlation matrix $\widehat{R}_k \in \mathbb{R}^{p\times p}$, which naturally reside on the SPD manifold rather than in a Euclidean vector space. To respect the non-Euclidean geometry of SPD matrices, we measure dissimilarity between two rolling correlation matrices $\widehat{R}_i$ and $\widehat{R}_j$ using a Cholesky-based metric $d(\widehat{R}_i, \widehat{R}_j) = \|\mathrm{chol}(\widehat{R}_i) - \mathrm{chol}(\widehat{R}_j)\|_F$, where $\mathrm{chol}(\cdot)$ denotes the upper-triangular Cholesky factor and $\|\cdot\|_F$ is the Frobenius norm. The resulting pairwise distance matrix $\Delta = (d(\widehat{R}_i, \widehat{R}_j))_{i,j=1}^n$ is used as the input for our distance-based multiple change-point procedure.

Applying the proposed DPDCP to Δ identifies three change points, corresponding to the calendar dates 2016-11-07, 2018-02-01, and 2024-01-19. These dates partition the rolling correlation sequence into four regimes with distinct dependence structures. Figure 3 complements this geometric view by displaying the mean correlation matrix within each estimated regime, highlighting systematic differences in overall correlation levels and block patterns across segments. Additional geometric visualization is provided in Figure 7 in the Appendix.

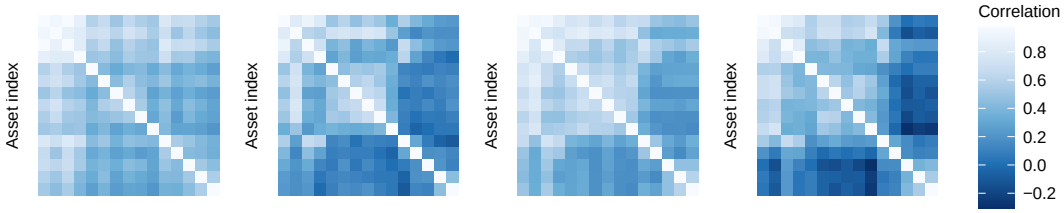

Figure 3: Mean rolling correlation matrices of U.S. equity returns within the regimes identified by the proposed method. Each panel reports the average correlation matrix over one segment, revealing clear shifts in overall dependence strength and cross-asset correlation patterns across market regimes.

From a market perspective, the change around early 2018 is consistent with a well-known volatility regime shift, during which dependence structures across assets changed markedly. The change around late 2016 is also compatible with a transition in macro-financial conditions, reflected here by a sustained reconfiguration of cross-asset correlations. The final change in early 2024 appears less tied to a single event but is supported by the emergence of a visually distinct cluster in the MDS embedding and a corresponding shift in the average correlation structure, suggesting a transition to a more heterogeneous dependence regime.

6 Conclusion

In this paper, we proposed DPDCP, a distance profile based framework for multiple change point detection for random objects in general metric spaces. The method operates directly on pairwise distances and is applicable once a meaningful metric on the object space is available. By combining distance profile losses with dynamic programming and penalized model selection, DPDCP estimates both the number and locations of multiple change points while optimizing the global segmentation structure. The proposed criterion targets general distributional changes, avoids reliance on an ambient vector space structure. Empirical results on simulations and real data demonstrate accurate localization and broad applicability across diverse non Euclidean data, while the oracle inequality provides finite sample theoretical support for the proposed criterion. In addition, further discussion of limitations and future work is provided in Section A.7 of the Appendix.

AI USE STATEMENT

In this work, we used generative AI tools for language polishing, improving clarity and expression, and checking parts of the mathematical proofs for possible inconsistencies or gaps. We have not used generative AI tools for generating the main research ideas, experimental results, or mathematical conclusions, and the other required disclosure tasks are not applicable to this work. Additionally, we used generative AI tools for minor editorial suggestions. We have reviewed all AI-assisted work. In particular, all mathematical arguments and proofs were independently checked and verified by the authors, and all AI-assisted language edits were reviewed before inclusion in the manuscript. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Sylvain Arlot, Alain Celisse, and Zaid Harchaoui. A kernel multiple change-point algorithm via model selection. *Journal of machine learning research*, 20(162):1–56, 2019.

Ivan E Auger and Charles E Lawrence. Algorithms for the optimal identification of segment neighborhoods. *Bulletin of mathematical biology*, 51(1):39–54, 1989.

Jean-Patrick Baudry, Cathy Maugis, and Bertrand Michel. Slope heuristics: overview and implementation. *Statistics and Computing*, 22(2):455–470, 2012. doi: 10.1007/s11222-011-9236-1.

Aurélien Bellet, Amaury Habrard, and Marc Sebban. A survey on metric learning for feature vectors and structured data. *arXiv preprint arXiv:1306.6709*, 2013.

Lucien Birgé and Pascal Massart. Minimal penalties for Gaussian model selection. *Probability Theory and Related Fields*, 138(1–2):33–73, 2007. doi: 10.1007/s00440-006-0011-8.

Shubhadeep Chakraborty and Xianyang Zhang. High-dimensional change-point detection using generalized homogeneity metrics. *arXiv preprint arXiv:2105.08976*, 2021.

Wei-Cheng Chang, Chun-Liang Li, Yiming Yang, and Barnabás Póczos. Kernel change-point detection with auxiliary deep generative models. *arXiv preprint arXiv:1901.06077*, 2019.

Hao Chen and Lynna Chu. Graph-based change-point analysis. *Annual Review of Statistics and Its Application*, 10(1):475–499, 2023.

Hao Chen and Nancy Zhang. Graph-based change-point detection. *The Annals of Statistics*, 43(1): 139–176, 2015.

Tianyi Chen, Youngser Park, Ali Saad-Eldin, Zachary Lubberts, Avanti Athreya, Benjamin D Pedigo, Joshua T Vogelstein, Francesca Puppo, Gabriel A Silva, Alysson R Muotri, et al. Discovering a change point and piecewise linear structure in a time series of organoid networks via the iso-mirror. *Applied Network Science*, 8(1):45, 2023.

Paromita Dubey and Minxing Zheng. Change point inference for non-euclidean data sequences using distance profiles, 2026. URL https://arxiv.org/abs/2311.16025.

Paromita Dubey, Yaqing Chen, and Hans-Georg Müller. Metric statistics: Exploration and inference for random objects with distance profiles. *The Annals of Statistics*, 52(2):757–792, 2024.

Nathan Eagle and Alex Pentland. Reality mining: sensing complex social systems. *Personal and ubiquitous computing*, 10(4):255–268, 2006.

Aasa Feragen, Francois Lauze, and Soren Hauberg. Geodesic exponential kernels: When curvature and linearity conflict. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3032–3042, 2015.

Damien Garreau and Sylvain Arlot. Consistent change-point detection with kernels. *Electronic Journal of Statistics*, 12:4440–4486, 2018.

Gregory B Gloor, Jia Rong Wu, Vera Pawlowsky-Glahn, and Juan José Egozcue. It's all relative: analyzing microbiome data as compositions. *Annals of epidemiology*, 26(5):322–329, 2016.

Zaid Harchaoui and Olivier Cappé. Retrospective mutiple change-point estimation with kernels. In *2007 IEEE/SP 14th Workshop on Statistical Signal Processing*, pp. 768–772. IEEE, 2007.

Zaid Harchaoui, Eric Moulines, and Francis Bach. Kernel change-point analysis. *Advances in neural information processing systems*, 21, 2008.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963. doi: 10.1080/01621459.1963.10500830.

Lajos Horváth, Piotr Kokoszka, and Shixuan Wang. Monitoring for a change point in a sequence of distributions. *The Annals of Statistics*, 49(4):2271–2291, 2021.

Seok-Oh Jeong, Chongwon Pae, and Hae-Jeong Park. Connectivity-based change point detection for large-size functional networks. *Neuroimage*, 143:353–363, 2016.

Brian Kulis. Metric learning: A survey. *Foundations and Trends® in Machine Learning*, 5(4): 287–364, 2013.

Marc Lavielle and Eric Moulines. Least-squares estimation of an unknown number of shifts in a time series. *Journal of time series analysis*, 21(1):33–59, 2000.

S. Letzgus. Change-point detection in wind turbine scada data for robust condition monitoring with normal behaviour models. *Wind Energy Science*, 5(4):1375–1397, 2020. doi: 10.5194/wes-5-1375-2020. URL `https://wes.copernicus.org/articles/5/1375/2020/`.

Jun Li. Asymptotic distribution-free change-point detection based on interpoint distances for high-dimensional data. *Journal of Nonparametric Statistics*, 32(1):157–184, 2020.

Shuang Li, Yao Xie, Hanjun Dai, and Le Song. M-statistic for kernel change-point detection. *Advances in Neural Information Processing Systems*, 28, 2015.

Marcos Matabuena, Alexander Petersen, Juan C Vidal, and Francisco Gude. Glucodensities: A new representation of glucose profiles using distributional data analysis. *Statistical methods in medical research*, 30(6):1445–1464, 2021.

David S Matteson and Nicholas A James. A nonparametric approach for multiple change point analysis of multivariate data. *Journal of the American Statistical Association*, 109(505):334–345, 2014.

Colin McDiarmid. On the method of bounded differences. In J. Siemons (ed.), *Surveys in Combinatorics, 1989*, volume 141 of *London Mathematical Society Lecture Note Series*, pp. 148–188. Cambridge University Press, 1989. doi: 10.1017/CBO9781107359949.008.

Colin McDiarmid. Concentration. In Michel Habib, Colin McDiarmid, Jorge Ramirez-Alfonsin, and Bruce Reed (eds.), *Probabilistic Methods for Algorithmic Discrete Mathematics*, volume 16 of *Algorithms and Combinatorics*, pp. 195–248. Springer, Berlin, Heidelberg, 1998. doi: 10.1007/978-3-662-12788-9_6.

Colin McDiarmid. Concentration for independent permutations. *Combinatorics, Probability and Computing*, 11(2):163–178, 2002. doi: 10.1017/S0963548301005089.

Lizhen Nie and Dan L Nicolae. Weighted-graph-based change point detection. *arXiv preprint arXiv:2103.02680*, 2021.

Yunlong Nie, LiangLiang Wang, and Jiguo Cao. Estimating time-varying directed gene regulation networks. *Biometrics*, 73(4):1231–1242, 2017.

Leto Peel and Aaron Clauset. Detecting change points in the large-scale structure of evolving networks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 29, 2015.

K. J. Prabuchandran, Nitin Singh, Pankaj Dayama, Ashutosh Agarwal, and Vinayaka Pandit. Change point detection for compositional multivariate data. *Applied Intelligence*, 52(2):1930–1955, 2022.

Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. *Advances in neural information processing systems*, 20, 2007.

Xiaoping Shi, Yuehua Wu, and Calyampudi Radhakrishna Rao. Consistent and powerful graph-based change-point test for high-dimensional data. *Proceedings of the National Academy of Sciences*, 114(15):3873–3878, 2017.

Olaf Sporns. Structure and function of complex brain networks. *Dialogues in clinical neuroscience*, 15(3):247–262, 2013.

Aad W. van der Vaart and Jon A. Wellner. *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer Series in Statistics. Springer, New York, 1996. doi: 10.1007/978-1-4757-2545-2.

Daren Wang, Yi Yu, and Alessandro Rinaldo. Optimal change point detection and localization in sparse dynamic networks. *The Annals of Statistics*, 49(1):203–232, 2021.

Heng Wang, Minh Tang, Youngser Park, and Carey E Priebe. Locality statistics for anomaly detection in time series of graphs. *IEEE Transactions on Signal Processing*, 62(3):703–717, 2014.

Kilian Q. Weinberger and Lawrence K. Saul. Distance metric learning for large margin nearest neighbor classification. *Journal of Machine Learning Research*, 10(9):207–244, 2009. URL `http://jmlr.org/papers/v10/weinberger09a.html`.

Eric Xing, Michael Jordan, Stuart J Russell, and Andrew Ng. Distance metric learning with application to clustering with side-information. *Advances in neural information processing systems*, 15, 2002.

Yi-Ching Yao. Estimating the number of change-points via schwarz'criterion. *Statistics & Probability Letters*, 6(3):181–189, 1988.

A Technical appendices and supplementary material

A.1 Proofs

A.1.1 Proofs of Theorems 1, 2, and 3

We first establish the projection and kernel identities, and then prove the probability bound used for the oracle comparison. All integrations use the empirical weights of Section 2.

For a realized sample and $u \in \mathcal{H}^n$, the coordinate formula for Π_{T_K} in Section 3.1 gives

$$\sum_{i \in S_k} \{u_i - (\Pi_{T_K} u)_i\} = \sum_{i \in S_k} u_i - n_k \left(\frac{1}{n_k} \sum_{i \in S_k} u_i \right) = 0.$$

For $v \in \mathcal{F}_{T_K}$, every coordinate on S_k equals v_{t_k+1}. This identity implies

$$\langle u - \Pi_{T_K} u, v \rangle_{\mathcal{H}^n} = \sum_{k=0}^{K} \left\langle \sum_{i \in S_k} \{u_i - (\Pi_{T_K} u)_i\}, v_{t_k+1} \right\rangle_{\mathcal{H}} = 0.$$

Furthermore, applying the averaging formula a second time leaves every coordinate unchanged. Thus $\Pi_{T_K} u \in \mathcal{F}_{T_K}$, the residual is orthogonal to that subspace, and Π_{T_K} is its orthogonal projection. The subspace is closed: convergence in the product norm implies convergence of every coordinate, so the equalities of coordinates within each segment pass to the limit. For $v \in \mathcal{F}_{T_K}$, expanding the squared norm and using the displayed orthogonality gives

$$\|u - v\|_{\mathcal{H}^n}^2 = \|u - \Pi_{T_K} u\|_{\mathcal{H}^n}^2 + \|\Pi_{T_K} u - v\|_{\mathcal{H}^n}^2.$$

This proves the least squares minimizing property stated in Section 3.1, including its application to $u = Y$.

In the following calculation, k denotes k_{H1} or k_{H2} from Theorems 1 and 2, respectively. Both have the inner product representation stated there. For a segment S_k, expanding each squared distance from its mean yields

$$\sum_{i\in S_k}\left\|Y_i-\frac{1}{n_k}\sum_{j\in S_k}Y_j\right\|_{\mathcal{H}}^2=\sum_{i\in S_k}\|Y_i\|_{\mathcal{H}}^2-\frac{2}{n_k}\sum_{i,j\in S_k}\langle Y_i,Y_j\rangle_{\mathcal{H}}+\frac{n_k}{n_k^2}\sum_{j,r\in S_k}\langle Y_j,Y_r\rangle_{\mathcal{H}}$$
$$=\sum_{i\in S_k}k(X_i,X_i)-\frac{1}{n_k}\sum_{i,j\in S_k}k(X_i,X_j).$$

Adding these equalities over the segments and dividing by n proves the kernel representation of the empirical least squares criterion in both kernel theorems.

The common indicator representation also proves the kernel properties. The product $\Phi(x)(w,t)\Phi(y)(w,t)$ is symmetric in x, y and belongs to $[0,1]$. Integration against the probability measure H therefore gives $k(x,y)=k(y,x)$ and $0\le k(x,y)\le 1$. For any finite list $x_1,\dots,x_s\in\Omega$ and real coefficients $a_1,\dots,a_s$,

$$\sum_{i,j=1}^{s}a_ia_jk(x_i,x_j)=\int\left\{\sum_{i=1}^{s}a_i\Phi(x_i)(w,t)\right\}^2H(dw,dt)\ge 0.$$

This proves positive semidefiniteness for each realized empirical measure H.

Proof of Theorem 1. The anchor measure μ and the empirical distribution function $\widehat{F}_w$ are those of **(H1)** in Section 2. For any bounded measurable function f on the anchor and radius domain, their definitions give the successive finite sum evaluations

$$\begin{aligned}\int f(w,t)\,H(dw,dt)&=\int_\Omega\left\{\int_{[0,\infty)}f(w,t)\,d\widehat{F}_w(t)\right\}\mu(dw)\\&=\frac{1}{n}\sum_{m=1}^{n}\int_{[0,\infty)}f(X_m,t)\,d\widehat{F}_{X_m}(t)\\&=\frac{1}{n^2}\sum_{m=1}^{n}\sum_{r=1}^{n}f\big(X_m,d(X_m,X_r)\big).\end{aligned}\tag{12}$$

The anchor remains the same in the function and in its associated radius distribution in every line. In particular, taking $f(w,t)=\Phi(x)(w,t)\Phi(y)(w,t)$ gives

$$k_{\mathrm{H1}}(x,y)=\frac{1}{n^2}\sum_{m=1}^{n}\sum_{r=1}^{n}\mathbb{I}\{d(X_m,x)\le d(X_m,X_r)\}\mathbb{I}\{d(X_m,y)\le d(X_m,X_r)\}.$$

For each fixed m, the set of indices r whose radius is at least $d(X_m,x)$ has cardinality $w_m(x)$ from Theorem 1. If $d(X_m,x)\ge d(X_m,y)$, that set is contained in the corresponding set for y, and the intersection has cardinality $w_m(x)$. If $d(X_m,x)\le d(X_m,y)$, the intersection instead has cardinality $w_m(y)$. When the two thresholds coincide, both sets coincide. In all three cases the inner sum in the preceding display equals $\min\{w_m(x),w_m(y)\}$. This proves the stated closed form, including tied radii.

We next evaluate the original loss itself. Applying (12) to its squared residual, and using the empirical profile from Section 2, yields

$$L_n(T_K)=\frac{1}{n^3}\sum_{m=1}^{n}\sum_{r=1}^{n}\sum_{k=0}^{K}\sum_{i\in S_k}\left\{\mathbb{I}\{d(X_m,X_i)\le d(X_m,X_r)\}-\frac{1}{n_k}\sum_{j\in S_k}\mathbb{I}\{d(X_m,X_j)\le d(X_m,X_r)\}\right\}^2.$$

For fixed m, r, k, expansion of this square, the identity $\mathbb{I}^2 = \mathbb{I}$, and summation over i give

$$\sum_{i \in S_k} \left\{ \mathbb{I}\{d(X_m, X_i) \le d(X_m, X_r)\} - \frac{1}{n_k} \sum_{j \in S_k} \mathbb{I}\{d(X_m, X_j) \le d(X_m, X_r)\} \right\}^2$$
$$= \sum_{i \in S_k} \mathbb{I}\{d(X_m, X_i) \le d(X_m, X_r)\}$$
$$- \frac{1}{n_k} \sum_{i,j \in S_k} \mathbb{I}\{d(X_m, X_i) \le d(X_m, X_r)\}\mathbb{I}\{d(X_m, X_j) \le d(X_m, X_r)\}.$$

The coefficient of the double sum is $-2/n_k + n_k/n_k^2 = -1/n_k$, which accounts for both the cross term and the square of the empirical mean. Summing over r transforms the single indicator into $w_m(i)$ from Algorithm 1, and transforms the product into $\min\{w_m(i), w_m(j)\}$. Summing over m then gives

$$L_n(T_K) = \frac{1}{n^3} \sum_{k=0}^{K} \left\{ \sum_{i \in S_k} M_{ii} - \frac{1}{n_k} \sum_{i,j \in S_k} M_{ij} \right\}.$$

Here M is the aggregated matrix of Algorithm 1, and the expression in braces is its cost $C^{\mathrm{H1}}(t_k + 1, t_{k+1})$ on a feasible segment. This proves (5).

For completeness, the prefix sums in that algorithm compute this same expression. Its variable s is $\sum_{i=a}^{b} M_{ii}$. Its running variable Q starts at M_{aa} and increases, when b is added, by $2\sum_{i=a}^{b-1} M_{ib} + M_{bb}$. The increase is the contribution of the new row and column to $\sum_{i,j=a}^{b} M_{ij}$. Thus at every endpoint b, its final assignment $s - Q/(b-a+1)$ equals the displayed segment cost. This establishes the correspondence with the computed quantity as well as the kernel identity.

Proof of Theorem 2. Under **(H2)**, consider a realization with $t_{\max} > t_{\min}$. The measure ν from Section 2 is uniform on the realized interval $[t_{\min}, t_{\max}]$. Consequently,

$$k_{\mathrm{H2}}(x, y) = \frac{1}{n(t_{\max} - t_{\min})} \sum_{m=1}^{n} \int_{t_{\min}}^{t_{\max}} \mathbb{I}\{d(X_m, x) \le t\}\mathbb{I}\{d(X_m, y) \le t\}\, dt.$$

The product of the two indicators equals one if and only if the integration variable is at least both distances. Within the integration interval, this set of radii has lower endpoint $\max\{t_{\min}, d(X_m, x), d(X_m, y)\}$ and upper endpoint $t_{\max}$ whenever the lower endpoint is at most the upper endpoint. Its length is therefore

$$\big(t_{\max} - \max\{t_{\min}, d(X_m, x), d(X_m, y)\}\big)_+.$$

If the lower endpoint exceeds the upper endpoint, the intersection is empty and its length is zero, which is also the value of this expression. Substitution proves the asserted kernel formula.

To derive the cost conversion, consider the distance entries $D_{m,i}$ already used in Algorithm 2. For $t \in [t_{\min}, t_{\max}]$,

$$\mathbb{I}\{D_{m,i} \le t\} = \mathbb{I}\{\max\{t_{\min}, D_{m,i}\} \le t\}.$$

Both $D_{m,i}$ and its truncated value are at most $t_{\max}$, including the diagonal value $D_{m,m} = 0$. The squared difference of two such threshold indicators is the indicator of the interval between the thresholds, including its lower endpoint and excluding its upper endpoint. It follows that

$$\int_{t_{\min}}^{t_{\max}} \{\mathbb{I}\{D_{m,i} \le t\} - \mathbb{I}\{D_{m,j} \le t\}\}^2\, dt = \big|\max\{t_{\min}, D_{m,i}\} - \max\{t_{\min}, D_{m,j}\}\big|. \tag{13}$$

This expression also covers equal thresholds, for which the interval has zero length.

We now compare the sum of these lengths to the entry A_{ij} of the aggregated matrix in Algorithm 2, for $i < j$. If $t_{\min} = 0$, every distance is already at least the lower endpoint, and summing (13) over m gives A_{ij}. If $t_{\min} > 0$, all off diagonal distances are at least $t_{\min}$. For m different from both i and j, both entries are off diagonal, and the summand equals $|D_{m,i} - D_{m,j}|$. For $m = i$, the original difference is $|0 - D_{i,j}| = D_{i,j}$, whereas the truncated difference is $|t_{\min} - D_{i,j}| = D_{i,j} - t_{\min}$.

For $m = j$, the same calculation subtracts a further $t_{\min}$. Combining all anchors, both cases give the single identity

$$\sum_{m=1}^{n} \int_{t_{\min}}^{t_{\max}} \{\mathbb{I}\{D_{m,i} \le t\} - \mathbb{I}\{D_{m,j} \le t\}\}^2 \, dt = A_{ij} - 2t_{\min}, \qquad i < j. \tag{14}$$

For real values $z_1, \dots, z_L$ with average $\bar{z} = L^{-1} \sum_i z_i$, direct expansion gives

$$\sum_{i=1}^{L} (z_i - \bar{z})^2 = \sum_{i=1}^{L} z_i^2 - \frac{1}{L} \left(\sum_{i=1}^{L} z_i \right)^2 = \frac{1}{L} \sum_{1 \le i < j \le L} (z_i - z_j)^2.$$

Indeed, expansion of the last pair sum gives $(L-1) \sum_i z_i^2 - 2 \sum_{i<j} z_i z_j = L \sum_i z_i^2 - (\sum_i z_i)^2$ before division by L. Apply this identity to the threshold indicators for each fixed anchor, radius, and segment S_k. The empirical profile in (3) is their average. Integrating and using (14) therefore yields

$$\begin{aligned} L_n(T_K) &= \frac{1}{n^2(t_{\max} - t_{\min})} \sum_{k=0}^{K} \frac{1}{n_k} \sum_{\substack{i,j \in S_k \\ i<j}} (A_{ij} - 2t_{\min}) \\ &= \kappa_{\mathrm{H2}} \sum_{k=0}^{K} \left\{ C^{\mathrm{H2}}(t_k + 1, t_{k+1}) - \frac{2t_{\min}}{n_k} \frac{n_k(n_k - 1)}{2} \right\} \\ &= \kappa_{\mathrm{H2}} \sum_{k=0}^{K} \{ C^{\mathrm{H2}}(t_k + 1, t_{k+1}) - t_{\min}(n_k - 1) \}. \end{aligned}$$

Since $\sum_{k=0}^{K} n_k = n$, summing $n_k - 1$ gives $n - (K+1)$, proving (6). The running sum R in Algorithm 2 increases by $\sum_{i=a}^{b-1} A_{ib}$ as a new endpoint b is added. It is therefore $\sum_{a \le i < j \le b} A_{ij}$, and division by $L = b - a + 1$ gives the same algorithmic C^{H2} used here.

The penalty conversion follows directly. In **(H1)**, the normalized algorithmic cost equals L_n. In **(H2)**, adding $\mathrm{pen}(T_K)$ to the normalized algorithmic cost gives

$$\begin{aligned} \kappa_{\mathrm{H2}} \sum_{k=0}^{K} C^{\mathrm{H2}}(t_k + 1, t_{k+1}) + \mathrm{pen}(T_K) &= L_n(T_K) + \mathrm{pen}(T_K) - \kappa_{\mathrm{H2}} t_{\min}(K+1) + \kappa_{\mathrm{H2}} t_{\min} n \\ &= L_n(T_K) + \mathfrak{p}(T_K) + \kappa_{\mathrm{H2}} t_{\min} n. \end{aligned}$$

The final term is common to every candidate configuration. This proves the equivalence claimed after (7), with the original costs and slope coefficients.

We detail the identification with the complete output of Algorithm 3, using its collection $\mathcal{P}$ from Section 3.1 and its original fitted penalty. For a prefix ending at t with $\ell_{\min} \le t \le n$, its entry $\mathrm{DP}(1, t) = C(1, t)$ is the cost of the unique one segment configuration. For $K \ge 1$ and $(K+1)\ell_{\min} \le t \le n$, suppose that $\mathrm{DP}(K, s)$ is the minimum original cost of a configuration with K segments ending at s. In any feasible configuration with $K+1$ segments ending at t, the final boundary s satisfies $K\ell_{\min} \le s \le t - \ell_{\min}$. Its first K segments have cost at least $\mathrm{DP}(K, s)$, and its last segment has cost $C(s+1, t)$. Conversely, each s in this range can be combined with a minimizing prefix configuration to obtain a feasible configuration ending at t. Thus the minimum is

$$\mathrm{DP}(K+1, t) = \min_{K\ell_{\min} \le s \le t - \ell_{\min}} \{\mathrm{DP}(K, s) + C(s+1, t)\},$$

which is the original recurrence. Induction proves its minimizing property for every computed entry. The corresponding entry of Prev records a last boundary attaining the minimum. At each recorded boundary s, the attained equality is $\mathrm{DP}(K+1, t) = \mathrm{DP}(K, s) + C(s+1, t)$. Replace t by this stored s and decrease the segment count by one, as in the algorithm. Repeating this equality decomposes the original value $\mathrm{DP}(K+1, n)$ into the costs of the recovered last segments and the remaining first segment. After K steps, the latter cost is the base entry $\mathrm{DP}(1, t) = C(1, t)$. Every stored boundary satisfies the displayed feasibility bounds, so all recovered segments have length at least $\ell_{\min}$. Their

total cost is therefore $\mathrm{DP}(K+1, n)$. For $K = 0$, the recovery consists of the unique segment containing all n observations.

The scaling constant κ is positive on the scheme's domain, so it preserves the ordering of these costs. Hence the algorithm's response is

$$\widehat{\mathcal{R}}(K+1) = \kappa \, \mathrm{dp}_{\mathrm{end}}(K+1) = \min_{T_K \in \mathcal{P}} \kappa \sum_{k=0}^{K} C(t_k + 1, t_{k+1}),$$

where this minimum ranges over configurations with K boundaries. Under **(H1)** the last objective equals $L_n(T_K)$. Under **(H2)** it equals $L_n(T_K) + \kappa_{\mathrm{H2}} t_{\min}(n - K - 1)$, by (6). The second term is common to every configuration with that fixed K. Therefore the configurations recovered by the original backtracking also minimize the integrated loss among all configurations with that K.

The original fitted penalty depends only on K. Adding it to the last display therefore gives

$$\mathrm{crit}(K+1) = \min_{T_K \in \mathcal{P}} \{L_n(T_K) + \mathfrak{p}(T_K)\} + \begin{cases} 0, & \text{under (H1)}, \\ \kappa_{\mathrm{H2}} t_{\min} n, & \text{under (H2)}. \end{cases}$$

Again, the minimum is over configurations with K boundaries. The added quantity is constant across K, so minimizing crit selects a minimum of the integrated objective over $\mathcal{P}$, using the coefficients fitted to the optimized algorithmic response above.

The boundary rule agrees as well. At each entry the original loop scans s in increasing order and updates the stored boundary only for a strict improvement. Hence it retains the smallest last boundary among all boundaries attaining that minimum. Applying the same statement recursively to its stored prefix gives the first minimizing boundary vector in the lexicographic order of $(t_K, t_{K-1}, \ldots, t_1)$: first choose the smallest last boundary, then the smallest penultimate boundary among these choices, and continue towards the first boundary. Use this same order for tied configurations at each fixed K, and the same minimizing K as the algorithm when several values of crit coincide. The theoretical selector then equals the algorithm's output at every sample in the scheme's domain.

We next establish the scalar exponential bounds used for the empirical distance profile weights. The calculation begins with the moments of a symmetric sign path and keeps the resulting constants throughout the probability argument.

Lemma 1 *Let $\xi_1, \ldots, \xi_s$ be independent nonnegative real random variables, where $s \geq 1$, and let $\mathcal{D}_j(t) = \mathbb{P}(\xi_j \leq t)$ denote their respective distribution functions. Put*

$$\mathcal{Z} = \sup_{t \geq 0} \left| \sum_{j=1}^{s} \{\mathbb{I}\{\xi_j \leq t\} - \mathcal{D}_j(t)\} \right|.$$

Then $\mathcal{Z}$ is measurable, and the following inequalities hold for every $\eta \in \mathbb{R}$ and every $0 < \vartheta < 2$:

$$\begin{aligned} \mathbb{E}\mathcal{Z} &\leq \sqrt{5s}, \\ \mathbb{E}e^{\eta(\mathcal{Z} - \mathbb{E}\mathcal{Z})} &\leq e^{s\eta^2/8}, \\ \mathbb{E}e^{\vartheta \mathcal{Z}^2/s} &\leq (1 - \vartheta/2)^{-1/2} \exp\left\{ \frac{5\vartheta}{1 - \vartheta/2} \right\}. \end{aligned} \tag{15}$$

In particular, for every $x \geq 0$,

$$\mathbb{P}\{\mathcal{Z} > \sqrt{5s} + \sqrt{sx/2}\} \leq e^{-x}.$$

Remark 5 *Lemma 1 treats independent distance variables through their individual distribution functions. It therefore applies to intervals that cross true segment boundaries. The symmetrization argument, as developed in van der Vaart & Wellner (1996, Section 2.3), is combined with a centered sign path whose moments produce the coefficient 5. The bounded differences argument of McDiarmid (1989) gives the exponential bound for every real parameter. A Gaussian integral then converts this bound directly into the squared exponential moment used below. The entire calculation is given here, including the elementary exponential estimate from Hoeffding (1963, Section 4).*

Proof. For each outcome, the indicators and their distribution functions are right continuous in t. Given any real $t \geq 0$, choose nonnegative rational numbers decreasing to t. The corresponding sums converge to the sum at t. Hence the supremum defining $\mathcal{Z}$ equals its supremum over the nonnegative rational numbers. This is a countable supremum of measurable random variables and is therefore measurable. The bound $0 \leq \mathcal{Z} \leq s$ also ensures that all the moments used below are finite.

Let $\chi_1, \ldots, \chi_s$ be independent of $\xi_1, \ldots, \xi_s$ and of one another, taking the values 1 and -1 with equal probabilities, and set $\mathcal{M}_0 = 0$. Define

$$\mathcal{S}_0 = 0, \qquad \mathcal{S}_j = \sum_{i=1}^{j} \chi_i, \qquad \mathcal{M}_j = \max_{0 \leq i \leq j} \mathcal{S}_i, \qquad 1 \leq j \leq s.$$

The inclusion of the initial value $\mathcal{S}_0 = 0$ gives $\mathcal{M}_j \geq 0$. We shall prove the identities

$$\mathbb{E}(\mathcal{M}_s^2 + \mathcal{M}_s) = s, \qquad \mathbb{E}(\mathcal{M}_s \mathcal{S}_s) = \frac{s}{2}.$$

For a positive integer h, every path ending at or above h has visited h. Consider a path that visits h and ends at a value $r < h$. Keep its increments up to its first visit to h and change the signs of all subsequent increments. The transformed path ends at $2h - r > h$. Conversely, every path ending above h visits h, and the same transformation produces a path ending below h whose maximum is at least h. Applying the transformation twice recovers the original path. Since every path of s signs has probability 2^{-s}, these two classes have the same probability. Separating the endpoints at or above h from those below h therefore gives

$$\mathbb{P}(\mathcal{M}_s \geq h) = \mathbb{P}(\mathcal{S}_s \geq h) + \mathbb{P}(\mathcal{S}_s \geq h + 1).$$

All random variables in this calculation take integer values. For a nonnegative integer r, the identity $r^2 + r = \sum_{h=1}^{r} 2h$ implies

$$\begin{aligned} \mathbb{E}(\mathcal{M}_s^2 + \mathcal{M}_s) &= \sum_{h=1}^{s} 2h \, \mathbb{P}(\mathcal{M}_s \geq h) \\ &= \mathbb{E}\left[\sum_{h=1}^{s} 2h \, \mathbb{I}\{\mathcal{S}_s \geq h\} + \sum_{h=1}^{s} 2h \, \mathbb{I}\{\mathcal{S}_s \geq h + 1\} \right]. \end{aligned}$$

If the endpoint is $r \geq 1$, the two sums in brackets are respectively $r(r+1)$ and $(r-1)r$, and their sum is $2r^2$. If the endpoint is at most zero, both sums vanish. Using the positive part notation of Theorem 2, this proves

$$\mathbb{E}(\mathcal{M}_s^2 + \mathcal{M}_s) = 2\mathbb{E}(\mathcal{S}_s)_+^2 = \mathbb{E}\mathcal{S}_s^2 = s.$$

The second equality follows from symmetry. The last follows from $\mathbb{E}\mathcal{S}_s^2 = \sum_{i=1}^{s} \mathbb{E}\chi_i^2 = s$, since the signs are independent and centered.

To evaluate the mixed moment, put $\mathfrak{h}_j = \mathcal{M}_j - \mathcal{M}_{j-1}$. The maximum increases by one if and only if $\mathcal{S}_{j-1} = \mathcal{M}_{j-1}$ and $\chi_j = 1$; otherwise its increment is zero. Expanding a product at consecutive times gives

$$\mathcal{M}_j \mathcal{S}_j - \mathcal{M}_{j-1} \mathcal{S}_{j-1} = \mathcal{M}_{j-1} \chi_j + \mathcal{S}_{j-1} \mathfrak{h}_j + \chi_j \mathfrak{h}_j.$$

The expectation of the first summand is zero because $\mathcal{M}_{j-1}$ is determined by the first $j-1$ signs and χ_j is independent of them with mean zero. The remaining two summands equal $(\mathcal{M}_{j-1} + 1)\mathfrak{h}_j$ on every path. Summing the product increments from $j = 1$ to s thus yields

$$\mathbb{E}(\mathcal{M}_s \mathcal{S}_s) = \mathbb{E} \sum_{j=1}^{s} (\mathcal{M}_{j-1} + 1)\mathfrak{h}_j.$$

On the other hand, $\mathfrak{h}_j \in \{0, 1\}$ gives

$$\begin{aligned} \mathcal{M}_j^2 - \mathcal{M}_{j-1}^2 &= 2\mathcal{M}_{j-1}\mathfrak{h}_j + (\mathfrak{h}_j)^2 \\ &= (2\mathcal{M}_{j-1} + 1)\mathfrak{h}_j. \end{aligned}$$

Adding the increments of $\mathcal{M}_j$ to these square increments and summing gives

$$\mathcal{M}_s^2 + \mathcal{M}_s = 2 \sum_{j=1}^{s} (\mathcal{M}_{j-1} + 1)\mathfrak{h}_j.$$

Taking expectations and using the already established identity proves

$$\mathbb{E}(\mathcal{M}_s\mathcal{S}_s) = \frac{1}{2}\mathbb{E}(\mathcal{M}_s^2 + \mathcal{M}_s) = \frac{s}{2}.$$

We now bound the centered maximum

$$\mathcal{B} = \max_{0\le j\le s} |\mathcal{S}_j - \mathcal{S}_s/2|.$$

Let $\mathcal{L} = \min_{0\le j\le s} \mathcal{S}_j$. The largest positive and negative deviations from $\mathcal{S}_s/2$ are

$$\mathcal{M}_s - \mathcal{S}_s/2, \qquad \mathcal{S}_s/2 - \mathcal{L},$$

respectively. Each is at least $|\mathcal{S}_s|/2$: both endpoints 0 and $\mathcal{S}_s$ belong to the list of partial sums. Their smaller squared value is therefore at least $\mathcal{S}_s^2/4$. The larger squared value is $\mathcal{B}^2$, so

$$\mathcal{B}^2 \le (\mathcal{M}_s - \mathcal{S}_s/2)^2 + (\mathcal{S}_s/2 - \mathcal{L})^2 - \mathcal{S}_s^2/4.$$

Changing every sign maps the pair $(\mathcal{S}_s, \mathcal{M}_s)$ to $(-\mathcal{S}_s, -\mathcal{L})$ and preserves its probability law. Thus the first two squared terms on the right have the same expectation. Expansion, followed by the identities above, gives

$$\begin{aligned}\mathbb{E}\mathcal{B}^2 &\le 2\mathbb{E}(\mathcal{M}_s - \mathcal{S}_s/2)^2 - \frac{1}{4}\mathbb{E}\mathcal{S}_s^2 = 2\mathbb{E}\mathcal{M}_s^2 - 2\mathbb{E}(\mathcal{M}_s\mathcal{S}_s) + \frac{s}{4} = 2\{s - \mathbb{E}\mathcal{M}_s\} - s + \frac{s}{4} \\ &= \frac{5s}{4} - 2\mathbb{E}\mathcal{M}_s \le \frac{5s}{4}.\end{aligned}$$

Cauchy Schwarz consequently yields $\mathbb{E}\mathcal{B} \le \sqrt{5s}/2$.

Let $\xi_1', \ldots, \xi_s'$ be independent copies with the same respective marginal distributions as $\xi_1, \ldots, \xi_s$, independent of the original variables and of the signs. Write $\mathbb{E}'$ for expectation over these copies while keeping the original variables fixed. For each threshold,

$$\sum_{j=1}^{s}\{\mathbb{I}\{\xi_j \le t\} - \mathcal{D}_j(t)\} = \mathbb{E}'\sum_{j=1}^{s}\{\mathbb{I}\{\xi_j \le t\} - \mathbb{I}\{\xi_j' \le t\}\}.$$

For each member of the countable threshold set, the absolute value of this expectation is at most the expectation of the supremum of the absolute values. Taking the supremum and then the expectation over the original sample consequently gives

$$\mathbb{E}\mathcal{Z} \le \mathbb{E}\sup_{t\ge 0}\left|\sum_{j=1}^{s}\{\mathbb{I}\{\xi_j \le t\} - \mathbb{I}\{\xi_j' \le t\}\}\right|.$$

Each pair (ξ_j, ξ_j') has an exchangeable distribution, and these pairs are mutually independent. Independently swapping the two entries of each pair with probability $1/2$ preserves their joint distribution and multiplies each difference by an independent sign. Therefore the last expected supremum equals

$$\mathbb{E}\sup_{t\ge 0}\left|\sum_{j=1}^{s}\chi_j\{\mathbb{I}\{\xi_j \le t\} - \mathbb{I}\{\xi_j' \le t\}\}\right|.$$

Subtracting $1/2$ from each of the two indicators leaves their difference unchanged. The triangle inequality, followed by equality of the primed and unprimed sample laws, yields

$$\begin{aligned}\mathbb{E}\mathcal{Z} &\le \mathbb{E}\sup_{t\ge 0}\left|\sum_{j=1}^{s}\chi_j\{\mathbb{I}\{\xi_j \le t\} - 1/2\}\right| + \mathbb{E}\sup_{t\ge 0}\left|\sum_{j=1}^{s}\chi_j\{\mathbb{I}\{\xi_j' \le t\} - 1/2\}\right| \\ &= 2\mathbb{E}\sup_{t\ge 0}\left|\sum_{j=1}^{s}\chi_j\{\mathbb{I}\{\xi_j \le t\} - 1/2\}\right|.\end{aligned}$$

Condition on $\xi_1, \ldots, \xi_s$ and order their indices by increasing observed value, resolving ties by index. At every threshold, the selected indices form a prefix in this ordering, with tied values entering

together. For a prefix of length j, the signed sum after subtracting $1/2$ from every indicator is the sum of its first j reordered signs minus one half of the sum of all s signs. The threshold supremum is bounded by the maximum of the absolute values of these centered partial sums over all $j = 0, \dots, s$. Conditional on the observations, the ordering is fixed and the reordered signs have the joint distribution of s independent symmetric signs. This maximum therefore has, conditionally on the observations, the distribution of $\mathcal{B}$ defined above. In particular,

$$\mathbb{E}\left[\sup_{t\geq 0}\left|\sum_{j=1}^{s}\chi_j\{\mathbb{I}\{\xi_j \leq t\} - 1/2\}\right| \middle| \xi_1, \dots, \xi_s\right] \leq \mathbb{E}\mathcal{B}.$$

Taking expectations over the observations and using the preceding symmetrization bound gives

$$\mathbb{E}\mathcal{Z} \leq 2\mathbb{E}\mathcal{B} \leq \sqrt{5s}.$$

We next prove the exponential estimate for both positive and negative parameters. Suppose that a centered real random variable ζ takes values in an interval $[\gamma, \gamma + 1]$, and define

$$\psi(\eta) = \log \mathbb{E}e^{\eta\zeta}, \qquad \eta \in \mathbb{R}.$$

Boundedness permits differentiation under the expectation. Let $\mathbb{E}_\eta$ denote expectation under the probability measure having density $e^{\eta\zeta}/\mathbb{E}e^{\eta\zeta}$ with respect to the original law. Differentiating gives

$$\psi'(\eta) = \mathbb{E}_\eta\zeta, \qquad \psi''(\eta) = \mathbb{E}_\eta\zeta^2 - (\mathbb{E}_\eta\zeta)^2.$$

Expansion of the square about the mean shows that the mean minimizes expected squared distance among real constants. Since the same interval contains ζ under the tilted measure, it follows that

$$0 \leq \psi''(\eta) \leq \mathbb{E}_\eta(\zeta - \gamma - 1/2)^2 \leq \frac{1}{4}, \qquad \eta \in \mathbb{R}.$$

We have $\psi(0) = 0$ and $\psi'(0) = \mathbb{E}\zeta = 0$. Applying the fundamental theorem of calculus twice along the segment from 0 to η gives the identity

$$\psi(\eta) = \eta^2 \int_0^1 (1-v)\psi''(v\eta)\, dv \leq \frac{\eta^2}{4}\int_0^1 (1-v)\, dv = \frac{\eta^2}{8}, \qquad \eta \in \mathbb{R}.$$

Thus $\mathbb{E}e^{\eta\zeta} \leq e^{\eta^2/8}$ for every real η. Rescaling proves that a centered variable whose support lies in an interval of length $b > 0$ satisfies the bound $e^{b^2\eta^2/8}$; a variable with zero interval length is zero and satisfies the same formula with $b = 0$.

Changing one coordinate ξ_j in $\mathcal{Z}$ changes each indexed sum by at most one. For any bounded functions f, g, the triangle inequality gives

$$\left|\sup_t |f(t)| - \sup_t |g(t)|\right| \leq \sup_t |f(t) - g(t)|.$$

The value of $\mathcal{Z}$ therefore changes by at most one under a change of one sample coordinate. Let $\mathcal{G}_j = \sigma(\xi_1, \dots, \xi_j)$ be the sigma field generated by the first j variables, and let $\mathcal{G}_0$ be the trivial sigma field. Define the martingale differences

$$\mathfrak{d}_j = \mathbb{E}[\mathcal{Z} \mid \mathcal{G}_j] - \mathbb{E}[\mathcal{Z} \mid \mathcal{G}_{j-1}], \qquad 1 \leq j \leq s.$$

Conditionally on $\mathcal{G}_{j-1}$, each difference has mean zero. Holding the first $j-1$ coordinates fixed, compare two values of ξ_j while keeping the remaining coordinates the same. The corresponding values of $\mathcal{Z}$ differ by at most one. Independence gives the same product of marginal laws for the remaining coordinates in both averages, so the conditional expectations also differ by at most one. Subtracting their conditional mean preserves this range length. Thus the conditional support of $\mathfrak{d}_j$ lies in an interval of length at most one, and the preceding exponential estimate gives

$$\mathbb{E}[e^{\eta\mathfrak{d}_j} \mid \mathcal{G}_{j-1}] \leq e^{\eta^2/8}, \qquad \eta \in \mathbb{R}.$$

The sum of the differences telescopes to $\mathcal{Z} - \mathbb{E}\mathcal{Z}$. Iterating conditional expectations gives

$$\mathbb{E}e^{\eta(\mathcal{Z}-\mathbb{E}\mathcal{Z})} = \mathbb{E}\left[e^{\eta\sum_{j=1}^{s-1}\mathfrak{d}_j}\mathbb{E}(e^{\eta\mathfrak{d}_s} \mid \mathcal{G}_{s-1})\right] \leq e^{\eta^2/8}\mathbb{E}e^{\eta\sum_{j=1}^{s-1}\mathfrak{d}_j} \leq e^{s\eta^2/8}, \qquad \eta \in \mathbb{R}.$$

For completeness, if $u > 0$ and $\eta > 0$, exponential Markov inequality gives

$$\mathbb{P}\{\mathcal{Z} - \mathbb{E}\mathcal{Z} > u\} \le \exp\{-\eta u + s\eta^2/8\}.$$

Differentiating the quadratic exponent in η shows that its minimum is attained at $\eta = 4u/s$, where its value is $-2u^2/s$. Substituting $u = \sqrt{sx/2}$ and using $\mathbb{E}\mathcal{Z} \le \sqrt{5s}$ proves the stated probability bound for $x > 0$. At $x = 0$ its right side equals one and the bound follows directly.

We now compute a squared exponential moment directly. The following calculation applies to any bounded scalar random variable $\mathfrak{V}$ satisfying

$$\mathbb{E}e^{\eta(\mathfrak{V}-\mathbb{E}\mathfrak{V})} \le e^{\eta^2/8}, \qquad \eta \in \mathbb{R}.$$

Let ω be a standard normal random variable independent of $\mathfrak{V}$. For a fixed real value v, completing the square in the normal density gives

$$\mathbb{E}_\omega e^{\sqrt{2\vartheta}\,\omega v} = \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} e^{-u^2/2+\sqrt{2\vartheta}\,uv}\, du = e^{\vartheta v^2},$$

where $\mathbb{E}_\omega$ denotes expectation with respect to ω. All integrands are nonnegative, so Tonelli's theorem permits interchanging the two expectations. The exponential bound for $\mathfrak{V}$ holds for every real parameter and hence applies at $\eta = \sqrt{2\vartheta}\,u$ for each real u. It follows that

$$\mathbb{E}e^{\vartheta\mathfrak{V}^2} = \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} e^{-u^2/2}\mathbb{E}e^{\sqrt{2\vartheta}\,u\mathfrak{V}}\, du \le \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} \exp\left\{-\frac{1-\vartheta/2}{2}u^2 + \sqrt{2\vartheta}\,u\mathbb{E}\mathfrak{V}\right\} du.$$

For $0 < \vartheta < 2$, the quadratic coefficient $1 - \vartheta/2$ is positive. The exponent inside this last integral equals

$$-\frac{1-\vartheta/2}{2}\left(u - \frac{\sqrt{2\vartheta}\,\mathbb{E}\mathfrak{V}}{1-\vartheta/2}\right)^2 + \frac{\vartheta(\mathbb{E}\mathfrak{V})^2}{1-\vartheta/2}.$$

Evaluating this Gaussian integral yields

$$\mathbb{E}e^{\vartheta\mathfrak{V}^2} \le (1-\vartheta/2)^{-1/2} \exp\left\{\frac{\vartheta(\mathbb{E}\mathfrak{V})^2}{1-\vartheta/2}\right\}. \tag{16}$$

Apply this calculation to $\mathfrak{V} = \mathcal{Z}/\sqrt{s}$. The centered exponential bound just proved gives its required exponential inequality, and its mean is at most $\sqrt{5}$. This proves the last assertion of (15).

Lemma 2 *Under Assumption 1, for an observed anchor index m and a nonempty deterministic interval J of sample indices define*

$$\mathcal{Z}_m(J) = \frac{1}{|J|} \sup_{t\ge 0} \left| \sum_{i\in J} \varepsilon_i(X_m, t) \right|^2, \tag{17}$$

where $|J|$ is the number of indices in J and the residuals are those in (2). These random variables are measurable and satisfy

$$\mathbb{E}\big[e^{\mathcal{Z}_m(J)/2} \mid X_m\big] \le \sqrt{\frac{4}{3}} \begin{cases} e^{10/3}, & m \notin J, \\ e^{23/6}, & m \in J. \end{cases} \tag{18}$$

For a deterministic partition into intervals, the variables $\mathcal{Z}_m(J)$ associated with its different intervals are conditionally independent given X_m.

Remark 6 *Lemma 2 retains the observation supplying the anchor in its original interval. Conditional on that observation, its residual is a bounded deterministic function. The exponential moment bound for this interval is $e^{1/2}$ times the bound for each other interval. Conditional independence across the disjoint intervals permits their exponential moments to be multiplied. The individual population profiles remain in the centering, which includes intervals containing observations from several true segments.*

Proof. We first verify the measurable versions used in the conditional calculation. A separable metric space has a countable base: balls with rational radii and centers in a countable dense subset form such a base. Finite products of these bases show that the Borel sigma field on the relevant product spaces equals the product of their Borel sigma fields. The metric is jointly continuous. Consequently, the set

$$\{(w,t,x) \in \Omega \times [0,\infty) \times \Omega : d(w,x) \le t\}$$

is closed, and its indicator is jointly Borel measurable. Here x is the point of Ω integrated against the marginal distribution of X_i. Integrating a bounded jointly Borel function in this argument gives a Borel function of (w,t). One can verify this fact first for indicators of measurable rectangles, whose integrals are a measurable indicator multiplied by a fixed probability, then for their finite linear combinations, and finally for bounded measurable functions by the monotone class theorem and monotone convergence applied to their nonnegative parts. Applying this fact to the displayed distance indicator shows that $(w,t) \mapsto \mu_i^\star(w,t)$ is jointly measurable. These population profile functions are the versions specified in Section 3.1.

Including the sample coordinates preserves measurability of the residual function $\varepsilon_i(w,t)$. The empirical integrals used in the theorem are measurable as well. Under **(H1)**, (12) expresses them as finite sums of these measurable evaluations. Under **(H2)**, the endpoints are the minimum and maximum of finitely many measurable distances. For example, on their positive span domain,

$$\|\varepsilon\|_{\mathcal{H}^n}^2 = \frac{1}{n(t_{\max} - t_{\min})} \sum_{m=1}^{n} \sum_{i=1}^{n} \int_0^\infty \mathbb{I}\{t_{\min} \le t \le t_{\max}\} \varepsilon_i(X_m,t)^2 \, dt.$$

The integrand is a nonnegative jointly measurable function of the sample and t, so integration in t preserves measurability by Tonelli's theorem. Its integral is finite because $|\varepsilon_i(X_m,t)| \le 1$ and both endpoints are finite. Division by the positive span preserves measurability on this domain. Replacing a residual by one of its finite segment averages or by a candidate estimation error gives the same conclusion for the projection energies and quadratic risks used below. Finite minima and finite sums preserve measurability of the algorithmic responses. Under Assumption 2, the OLS weights are deterministic, so the fitted coefficients are measurable on their stated domains as well. Throughout these evaluations, the anchors and endpoints retain their full sample values.

For fixed w, the indicators and their respective distribution functions are right continuous in t. Hence the finite sums of residuals are right continuous too. The countable supremum argument from Lemma 1 proves measurability of (17), including after the substitution $w = X_m$.

Fix $X_m = w$ in the conditional calculation. Independence means that the joint law factors into the marginal law of X_m and the product of the other marginal laws. A version of the conditional expectation of any bounded measurable function of X_m and the other coordinates is obtained by holding its anchor argument at w and integrating its remaining arguments against those marginals. To verify the defining property of conditional expectation, multiply this integrated function by any bounded measurable function of X_m and integrate in the law of X_m. Fubini's theorem gives the original joint expectation of the same product. Thus this version has the required conditional expectation property. Applying the same product integration to products of bounded functions of distinct remaining coordinates proves their conditional independence. The distances $d(w,X_i)$, with $i \ne m$, therefore have conditional distribution functions $\mu_i^\star(w,t)$. The same reasoning applies when the values at two fixed indices are specified, by integrating against the product law of all remaining coordinates.

If $m \notin J$, the supremum in (17) contains $|J|$ centered threshold indicators of these independent distances. Lemma 1, applied conditionally with $s = |J|$ and $\vartheta = 1/2$, gives

$$\mathbb{E}\big[e^{\mathcal{Z}_m(J)/2} \mid X_m\big] \le (1 - 1/4)^{-1/2} \exp\left\{\frac{5/2}{1 - 1/4}\right\} = \sqrt{\frac{4}{3}}\, e^{10/3}.$$

Suppose next that $m \in J$. At the conditioned anchor we have $d(w,X_m) = d(w,w) = 0$, and all the thresholds in the supremum are nonnegative. Hence its residual equals

$$\varepsilon_m(w,t) = 1 - \mu_m^\star(w,t), \qquad 0 \le \varepsilon_m(w,t) \le 1.$$

If $|J| \ge 2$, the triangle inequality for the residual sum, followed by the mean bound in Lemma 1 for the other $|J| - 1$ coordinates, gives

$$\mathbb{E}\big[\sqrt{\mathcal{Z}_m(J)} \mid X_m\big] \le \frac{1 + \sqrt{5(|J| - 1)}}{\sqrt{|J|}}.$$

Replacing any one of the other $|J|-1$ coordinates changes each residual sum by at most one. Applying the supremum inequality used in Lemma 1 and dividing by $\sqrt{|J|}$ shows that the value of $\sqrt{\mathcal{Z}_m(J)}$ changes by at most $1/\sqrt{|J|}$. The conditional coordinate martingale argument therefore gives, for every real η,

$$\mathbb{E}\left[\exp\left\{\eta\left(\sqrt{\mathcal{Z}_m(J)}-\mathbb{E}[\sqrt{\mathcal{Z}_m(J)}\mid X_m]\right)\right\}\middle| X_m\right]\le\exp\left\{\frac{(|J|-1)\eta^2}{8|J|}\right\}.$$

Indeed, each of its $|J|-1$ conditional martingale differences has range length at most $1/\sqrt{|J|}$, so each conditional exponential bound contributes the factor $e^{\eta^2/(8|J|)}$; multiplying those factors gives the displayed bound. We retain this dependence on $|J|$ in the squared exponential moment.

The Gaussian identity used to prove (16), at $\vartheta=1/2$, expresses $e^{\mathcal{Z}_m(J)/2}$ as the integral of $e^{u\sqrt{\mathcal{Z}_m(J)}}$ against the standard normal density. Conditional Tonelli applies to its nonnegative integrand. Using the preceding conditional exponential bound at every real u yields

$$\begin{aligned}\mathbb{E}[e^{\mathcal{Z}_m(J)/2}\mid X_m]&=\frac{1}{\sqrt{2\pi}}\int_{\mathbb{R}}e^{-u^2/2}\mathbb{E}[e^{u\sqrt{\mathcal{Z}_m(J)}}\mid X_m]\,du\\&\le\frac{1}{\sqrt{2\pi}}\int_{\mathbb{R}}\exp\left\{-\frac{1}{2}\frac{3|J|+1}{4|J|}u^2+u\mathbb{E}[\sqrt{\mathcal{Z}_m(J)}\mid X_m]\right\}du.\end{aligned}$$

The quadratic coefficient follows from

$$1-\frac{|J|-1}{4|J|}=\frac{3|J|+1}{4|J|}>0.$$

To evaluate the last integral, its exponent can be written as

$$-\frac{1}{2}\frac{3|J|+1}{4|J|}\left(u-\frac{4|J|}{3|J|+1}\mathbb{E}[\sqrt{\mathcal{Z}_m(J)}\mid X_m]\right)^2+\frac{2|J|}{3|J|+1}\left(\mathbb{E}[\sqrt{\mathcal{Z}_m(J)}\mid X_m]\right)^2.$$

Integration of the centered Gaussian therefore gives

$$\begin{aligned}\mathbb{E}[e^{\mathcal{Z}_m(J)/2}\mid X_m]&\le\sqrt{\frac{4|J|}{3|J|+1}}\exp\left\{\frac{2|J|}{3|J|+1}\left(\mathbb{E}[\sqrt{\mathcal{Z}_m(J)}\mid X_m]\right)^2\right\}\\&\le\sqrt{\frac{4|J|}{3|J|+1}}\exp\left\{\frac{2\{1+\sqrt{5(|J|-1)}\}^2}{3|J|+1}\right\},\end{aligned}$$

where the last line uses the conditional mean bound above. Cauchy Schwarz applied to the vectors $(1/2,\sqrt{5/3})$ and $(2,\sqrt{3(|J|-1)})$ gives

$$\{1+\sqrt{5(|J|-1)}\}^2=\left\{\frac{1}{2}2+\sqrt{\frac{5}{3}}\sqrt{3(|J|-1)}\right\}^2\le\left(\frac{1}{4}+\frac{5}{3}\right)\{4+3(|J|-1)\}=\frac{23}{12}(3|J|+1).$$

In particular,

$$\frac{4|J|}{3|J|+1}\le\frac{4}{3},\qquad\frac{2\{1+\sqrt{5(|J|-1)}\}^2}{3|J|+1}\le\frac{23}{6}.$$

Substitution proves

$$\mathbb{E}[e^{\mathcal{Z}_m(J)/2}\mid X_m]\le\sqrt{\frac{4}{3}}\,e^{23/6}.$$

For the remaining case $J=\{m\}$, the single bounded anchor residual gives $\mathcal{Z}_m(J)\le 1$, so its conditional exponential moment is at most $e^{1/2}$, which is bounded by the same displayed constant. This proves both cases in (18).

Finally, conditionally on X_m, the variable $\mathcal{Z}_m(J)$ is a measurable function of w and the coordinates indexed by $J\setminus\{m\}$. Distinct intervals in a partition have disjoint sets of these coordinates. Their conditional laws are products by the original independence assumption. The anchor residual, when included, is deterministic under this conditioning. Consequently the variables associated with the different intervals are conditionally independent, as asserted.

Lemma 3 *For two candidate configurations $T_K, T'_{K'}$, let $T_K \vee T'_{K'}$ denote their common refinement, formed by taking the union of their boundary locations, and let $\mathcal{J}$ be the resulting collection of intervals. Its projection is the segment averaging projection of Section 3.1 on these intervals. Under Assumption 1, consider **(H1)** with its complexity $\mathfrak{c}$ in (8). For every $y > 0$, there is an event $\mathcal{E}_y$ with probability at least $1 - e^{-y}$ such that, for every pair of candidate configurations,*

$$\|\Pi_{T_K \vee T'_{K'}} \varepsilon\|^2_{\mathcal{H}^n} \leq \mathfrak{c}(T_K) + \mathfrak{c}(T'_{K'}) + \frac{5}{8} y.$$

The event applies to all candidate collections described in Section 3.1 and all measurable penalties considered in Section 3.3.

Remark 7 *Lemma 3 uses the finite sum representation of the original empirical anchor and radius measure in **(H1)**. Conditioning on the two observations that determine one anchor and one radius leaves independent scalar threshold indicators at the remaining coordinates. A Gaussian integral converts their linear exponential bound into the required squared sum bound. Averaging these exponential moments over both original empirical indices then gives the complexity in (8). Every observation remains in the integrated loss and in the projection energy.*

Proof. Fix two deterministic configurations $T_K, T'_{K'}$ and use the collection $\mathcal{J}$ just specified. Thus

$$|\mathcal{J}| \leq K + K' + 1, \qquad |\mathcal{J}| + 1 \leq (K+1) + (K'+1).$$

For two sample indices $m, r \in \{1, \ldots, n\}$, define

$$\mathcal{A}_{mr}(T_K, T'_{K'}) = \sum_{J \in \mathcal{J}} \frac{1}{|J|} \left| \sum_{i \in J} \varepsilon_i\big(X_m, d(X_m, X_r)\big) \right|^2.$$

This is a measurable nonnegative random variable, since it is a finite sum of evaluations of the jointly measurable residual functions established in Lemma 2.

For an interval J in the common refinement, its coordinates in the projection all equal the same average. Thus

$$\begin{aligned} \sum_{i \in J} \int |(\Pi_{T_K \vee T'_{K'}} \varepsilon)_i(w,t)|^2 H(dw, dt) &= |J| \int \left| \frac{1}{|J|} \sum_{i \in J} \varepsilon_i(w,t) \right|^2 H(dw, dt) \\ &= \frac{1}{|J|} \int \left| \sum_{i \in J} \varepsilon_i(w,t) \right|^2 H(dw, dt). \end{aligned}$$

Summing over the intervals and applying (12) to each integral gives the equality that follows.

$$\|\Pi_{T_K \vee T'_{K'}} \varepsilon\|^2_{\mathcal{H}^n} = \frac{1}{n^2} \sum_{m=1}^{n} \sum_{r=1}^{n} \mathcal{A}_{mr}(T_K, T'_{K'}). \tag{19}$$

Let $\mathcal{I} = \{m, r\}$ be the set of distinct indices determining this anchor and radius. Its cardinality is one when $m = r$ and two otherwise. Condition on the observations with indices in $\mathcal{I}$. In this conditional calculation write $w = X_m$ and $t = d(X_m, X_r)$ for their now fixed values. All the other observations retain their product of marginal distributions by Assumption 1.

When $m = r$, the conditioned set consists of one index and the radius is zero. For each remaining coordinate its success probability is $\mu_i^\star(w, 0) = \mathbb{P}(X_i = w)$, as follows from the defining property of a metric. When m and r are distinct indices whose observations have the same value, the conditioned set still has two elements and the radius is again zero. The product integration treats these two cases according to the distinct indices in $\mathcal{I}$ and therefore includes both diagonal index pairs and tied observations.

For one interval $J \in \mathcal{J}$, put $l = |J|$, $v = |J \cap \mathcal{I}|$, and $q = l - v$. These are nonnegative integers, with $l \geq 1$. Decompose its residual sum as

$$\sum_{i \in J} \varepsilon_i(w,t) = \mathcal{U} + \mathfrak{s}, \qquad \mathcal{U} = \sum_{i \in J \setminus \mathcal{I}} \varepsilon_i(w,t), \qquad \mathfrak{s} = \sum_{i \in J \cap \mathcal{I}} \varepsilon_i(w,t).$$

The term $\mathfrak{s}$ is deterministic under the conditioning. The conditioned indicators all equal one: the anchor has distance zero from itself, and the observation supplying the radius has distance t from the anchor. Each associated residual therefore belongs to $[0, 1]$, giving $0 \le \mathfrak{s} \le v$ and $|\mathfrak{s}| \le v$.

For $i \in J \setminus \mathcal{I}$, the indicator at the fixed (w, t) is Bernoulli with mean $\mu_i^\star(w, t)$ from Section 3.1. In the original sample notation, the product integration justified in Lemma 2 gives

$$\mathbb{E}[\mathbb{I}\{d(X_m, X_i) \le d(X_m, X_r)\}|(X_j)_{j\in\mathcal{I}}] = \mu_i^\star\big(X_m, d(X_m, X_r)\big), \qquad i \in J \setminus \mathcal{I}.$$

The anchor and radius on the right are the observed values from the original empirical measure. Subtracting this conditional mean shows that each remaining residual has mean zero. Its support lies in an interval of length one. The elementary exponential inequality proved in Lemma 1 therefore gives

$$\mathbb{E}[e^{h\mathcal{U}} \mid (X_i)_{i\in\mathcal{I}}] \le e^{qh^2/8}, \qquad h \in \mathbb{R}.$$

Conditional independence permits multiplication of the unit interval bounds for every real h. When $q = 0$, the sum is zero and the display holds with equality.

We calculate a squared exponential moment with the fixed exponent $\eta = 8/5$. For a real number a, completing the square in a Gaussian integral gives

$$e^{\eta a^2/l} = \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} \exp\left\{-\frac{z^2}{2} + \sqrt{\frac{2\eta}{l}}\, za\right\} dz.$$

Apply this identity to $a = \mathcal{U} + \mathfrak{s}$, use conditional Tonelli for the nonnegative integrand, and then use the preceding bound with $h = \sqrt{2\eta/l}\, z$. For this exponent, which satisfies $\eta < 2$, this yields

$$\begin{aligned}\mathbb{E}\Big[e^{\eta(\mathcal{U}+\mathfrak{s})^2/l}\Big|(X_i)_{i\in\mathcal{I}}\Big] &\le \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} \exp\left\{-\frac{1}{2}\left(1 - \frac{\eta q}{2l}\right) z^2 + \sqrt{\frac{2\eta}{l}}\, z\mathfrak{s}\right\} dz \\ &= \left(1 - \frac{\eta q}{2l}\right)^{-1/2} \exp\left\{\frac{\eta\mathfrak{s}^2}{l - \eta q/2}\right\}.\end{aligned}$$

The equality follows by completing the square once more and changing variables to a centered Gaussian with precision $1 - \eta q/(2l) > 0$. Indeed, $0 \le q \le l$ and $\eta = 8/5$ imply

$$1 - \frac{\eta q}{2l} \ge 1 - \frac{\eta}{2} = \frac{1}{5} > 0,$$

so the stated change of variables has a positive scaling factor for every interval.

We next give bounds that preserve the number v of conditioned coordinates in the interval. Since $\mathcal{I}$ contains either one or two indices, $v \in \{0, 1, 2\}$. We shall use two elementary upper bounds for the logarithm. For $0 \le z < 1$, monotonicity of the integrand gives

$$-\log(1 - z) = \int_0^z \frac{ds}{1 - s} \le \frac{z}{1 - z}.$$

A second bound follows from convexity. For $z > 0$ and $0 \le s \le z$, the function $s \mapsto (1 - s)^{-1}$ has second derivative $2(1 - s)^{-3} > 0$ and therefore lies below the line joining its endpoint values. Thus,

$$\frac{1}{1 - s} \le \left(1 - \frac{s}{z}\right) + \frac{s}{z}\frac{1}{1 - z}.$$

Integrating this inequality from 0 to z gives

$$-\log(1 - z) \le \frac{z}{2}\left(1 + \frac{1}{1 - z}\right) = \frac{z(2 - z)}{2(1 - z)}.$$

At $z = 0$ both logarithm bounds hold with equality.

First suppose $v = 0$. Then $q = l$ and $\mathfrak{s} = 0$. The squared moment bound is

$$\left(1 - \frac{\eta}{2}\right)^{-1/2} = \sqrt{5},$$

since $\eta = 8/5$. Thus the conditional squared moment is at most $\sqrt{5}$ on every interval containing no conditioned coordinate.

Now suppose $v \geq 1$ and put $z = \eta q/(2l)$. The Gaussian integral and $|\mathfrak{s}| \leq v$ bound the logarithm of the squared moment by

$$-\frac{1}{2}\log(1-z) + \frac{\eta v^2}{l(1-z)}.$$

We show that this expression is at most ηv. Since $l = v + q$, direct subtraction gives

$$\eta v - \frac{\eta v^2}{l(1-z)} = \frac{\eta v\{l(1-z) - v\}}{l(1-z)} = \frac{\eta v q(1-\eta/2)}{l(1-z)} = \frac{2v(1-\eta/2)z}{1-z} = \frac{2v}{5}\frac{z}{1-z}.$$

All denominators are positive by the preceding Gaussian precision bound. If $q = 0$, then $z = 0$, $l = v$, and the logarithmic term is zero; the asserted inequality follows with equality for its upper bound.

If $v = 2$, the first logarithm bound gives

$$-\frac{1}{2}\log(1-z) \leq \frac{z}{2(1-z)} \leq \frac{2v}{5}\frac{z}{1-z}.$$

The second inequality uses $2v/5 = 4/5 > 1/2$, and the last expression equals the difference just computed. For the remaining case $v = 1$ and $q \geq 1$, the integer q satisfies

$$z = \frac{4q}{5(1+q)} \geq \frac{2}{5}, \qquad \frac{2-z}{4} \leq \frac{2}{5}.$$

Indeed, $q \geq 1$ implies $q/(1+q) \geq 1/2$, which gives the first inequality; subtracting it from 2 and dividing by 4 gives the second. The second logarithm bound now yields

$$-\frac{1}{2}\log(1-z) \leq \frac{z(2-z)}{4(1-z)} \leq \frac{2}{5}\frac{z}{1-z}.$$

This is the computed difference for $v = 1$. These cases cover every possible (v, q) with $v \geq 1$. Consequently,

$$\mathbb{E}\Big[e^{\eta(\mathcal{U}+\mathfrak{s})^2/l}\Big|(X_i)_{i\in\mathcal{I}}\Big] \leq \begin{cases}\sqrt{5}, & v = 0,\\ e^{\eta v}, & v \geq 1.\end{cases}$$

Under the conditioning, the random sums $\mathcal{U}$ associated with different intervals depend on disjoint sets of the remaining independent sample coordinates. Their squared exponential moments therefore multiply. At least one interval meets $\mathcal{I}$. Let $\mathfrak{r}$ be the number of intervals meeting $\mathcal{I}$. The bounds just obtained give

$$\mathbb{E}\Big[e^{\eta\mathcal{A}_{mr}(T_K,T'_{K'})}\Big|(X_i)_{i\in\mathcal{I}}\Big] \leq (\sqrt{5})^{|\mathcal{J}|-\mathfrak{r}}\exp\left\{\eta\sum_{J\in\mathcal{J}}|J\cap\mathcal{I}|\right\} = (\sqrt{5})^{|\mathcal{J}|-\mathfrak{r}}e^{\eta|\mathcal{I}|} \leq (\sqrt{5})^{|\mathcal{J}|-1}e^{16/5}.$$

Here $\mathfrak{r} \geq 1$, $|\mathcal{I}| \leq 2$, and $\eta = 8/5$ give the last inequality. Taking unconditional expectations preserves the bound. Pointwise convexity of the exponential and (19) now give

$$\mathbb{E}\exp\left\{\frac{8}{5}\|\Pi_{T_K\vee T'_{K'}}\varepsilon\|^2_{\mathcal{H}^n}\right\} \leq \frac{1}{n^2}\sum_{m=1}^{n}\sum_{r=1}^{n}\mathbb{E}e^{(8/5)\mathcal{A}_{mr}(T_K,T'_{K'})} \leq (\sqrt{5})^{|\mathcal{J}|-1}e^{16/5} \leq (\sqrt{5})^{K+K'}e^{16/5}.$$

The final inequality uses $|\mathcal{J}| \leq K + K' + 1$. This averaging retains the sample dependence between every anchor, every radius, and every profile. It uses their original empirical weights $1/n^2$.

By the **(H1)** definition of $\mathfrak{c}$ in (8), the proposed threshold multiplied by the exponential parameter is

$$\frac{8}{5}\left\{\mathfrak{c}(T_K) + \mathfrak{c}(T'_{K'}) + \frac{5}{8}y\right\} = (K+K'+2)\log(\sqrt{5}+e^{8/5}) + \log\binom{n-1}{K} + \log\binom{n-1}{K'} + y.$$

Exponential Markov inequality, with the moment bound just proved, therefore yields

$$\begin{aligned}&\mathbb{P}\left\{\|\Pi_{T_K\vee T'_{K'}}\varepsilon\|^2_{\mathcal{H}^n} > \mathfrak{c}(T_K) + \mathfrak{c}(T'_{K'}) + \frac{5}{8}y\right\}\\ &\leq \frac{e^{16/5}}{5}\left(\frac{\sqrt{5}}{\sqrt{5}+e^{8/5}}\right)^{K+K'+2}\exp\left\{-\log\binom{n-1}{K} - \log\binom{n-1}{K'} - y\right\}.\end{aligned}$$

The factor $e^{16/5}/5$ follows from $(\sqrt{5})^{K+K'}e^{16/5} = (e^{16/5}/5)(\sqrt{5})^{K+K'+2}$.

Let $\mathcal{E}_y$ be the intersection of the corresponding nonstrict inequalities over all ordered pairs of configurations. Every constituent event is measurable by the finite sum representation above. There are $\binom{n-1}{K}$ configurations with K boundaries. Summing their exceptional probabilities cancels each model count and gives

$$\begin{aligned}
\mathbb{P}(\mathcal{E}_y^c) &\le \frac{e^{16/5}}{5}e^{-y}\left[\sum_{K=0}^{n-1}\left(\frac{\sqrt{5}}{\sqrt{5}+e^{8/5}}\right)^{K+1}\right]^2 \\
&= \frac{e^{16/5}}{5}e^{-y}\left[\frac{\sqrt{5}/(\sqrt{5}+e^{8/5})}{1-\sqrt{5}/(\sqrt{5}+e^{8/5})}\left\{1-\left(\frac{\sqrt{5}}{\sqrt{5}+e^{8/5}}\right)^{n}\right\}\right]^2 \\
&= \frac{e^{16/5}}{5}e^{-y}\left[\sqrt{5}e^{-8/5}\left\{1-\left(\frac{\sqrt{5}}{\sqrt{5}+e^{8/5}}\right)^{n}\right\}\right]^2 \\
&= e^{-y}\left\{1-\left(\frac{\sqrt{5}}{\sqrt{5}+e^{8/5}}\right)^{n}\right\}^2 < e^{-y}.
\end{aligned}$$

The finite geometric series formula gives the second line, since its ratio lies strictly between zero and one. Subtracting the numerator $\sqrt{5}$ from the denominator $\sqrt{5}+e^{8/5}$ gives

$$\frac{\sqrt{5}/(\sqrt{5}+e^{8/5})}{1-\sqrt{5}/(\sqrt{5}+e^{8/5})} = \sqrt{5}e^{-8/5}.$$

Its square multiplied by $e^{16/5}/5$ equals one, giving the last line. This calculation shows how the finite conditional moment and the model counting weights determine the coefficient $\log(\sqrt{5}+e^{8/5})$ in (8). Thus $\mathbb{P}(\mathcal{E}_y) \ge 1-e^{-y}$. The event includes every configuration before imposing candidate restrictions or choosing a penalty, which proves the final assertion.

Lemma 4 *Under Assumption 1, fix either weighting scheme. For every $y > 0$ there is an event $\mathcal{E}_y$ with $\mathbb{P}(\mathcal{E}_y) \ge 1-e^{-y}$ such that, for every pair of configurations, the following bound holds on that event, with the **(H2)** bound evaluated on its domain $t_{\max} > t_{\min}$:*

$$\|\Pi_{T_K \vee T'_{K'}}\varepsilon\|_{\mathcal{H}^n}^2 \le \mathfrak{c}(T_K) + \mathfrak{c}(T'_{K'}) + \mathfrak{a}y. \tag{20}$$

Remark 8 *Lemma 4 is the probability statement that controls the sample dependence of the kernel. At each observed anchor, the integral of the squared residual sum is bounded by its supremum over the radius. Exponential convexity then averages these bounds over the original empirical anchor measure. For **(H2)**, this controls the random integration interval through a single radius supremum. For **(H1)**, Lemma 3 uses the empirical radii directly and supplies its smaller constants. Its construction uses only the sample residuals and the candidate configurations, so it also supports penalties selected from the data.*

Proof. For **(H1)**, the conclusion is Lemma 3. We give the **(H2)** calculation in full, using its values of $\mathfrak{a}$ and $\mathfrak{c}$ in (8). Fix a pair $T_K, T'_{K'}$, and use its common refinement and interval collection $\mathcal{J}$ as specified in Lemma 3. Each configuration has K or K' internal boundaries, respectively. Their union has at most $K+K'$ internal boundaries, and therefore

$$|\mathcal{J}| \le K+K'+1 = (K+1)+(K'+1)-1.$$

For every anchor index m, there is a unique interval of $\mathcal{J}$ containing m. By conditional independence in Lemma 2, the conditional exponential moment factors as a product over intervals. Using

(18) for the unique interval containing m and for each of the other intervals gives

$$\begin{aligned}
&\mathbb{E}\left[\exp\left\{\frac{1}{2}\sum_{J\in\mathcal{J}}\mathcal{Z}_m(J)\right\}\,\middle|\,X_m\right] \\
&=\prod_{J\in\mathcal{J}}\mathbb{E}[e^{\mathcal{Z}_m(J)/2}\mid X_m]\le\left(\frac{2}{\sqrt{3}}e^{10/3}\right)^{|\mathcal{J}|-1}\left(\frac{2}{\sqrt{3}}e^{23/6}\right) \\
&=\exp\left\{\left(\frac{10}{3}+\frac{1}{2}\log\frac{4}{3}\right)|\mathcal{J}|+\frac{1}{2}\right\}.
\end{aligned}$$

Define the scalar envelope associated with this pair by

$$\mathcal{Q}(T_K,T'_{K'})=\frac{1}{n}\sum_{m=1}^{n}\sum_{J\in\mathcal{J}}\mathcal{Z}_m(J). \tag{21}$$

On the **(H2)** domain $t_{\max}>t_{\min}$, the averaging projection and empirical anchor measure give

$$\begin{aligned}
\|\Pi_{T_K\vee T'_{K'}}\varepsilon\|^2_{\mathcal{H}^n}&=\frac{1}{n}\sum_{m=1}^{n}\sum_{J\in\mathcal{J}}\frac{1}{|J|}\int_{t_{\min}}^{t_{\max}}\left|\sum_{i\in J}\varepsilon_i(X_m,t)\right|^2\nu(dt)\\
&\le\frac{1}{n}\sum_{m=1}^{n}\sum_{J\in\mathcal{J}}\mathcal{Z}_m(J).
\end{aligned}$$

Here ν is the sample dependent probability measure of Section 2. Its total mass is one and its random integration interval is contained in $[0,\infty)$, so the integral is bounded by the supremum in (17). Therefore, at each realization in the **(H2)** domain,

$$\|\Pi_{T_K\vee T'_{K'}}\varepsilon\|^2_{\mathcal{H}^n}\le\mathcal{Q}(T_K,T'_{K'}). \tag{22}$$

The anchor average in (21) permits a direct exponential bound. Pointwise convexity of the exponential yields

$$\exp\{\mathcal{Q}(T_K,T'_{K'})/2\}\le\frac{1}{n}\sum_{m=1}^{n}\exp\left\{\frac{1}{2}\sum_{J\in\mathcal{J}}\mathcal{Z}_m(J)\right\}.$$

Taking expectations, then applying the conditional exponential bound already obtained, gives

$$\begin{aligned}
\mathbb{E}\exp\{\mathcal{Q}(T_K,T'_{K'})/2\}&\le\frac{1}{n}\sum_{m=1}^{n}\mathbb{E}\left[\mathbb{E}\left[\exp\left\{\frac{1}{2}\sum_{J\in\mathcal{J}}\mathcal{Z}_m(J)\right\}\middle|X_m\right]\right]\\
&\le\exp\left\{\left(\frac{10}{3}+\frac{1}{2}\log\frac{4}{3}\right)|\mathcal{J}|+\frac{1}{2}\right\}.
\end{aligned}$$

This exponential average preserves the dependence among the anchors.

The moment bound retains the contribution of the unique interval containing the conditioned anchor. Since $|\mathcal{J}|\le K+K'+1$, it gives

$$\mathbb{E}e^{\mathcal{Q}(T_K,T'_{K'})/2}\le\exp\left\{\left(\frac{10}{3}+\frac{1}{2}\log\frac{4}{3}\right)(K+K'+2)-\frac{17}{6}-\frac{1}{2}\log\frac{4}{3}\right\}.$$

Indeed, replacing $|\mathcal{J}|$ by $K+K'+1$ and separating off one from $K+K'+2$ leaves the constant $1/2-10/3-(1/2)\log(4/3)=-17/6-(1/2)\log(4/3)$. This constant is retained in the model sum below.

For this calculation define the numerical ratio

$$\mathfrak{q}=\left(1+e^{-17/12}\left(\frac{3}{4}\right)^{1/4}\right)^{-1}.$$

Its added term is positive, so $0 < \mathfrak{q} < 1$. Factoring $(2/\sqrt{3})e^{10/3}$ out of the logarithm in (8) yields

$$\log\left(\frac{2}{\sqrt{3}}e^{10/3} + \sqrt[4]{\frac{4}{3}}\,e^{23/12}\right) = \frac{10}{3} + \frac{1}{2}\log\frac{4}{3} + \log\left(1 + e^{-17/12}\left(\frac{3}{4}\right)^{1/4}\right) = \frac{10}{3} + \frac{1}{2}\log\frac{4}{3} - \log\mathfrak{q}.$$

The ratio of the second summand to the first is $e^{23/12-10/3}(4/3)^{1/4}/(4/3)^{1/2} = e^{-17/12}(3/4)^{1/4}$, which verifies the factorization directly. Consequently, one half of the threshold in (20) is

$$\frac{1}{2}\{\mathfrak{c}(T_K) + \mathfrak{c}(T'_{K'}) + 2y\} = \left(\frac{10}{3} + \frac{1}{2}\log\frac{4}{3} - \log\mathfrak{q}\right)(K + K' + 2) + \log\binom{n-1}{K} + \log\binom{n-1}{K'} + y.$$

By exponential Markov inequality, dividing the preceding exponential moment by the exponential of this threshold gives

$$\mathbb{P}\{\mathcal{Q}(T_K, T'_{K'}) > \mathfrak{c}(T_K) + \mathfrak{c}(T'_{K'}) + 2y\} \le e^{-y-17/6}\left(\frac{3}{4}\right)^{1/2}\frac{\mathfrak{q}^{K+K'+2}}{\binom{n-1}{K}\binom{n-1}{K'}}.$$

Let $\mathcal{E}_y$ be the intersection, over all ordered pairs of configurations, of the events

$$\mathcal{Q}(T_K, T'_{K'}) \le \mathfrak{c}(T_K) + \mathfrak{c}(T'_{K'}) + 2y.$$

Each event is measurable by the countable supremum and finite sum representations already established. There are $\binom{n-1}{K}$ configurations with K boundaries, since their boundary positions are distinct elements of $\{1, \dots, n-1\}$. The union bound and cancellation of both model counts give

$$\begin{aligned}\mathbb{P}(\mathcal{E}_y^c) &\le e^{-y-17/6}\left(\frac{3}{4}\right)^{1/2}\sum_{K=0}^{n-1}\sum_{K'=0}^{n-1}\mathfrak{q}^{K+K'+2} \\ &= e^{-y-17/6}\left(\frac{3}{4}\right)^{1/2}\left[\sum_{K=0}^{n-1}\mathfrak{q}^{K+1}\right]^2.\end{aligned}$$

The two sums correspond to the two configuration sizes. Since $0 < \mathfrak{q} < 1$, the finite geometric series and the definition of $\mathfrak{q}$ give

$$\sum_{K=0}^{n-1}\mathfrak{q}^{K+1} = \frac{\mathfrak{q}(1-\mathfrak{q}^n)}{1-\mathfrak{q}}, \qquad \frac{\mathfrak{q}}{1-\mathfrak{q}} = e^{17/12}\left(\frac{4}{3}\right)^{1/4}.$$

Substituting these two identities yields the cancellation

$$\begin{aligned}\mathbb{P}(\mathcal{E}_y^c) &\le e^{-y-17/6}\left(\frac{3}{4}\right)^{1/2}\left[e^{17/12}\left(\frac{4}{3}\right)^{1/4}(1-\mathfrak{q}^n)\right]^2 \\ &= e^{-y}(1-\mathfrak{q}^n)^2 < e^{-y}.\end{aligned}$$

Thus $\mathbb{P}(\mathcal{E}_y) \ge 1 - e^{-y}$, with the **(H2)** complexity in (8). In particular, its coefficient of $K+1$ uses both the interval moment and the prefactor in the finite model sum.

Finally, (22) implies (20) on $\mathcal{E}_y$. For **(H2)**, the envelope defining this event is defined on the entire sample space and uses all candidate configurations. For **(H1)**, Lemma 3 has the same uniformity. Thus, for the chosen scheme, all deterministic candidate restrictions and all choices of penalty share its event.

Proof of Theorem 3. Fix the weighting scheme, and work on its event $\mathcal{E}_y$ from Lemma 4. The constants $\mathfrak{a}$ and the complexity $\mathfrak{c}$ are those specified for this scheme in Section 3.3. The event was defined by imposing the bound on every pair from a finite set of configurations. At a realization in that event, the selected configuration is one member of the same set, so its pair with each comparator satisfies the bound as well. This justifies its use for the data selected configuration with the same constants and the same event. Choose any $T_K \in \mathcal{P}$. The minimizing property and (4) give

$$\frac{1}{n}\|Y - \widehat{\mu}_{\widehat{T}_{\widehat{K}}}\|^2_{\mathcal{H}^n} + \mathfrak{p}(\widehat{T}_{\widehat{K}}) \le \frac{1}{n}\|Y - \widehat{\mu}_{T_K}\|^2_{\mathcal{H}^n} + \mathfrak{p}(T_K).$$

For any $u \in \mathcal{H}^n$, the identity $Y = \mu^\star + \varepsilon$ and direct expansion show

$$\frac{1}{n}\|Y-u\|_{\mathcal{H}^n}^2 = \frac{1}{n}\|\varepsilon\|_{\mathcal{H}^n}^2 + \mathcal{R}(u) - \frac{2}{n}\langle \varepsilon, u-\mu^\star\rangle_{\mathcal{H}^n}.$$

Apply this formula to the two estimators in the preceding comparison and cancel the common residual norm. This yields

$$\mathcal{R}(\widehat{\mu}_{\widehat{T}_{\widehat{K}}}) \le \mathcal{R}(\widehat{\mu}_{T_K}) + \mathfrak{p}(T_K) - \mathfrak{p}(\widehat{T}_{\widehat{K}}) + \frac{2}{n}\left\langle \varepsilon, \widehat{\mu}_{\widehat{T}_{\widehat{K}}} - \widehat{\mu}_{T_K}\right\rangle_{\mathcal{H}^n}. \tag{23}$$

Each fitted vector is constant on its own segments. On every interval of their common refinement, both vectors are constant, so their difference belongs to $\mathcal{F}_{T_K \vee \widehat{T}_{\widehat{K}}}$. Orthogonality of the projection established at the beginning of this appendix gives

$$\left\langle \varepsilon, \widehat{\mu}_{\widehat{T}_{\widehat{K}}} - \widehat{\mu}_{T_K}\right\rangle_{\mathcal{H}^n} = \left\langle \Pi_{T_K \vee \widehat{T}_{\widehat{K}}}\varepsilon, \widehat{\mu}_{\widehat{T}_{\widehat{K}}} - \widehat{\mu}_{T_K}\right\rangle_{\mathcal{H}^n}.$$

Subtract and add $\mu^\star$ inside the second vector. Cauchy Schwarz and the triangle inequality then yield

$$\frac{2}{n}\left\langle \varepsilon, \widehat{\mu}_{\widehat{T}_{\widehat{K}}} - \widehat{\mu}_{T_K}\right\rangle_{\mathcal{H}^n} \le \frac{2}{n}\|\Pi_{T_K \vee \widehat{T}_{\widehat{K}}}\varepsilon\|_{\mathcal{H}^n}\left\{\|\widehat{\mu}_{\widehat{T}_{\widehat{K}}} - \mu^\star\|_{\mathcal{H}^n} + \|\widehat{\mu}_{T_K} - \mu^\star\|_{\mathcal{H}^n}\right\}.$$

We bound the two products separately. For nonnegative real a, b, expansion of $(a/\sqrt{3} - \sqrt{3}b)^2 \ge 0$ gives $2ab \le a^2/3 + 3b^2$, and expansion of $(a-b)^2 \ge 0$ gives $2ab \le a^2 + b^2$. Applying the first inequality to the selected estimate and the second to the comparator gives

$$\frac{2}{n}\|\Pi_{T_K \vee \widehat{T}_{\widehat{K}}}\varepsilon\|_{\mathcal{H}^n}\|\widehat{\mu}_{\widehat{T}_{\widehat{K}}} - \mu^\star\|_{\mathcal{H}^n} \le \frac{1}{3}\mathcal{R}(\widehat{\mu}_{\widehat{T}_{\widehat{K}}}) + \frac{3}{n}\|\Pi_{T_K \vee \widehat{T}_{\widehat{K}}}\varepsilon\|_{\mathcal{H}^n}^2,$$
$$\frac{2}{n}\|\Pi_{T_K \vee \widehat{T}_{\widehat{K}}}\varepsilon\|_{\mathcal{H}^n}\|\widehat{\mu}_{T_K} - \mu^\star\|_{\mathcal{H}^n} \le \mathcal{R}(\widehat{\mu}_{T_K}) + \frac{1}{n}\|\Pi_{T_K \vee \widehat{T}_{\widehat{K}}}\varepsilon\|_{\mathcal{H}^n}^2.$$

The total contribution is therefore at most one third of the selected risk, the full comparator risk, and four times the projection energy divided by n. Lemma 4 applies to the selected configuration because its event is uniform over every pair. Substitute these bounds into (23) and subtract one third of the selected risk from both sides. The result is

$$\frac{2}{3}\mathcal{R}(\widehat{\mu}_{\widehat{T}_{\widehat{K}}}) \le 2\mathcal{R}(\widehat{\mu}_{T_K}) + \mathfrak{p}(T_K) - \mathfrak{p}(\widehat{T}_{\widehat{K}}) + \frac{4}{n}\{\mathfrak{c}(T_K) + \mathfrak{c}(\widehat{T}_{\widehat{K}}) + \mathfrak{a}y\}. \tag{24}$$

The definition of $\mathfrak{b}$ in Theorem 3 gives

$$\frac{4\mathfrak{c}(\widehat{T}_{\widehat{K}})}{n} - \mathfrak{p}(\widehat{T}_{\widehat{K}}) \le \mathfrak{b}.$$

Use this inequality in (24) and multiply by $3/2$. For every comparator,

$$\mathcal{R}(\widehat{\mu}_{\widehat{T}_{\widehat{K}}}) \le 3\mathcal{R}(\widehat{\mu}_{T_K}) + \frac{3}{2}\mathfrak{p}(T_K) + \frac{6\mathfrak{c}(T_K)}{n} + \frac{3}{2}\mathfrak{b} + \frac{6\mathfrak{a}y}{n}.$$

Taking the infimum over $\mathcal{P}$ proves (9). The event was determined through the projection energies, so this calculation applies to every finite realized penalty on that same event.

Under (10), the definition of $\mathfrak{b}$ gives $\mathfrak{b} = 0$. For the comparator, multiplying $\mathfrak{p}(T_K) \ge 4\mathfrak{c}(T_K)/n$ by $3/2$ gives

$$\frac{6\mathfrak{c}(T_K)}{n} \le \frac{3}{2}\mathfrak{p}(T_K).$$

The two comparator penalty terms therefore sum to at most $3\mathfrak{p}(T_K)$. Substitution yields

$$\mathcal{R}(\widehat{\mu}_{\widehat{T}_{\widehat{K}}}) \le 3\{\mathcal{R}(\widehat{\mu}_{T_K}) + \mathfrak{p}(T_K)\} + \frac{6\mathfrak{a}y}{n}.$$

Taking the infimum proves (11). Since $\mathfrak{a} = 5/8$ under **(H1)** and $\mathfrak{a} = 2$ under **(H2)**, the respective confidence terms are $15y/(4n)$ and $12y/n$.

Finally, Algorithm 3 uses

$$\mathrm{pen}(T_K) = \frac{\widehat{c}_1 \log\binom{n-1}{K} + \widehat{c}_2(K+1)}{n}.$$

Under **(H1)**, (7) leaves these coefficients unchanged, and the sufficient condition is

$$\widehat{c}_1 \geq \frac{5}{2}, \qquad \widehat{c}_2 \geq \frac{5}{2}\log\bigl(\sqrt{5}+e^{8/5}\bigr).$$

Indeed, multiplying its complexity by $4/n$ gives these two coefficients. Under **(H2)**, the effective penalty is

$$\mathfrak{p}(T_K) = \frac{1}{n}\left[\widehat{c}_1 \log\binom{n-1}{K} + \left\{\widehat{c}_2 - \frac{t_{\min}}{n(t_{\max}-t_{\min})}\right\}(K+1)\right].$$

Multiplication of the **(H2)** complexity by $4/n$ gives the coefficients 8 and $8\log\left(\frac{2}{\sqrt{3}}e^{10/3} + \sqrt[4]{\frac{4}{3}}\,e^{23/12}\right)$. Hence sufficient values are

$$\widehat{c}_1 \geq 8, \qquad \widehat{c}_2 \geq 8\log\left(\frac{2}{\sqrt{3}}e^{10/3} + \sqrt[4]{\frac{4}{3}}\,e^{23/12}\right) + \frac{t_{\min}}{n(t_{\max}-t_{\min})}.$$

The quantities $\log\binom{n-1}{K}$ and $K+1$ are nonnegative, so comparison of the coefficients implies (10) for every candidate. This verifies the penalty conditions and completes the proof using the original normalized losses, empirical weights, and algorithmic costs.

A.1.2 Slope Regression Calibration and Proof of Theorem 4

The following result analyzes the original slope regression calibration; Baudry et al. (2012, Section 4) discuss slope estimation for penalty calibration.

For slope calibration, use the full feasible collection $\mathcal{P}$ and updated $K_{\max}$ from Section 3.1, together with K_0, $x_1(K)$, $x_2(K)$, $\widehat{\mathcal{R}}$, $(\widehat{\beta}_0, \widehat{\beta}_1, \widehat{\beta}_2)$, and $(\widehat{c}_1, \widehat{c}_2)$ from Algorithm 3. In particular, $\widehat{\mathcal{R}}(K+1)$ is the optimized algorithmic response for a specified number of changes, while $\widehat{\mathcal{R}}_n(T_K)$ in (4) is the integrated loss for a specified configuration. All fitted responses are functions of the same observations.

Assumption 2 *The inputs $\ell_{\min}$, $K_{\max}$ before its feasibility update,* frac*, and α of Algorithm 3 are deterministic and satisfy its stated input ranges, with $\ell_{\min}$ and $K_{\max}$ integer valued. Its resulting fitting range satisfies $K_{\max} - K_0 + 1 \geq 3$.*

Remark 9 *The feasibility update gives finite optimized costs, and three consecutive fitted model sizes identify the intercept and two slopes uniquely. Lemma 5 proves this through strict discrete concavity of the logarithmic binomial coefficient. Deterministic inputs make the OLS weight matrix fixed; they may depend on n, and the actual matrix enters the concentration radius. The three fitted model sizes imply $K_{\max} \geq 2$, so feasibility and $\ell_{\min} \geq 2$ give $n \geq (K_{\max}+1)\ell_{\min} \geq 6$. In particular, the factor $1 - 2/n$ in that radius is positive. It retains the segment length contribution in the exchange bound proved in Lemma 6. The iid structure within true segments from Assumption 1 supplies invariance under their permutations. Lemma 6 combines this invariance with sequential conditional expectations and range bounds, following the concentration principles in McDiarmid (1998, Section 3.2). Related concentration for independent permutations is studied by McDiarmid (2002).*

Let $\mathcal{X}$ be the design matrix with rows $(1, x_1(K), x_2(K))$, indexed by $K = K_0, \ldots, K_{\max}$, and set $\mathcal{W} = (\mathcal{X}^\top\mathcal{X})^{-1}\mathcal{X}^\top$. The row indices of $\mathcal{W}$ are $j = 0, 1, 2$, corresponding to the intercept and the two slopes, and its column indices are the fitted values of K. Thus ordinary least squares gives, for $j = 0, 1, 2$,

$$\widehat{\beta}_j = \sum_{K=K_0}^{K_{\max}} \mathcal{W}_{jK}\widehat{\mathcal{R}}(K+1), \qquad \|\mathcal{W}_j\|_1 = \sum_{K=K_0}^{K_{\max}} |\mathcal{W}_{jK}|.$$

Let $\mathfrak{S}$ be the finite group of permutations of $\{1, \ldots, n\}$ that map every true segment into itself. Its cardinality is $|\mathfrak{S}| = \prod_{k=0}^{K^\star}(t^\star_{k+1} - t^\star_k)!$. For $\varsigma \in \mathfrak{S}$, a superscript ς on an algorithmic quantity denotes its value computed from $(X_{\varsigma(1)}, \ldots, X_{\varsigma(n)})$. Define the finite orbit averages

$$\overline{\mathcal{R}}(K+1) = \frac{1}{|\mathfrak{S}|}\sum_{\varsigma\in\mathfrak{S}} \widehat{\mathcal{R}}^{\varsigma}(K+1), \qquad \overline{\beta}_j = \sum_{K=K_0}^{K_{\max}} \mathcal{W}_{jK}\overline{\mathcal{R}}(K+1).$$

The response average is defined for $K = 0, \ldots, K_{\max}$ and the coefficient average for $j = 0, 1, 2$. The group average includes every reordering within the true segments, with multiplicities, and defines the orbit averaged response curve used for the analysis. Algorithm 3 fits its coefficients to the observed ordering.

Using the observations Y_i of Section 3.1, define

$$\begin{aligned}
\mathfrak{v} &= \sum_{k=0}^{K^\star} (t^\star_{k+1} - t^\star_k - 1) \max_{t^\star_k < i, l \le t^\star_{k+1}} \|Y_i - Y_l\|^4_{\mathcal{H}}, \\
\mathfrak{e}_j(x) &= \left(1 - \frac{2}{n}\right) \frac{\|\mathcal{W}_j\|_1}{n} \sqrt{2\mathfrak{v}(x + \log 2)}, \qquad j = 1, 2, \quad x > 0, \\
\underline{c}_j(x) &= \left(-\alpha\{\overline{\beta}_j + \mathfrak{e}_j(x)\}\right)_+, \qquad j = 1, 2.
\end{aligned}$$

The maximum includes $i = l$, so a true segment of length one contributes zero. The quantity $\mathfrak{v}$ measures the dispersion within the observed true segments, $\mathfrak{e}_j(x)$ is the corresponding upper deviation radius for the jth slope, and $\underline{c}_j(x)$ is its induced lower calibration value. The positive part is the one specified in Theorem 2. For the **(H2)** adjustment, write

$$\mathfrak{t} = \begin{cases} 0, & \text{under (H1)}, \\ \dfrac{t_{\min}}{n(t_{\max} - t_{\min})}, & \text{under (H2)}. \end{cases}$$

It follows from (7) and the original penalty formula that $\mathfrak{p}(T_K) = \widehat{c}_1 x_1(K) + (\widehat{c}_2 - \mathfrak{t}) x_2(K)$.

Let $\mathcal{C}_x$ be the event on which the orbit averaged response curve satisfies

$$\underline{c}_1(x) x_1(K) + \{\underline{c}_2(x) - \mathfrak{t}\} x_2(K) \ge \frac{4\mathfrak{c}(T_K)}{n}, \qquad K = 0, \ldots, K_{\max}. \tag{25}$$

In **(H2)**, include $t_{\max} > t_{\min}$ in the definition of $\mathcal{C}_x$. The right hand side depends only on K, so (25) uses any feasible configuration with that number of changes. This condition places the penalty induced by the orbit averaged slopes above the sufficient bound after accounting for their quantified fluctuations, retaining the joint contribution of both design columns.

Theorem 4 *Suppose Assumptions 1 and 2 hold, and fix a weighting scheme and $x, y > 0$. The probability that the sample belongs to $\mathcal{C}_x$, the coefficients fitted by Algorithm 3 satisfy (10), and its selected configuration satisfies the oracle inequality (11) is at least*

$$(1 - e^{-x})\mathbb{P}(\mathcal{C}_x) - e^{-y}. \tag{26}$$

In particular, if the sampling distribution and the algorithmic inputs satisfy $\mathbb{P}(\mathcal{C}_x) \ge 1 - \epsilon$ for a specified $\epsilon \in [0, 1]$, then this oracle inequality holds with probability at least $(1-\epsilon)(1-e^{-x}) - e^{-y}$.

Remark 10 *Theorem 4 connects the two coefficients fitted by the original slope regression to the oracle risk bound for the sample dependent kernel. On the simultaneous concentration event of Lemma 6, the fitted slopes yield $\widehat{c}_j \ge \underline{c}_j(x)$. Intersecting this event with the calibration event (25) supplies the effective penalty required by Theorem 3. Invariance of the calibration event $\mathcal{C}_x$ under permutations within true segments gives the probability bound (26). The condition $\mathbb{P}(\mathcal{C}_x) \ge 1 - \epsilon$ specifies the sampling distributions and algorithmic inputs covered by this calibration guarantee. True segments enter the probability analysis through the finite group averages; the estimator uses the original response, fitting range, multiplier α, and positive part operation in Algorithm 3.*

The coefficientwise conditions in the remark after Theorem 3 imply (25) when $\widehat{c}_j$ is replaced by $\underline{c}_j(x)$, $j = 1, 2$. The joint condition (25) also permits the two design columns to compensate for each other. Appendix A.1 derives the concentration bounds and combines their simultaneous event with $\mathcal{C}_x$.

Lemma 5 *Under Assumption 2, every fitted response in Algorithm 3 is finite on the domain of the selected weighting scheme, and the design matrix $\mathcal{X}$ in Section A.1.2 has column rank three. Its original OLS coefficients consequently have the weight representation given there.*

Remark 11 *Lemma 5 identifies a checkable condition on the original fitting range that determines all three regression coefficients uniquely. The strict curvature of the logarithmic model count distinguishes its design column from the intercept and the linear segment count. The representation through $\mathcal{W}$ makes each original fitted slope a fixed linear combination of the jointly observed response values.*

Proof. The initial feasibility update in Algorithm 3 ensures $(K+1)\ell_{\min} \le n$ for $0 \le K \le K_{\max}$. For each such K, use K segments of length $\ell_{\min}$ followed by a final segment of length $n - K\ell_{\min} \ge \ell_{\min}$. This gives a feasible configuration. All within sample distances are finite. Under **(H1)**, each original reverse rank count is between zero and n, and every feasible cost in Algorithm 1 is therefore finite. Under **(H2)**, its original segment costs are finite sums of finite distances, and its scaling constant is finite on $t_{\max} > t_{\min}$. The dynamic program takes minima over finite sets containing a finite feasible cost, so each of its final optimized responses is finite.

The number of fitted rows can be read directly from the original inputs. For a nonempty feasible model range, $K_{\max} \ge 0$ and $0 < \mathrm{frac} < 1$ imply

$$1 \le \lceil \mathrm{frac}(K_{\max}+1) \rceil \le K_{\max} + 1.$$

Consequently, the maximum with zero in the original definition of K_0 leaves the value $K_0 = \lceil \mathrm{frac}(K_{\max}+1) \rceil - 1$, and the number of rows is

$$K_{\max} - K_0 + 1 = K_{\max} + 2 - \lceil \mathrm{frac}(K_{\max}+1) \rceil .$$

Having at least three rows is therefore equivalent to

$$\lceil \mathrm{frac}(K_{\max}+1) \rceil \le K_{\max} - 1.$$

The right hand side is an integer, so this inequality is equivalent to $\mathrm{frac}(K_{\max}+1) \le K_{\max} - 1$. Since the left hand side is positive, this yields the input form

$$K_{\max} \ge 2, \qquad 0 < \mathrm{frac} \le \frac{K_{\max}-1}{K_{\max}+1}.$$

For example, at the input $\mathrm{frac} = 3/5$, it is equivalent to $3(K_{\max}+1) \le 5(K_{\max}-1)$, or $K_{\max} \ge 4$. All these statements refer to the updated $K_{\max}$ of Algorithm 3. Three is also the row count required by column rank three, since the span of the rows has dimension at most the number of rows. The following strict curvature argument establishes column rank three throughout this input range.

There are at least three consecutive rows in the fitting range by Assumption 2. Suppose that real numbers a, b, c give a zero linear combination of the three design columns. This means

$$a + b\,x_1(K) + c\,x_2(K) = 0, \qquad K_0 \le K \le K_{\max}.$$

Use the first three rows and put $h = K_0 + 1$, their middle index. They satisfy $1 \le h \le n-2$ because all three model sizes lie between zero and $n-1$. From the factorial expression for the binomial coefficient,

$$\frac{\binom{n-1}{h+1}}{\binom{n-1}{h}} = \frac{n-1-h}{h+1}, \qquad \frac{\binom{n-1}{h}}{\binom{n-1}{h-1}} = \frac{n-h}{h}.$$

Subtracting the logarithms of these two identities and using the design definition in Algorithm 3 gives

$$x_1(h+1) - 2x_1(h) + x_1(h-1) = \frac{1}{n} \log \frac{h(n-1-h)}{(h+1)(n-h)} < 0.$$

To verify the strict sign, both numerator and denominator are positive, and the denominator minus the numerator is $(h+1)(n-h) - h(n-1-h) = n > 0$. The same second difference of the constant column and of $x_2(K) = (K+1)/n$ is zero. Taking the second difference in the supposed zero linear combination therefore gives $b = 0$. Taking the first difference of the two remaining terms gives $c/n = 0$, so $c = 0$. Substituting into any row then gives $a = 0$. This proves column independence.

For a nonzero column vector $u \in \mathbb{R}^3$, column independence gives $\mathcal{X}u \ne 0$, and hence $u^\top \mathcal{X}^\top \mathcal{X} u = \|\mathcal{X}u\|_2^2 > 0$. Thus $\mathcal{X}^\top \mathcal{X}$ is invertible. Here $\|\cdot\|_2$ is the usual Euclidean norm. To obtain the original

OLS solution, differentiate its sum of squared residuals with respect to its three coefficients. The gradient at a column vector u is

$$2\mathcal{X}^\top \mathcal{X}u - 2\mathcal{X}^\top(\widehat{\mathcal{R}}(K_0+1),\ldots,\widehat{\mathcal{R}}(K_{\max}+1))^\top.$$

Setting this gradient to zero gives $u = \mathcal{W}(\widehat{\mathcal{R}}(K_0+1),\ldots,\widehat{\mathcal{R}}(K_{\max}+1))^\top$. The positive definite Hessian $2\mathcal{X}^\top\mathcal{X}$ makes this zero the unique minimizer. By the definition of $\mathcal{W}$ in Section A.1.2, its jth component is the stated sum with weights $\mathcal{W}_{jK}$.

The endpoint contribution can also be followed through the original regression. Use $\mathfrak{t}$ from Section A.1.2. Under **(H2)**,

$$\begin{aligned}\kappa_{\mathrm{H2}} t_{\min}(n-K-1) &= \frac{t_{\min}}{n(t_{\max}-t_{\min})}\left(1-\frac{K+1}{n}\right)\\ &= \mathfrak{t}\{1-x_2(K)\}.\end{aligned}$$

Under **(H1)** the corresponding endpoint term is zero, consistent with $\mathfrak{t}=0$. The original optimized response consequently satisfies, for either weighting scheme,

$$\widehat{\mathcal{R}}(K+1) = \min_{T_K\in\mathcal{P}} L_n(T_K) + \mathfrak{t}\{1-x_2(K)\}, \qquad K=0,\ldots,K_{\max},$$

where the minimum in each entry is over configurations with K boundaries. The multiplier $\mathfrak{t}$ depends on the realized full sample and has the same value at every fitted model size.

The definition of $\mathcal{W}$ gives $\mathcal{W}\mathcal{X} = (\mathcal{X}^\top\mathcal{X})^{-1}(\mathcal{X}^\top\mathcal{X})$, which acts as the identity on $\mathbb{R}^3$. Its action on the intercept column therefore gives

$$\sum_{K=K_0}^{K_{\max}} \mathcal{W}_{0K} = 1, \qquad \sum_{K=K_0}^{K_{\max}} \mathcal{W}_{1K} = 0, \qquad \sum_{K=K_0}^{K_{\max}} \mathcal{W}_{2K} = 0.$$

Its action on the column with entries $x_2(K)$ similarly gives

$$\sum_{K=K_0}^{K_{\max}} \mathcal{W}_{0K}x_2(K) = 0, \qquad \sum_{K=K_0}^{K_{\max}} \mathcal{W}_{1K}x_2(K) = 0, \qquad \sum_{K=K_0}^{K_{\max}} \mathcal{W}_{2K}x_2(K) = 1.$$

Apply these identities to the displayed response decomposition and use the original OLS weight representation. Subtracting its intercept and second column contributions yields

$$\begin{aligned}\sum_{K=K_0}^{K_{\max}} \mathcal{W}_{0K} \min_{T_K\in\mathcal{P}} L_n(T_K) &= \widehat{\beta}_0 - \mathfrak{t},\\ \sum_{K=K_0}^{K_{\max}} \mathcal{W}_{1K} \min_{T_K\in\mathcal{P}} L_n(T_K) &= \widehat{\beta}_1,\\ \sum_{K=K_0}^{K_{\max}} \mathcal{W}_{2K} \min_{T_K\in\mathcal{P}} L_n(T_K) &= \widehat{\beta}_2 + \mathfrak{t}.\end{aligned}$$

Each minimum again uses its displayed fixed K. These are identities for the coefficients already computed from $\widehat{\mathcal{R}}$ by Algorithm 3. Its original positive part remains $\widehat{c}_j = (-\alpha\widehat{\beta}_j)_+$, and its effective penalty is $\widehat{c}_1 x_1(K) + (\widehat{c}_2 - \mathfrak{t})x_2(K)$.

Lemma 6 *Under Assumptions 1 and 2, use the OLS design $\mathcal{X}$, its weight matrix $\mathcal{W}$, the finite permutation group $\mathfrak{S}$, the orbit averaged slopes $\overline{\beta}_j$, and the quantities $\mathfrak{v}$ and $\mathfrak{e}_j$ defined in Section A.1.2. The design matrix $\mathcal{X}$ has column rank three by Lemma 5. For every $x>0$, under **(H1)**,*

$$\mathbb{P}\Big\{\widehat{\beta}_j > \overline{\beta}_j + \mathfrak{e}_j(x) \text{ for some } j\in\{1,2\}\Big\} \le e^{-x}. \tag{27}$$

*Under **(H2)**, the same bound holds for the intersection of the event in braces with the original sample domain $t_{\max} > t_{\min}$.*

Remark 12 *Lemma 6 gives a simultaneous upper bound for the two original OLS slopes. The empirical anchor measure, the empirical radius distributions in **(H1)**, and both radius endpoints in **(H2)** are invariant under permutations within the true segments. This property permits a finite permutation calculation while preserving their full sample dependence. The quantity $\mathfrak{v}$ records the dispersion of the observations Y_i within the true segments and satisfies $0 \le \mathfrak{v} \le n - K^\star - 1$. The lemma therefore supplies both the stated sample dispersion bound and its universal version obtained by replacing $\mathfrak{v}$ with $n - K^\star - 1$. The response vector is the one computed by Algorithm 3, including its original **(H2)** normalization.*

Proof. Assumption 2 gives at least three fitted model sizes, so $K_{\max} \ge 2$. The feasibility update and $\ell_{\min} \ge 2$ then give $n \ge (K_{\max}+1)\ell_{\min} \ge 6$. Hence $1 - 2/n > 0$. Fix a realization of $X_1, \dots, X_n$, taking a realization in the domain $t_{\max} > t_{\min}$ for **(H2)**. We first verify the invariance used in the finite permutation calculation. For every $\varsigma \in \mathfrak{S}$ and every deterministic anchor w and radius t, reindexing finite sums gives

$$\frac{1}{n}\sum_{i=1}^{n} \delta_{X_{\varsigma(i)}} = \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i}, \qquad \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}\{d(w, X_{\varsigma(i)}) \le t\} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}\{d(w, X_i) \le t\}.$$

These are the empirical anchor measure and the full sample radius distribution from Section 2. Their invariance proves invariance of the complete **(H1)** measure H. A permutation also preserves the multiset of all pairwise distances between distinct observation indices. It therefore preserves $t_{\min}$, $t_{\max}$, their difference, and the radius probability measure in **(H2)**. Combined with the first finite sum identity, this proves invariance of the complete **(H2)** measure as well. Consequently, within this finite orbit, the observations Y_i of Section 3.1 are permuted in the same realized Hilbert space, and all its norms use the original full sample measure H.

We next evaluate the change in the loss when two observation positions i, ℓ in the same true segment are interchanged. Fix a candidate partition T_K and one anchor and radius. Its profile coordinates are the binary values $\mathbb{I}\{d(w, X_l) \le t\}$, for $l = 1, \dots, n$. If the two positions belong to the same candidate segment, the interchange preserves its count of ones and its residual sum of squares. The same is true if their two binary values agree.

Consider the remaining case, with the value zero at i and the value one at ℓ, and with the two positions in different candidate segments. Let their lengths be r, s and their counts of ones before the interchange be a, b, respectively. These numbers satisfy

$$0 \le a \le r - 1, \qquad 1 \le b \le s.$$

For a binary segment of length r with a ones, its mean is a/r. Expanding its squared deviations from that mean gives

$$a(1 - a/r)^2 + (r - a)(a/r)^2 = a - \frac{a^2}{r}.$$

After the interchange, the two counts become $a + 1$ and $b - 1$. The change in the sum of the two residual sums of squares is therefore

$$\left(a + 1 - \frac{(a+1)^2}{r} + b - 1 - \frac{(b-1)^2}{s}\right) - \left(a - \frac{a^2}{r} + b - \frac{b^2}{s}\right)$$
$$= -\frac{2a+1}{r} + \frac{2b-1}{s}.$$

The constraints on the counts imply

$$\frac{1}{r} \le \frac{2a+1}{r} \le 2 - \frac{1}{r}, \qquad \frac{1}{s} \le \frac{2b-1}{s} \le 2 - \frac{1}{s}.$$

The absolute difference of these two fractions is at most $2 - 1/r - 1/s$. The two candidate segments are disjoint, so $r + s \le n$. Moreover,

$$\frac{1}{r} + \frac{1}{s} - \frac{4}{r+s} = \frac{(r-s)^2}{rs(r+s)} \ge 0.$$

All denominators are positive. Consequently,

$$\frac{1}{r} + \frac{1}{s} \ge \frac{4}{r+s} \ge \frac{4}{n}, \qquad 2 - \frac{1}{r} - \frac{1}{s} \le 2 - \frac{4}{n} = 2\left(1 - \frac{2}{n}\right).$$

Interchanging the labels of the two positions covers the case with a one at i and a zero at ℓ. When both positions lie in the same candidate segment or their binary values agree, the change is zero. Since $1 - 2/n > 0$, these cases satisfy the same upper bound. Combining all cases, the pointwise change in the total residual sum of squares is bounded in absolute value by

$$2\left(1 - \frac{2}{n}\right) \left|\mathbb{I}\{d(w, X_i) \le t\} - \mathbb{I}\{d(w, X_\ell) \le t\}\right|.$$

A difference of binary values takes values in $\{-1, 0, 1\}$, so its absolute value equals its square. Integration against the same H before and after the interchange gives

$$\int \left|\mathbb{I}\{d(w, X_i) \le t\} - \mathbb{I}\{d(w, X_\ell) \le t\}\right| H(dw, dt) = \|Y_i - Y_\ell\|_{\mathcal{H}}^2.$$

The normalization in L_n is $1/n$. Thus the absolute change in the loss of every fixed candidate partition is at most

$$\frac{2}{n}\left(1 - \frac{2}{n}\right) \|Y_i - Y_\ell\|_{\mathcal{H}}^2.$$

This bound is uniform in the candidate partition.

Taking the minimum over the same finite feasible candidate set preserves the bound. Indeed, if every pair of corresponding objective values before and after the interchange differs by at most a number $c \ge 0$, then every value after the interchange is at least the corresponding value before it minus c. Taking minima gives the same lower bound for the minima. Applying this argument with the two samples in reversed order gives the upper bound. Hence the two minimum values differ in absolute value by at most c.

Under **(H1)**, the loss and normalized algorithmic cost agree by (5). Under **(H2)**, (6) gives, for every fixed K,

$$\widehat{\mathcal{R}}(K+1) = \min_{T_K \in \mathcal{P}} L_n(T_K) + \kappa_{\mathrm{H2}} t_{\min}\{n - (K+1)\}.$$

Here the minimum is taken over the configurations in $\mathcal{P}$ with K boundaries, as in Algorithm 3. Both κ_{H2} and $t_{\min}$ are invariant under the interchange, so the second term is identical for the two samples. This proves that the original response $\widehat{\mathcal{R}}(K+1)$, in either weighting scheme, has the same interchange bound.

By the OLS identity in Section A.1.2, the fitted slope $\widehat{\beta}_j$ is the inner product of row $\mathcal{W}_j$ with the original response vector. Summing the response bounds with the absolute values of the row entries shows that an interchange of i, ℓ changes this slope by at most

$$\frac{2\|\mathcal{W}_j\|_1}{n}\left(1 - \frac{2}{n}\right) \|Y_i - Y_\ell\|_{\mathcal{H}}^2, \qquad j = 1, 2.$$

In particular, for an interchange of two positions inside the true segment $(t_k^\star, t_{k+1}^\star]$, it changes the slope by at most

$$\frac{2\|\mathcal{W}_j\|_1}{n}\left(1 - \frac{2}{n}\right) \max_{t_k^\star < i, l \le t_{k+1}^\star} \|Y_i - Y_l\|_{\mathcal{H}}^2.$$

The maxima in this expression are themselves invariant under the true segment permutations. Each squared distance between two observations Y_i and Y_l is at most one because their indicator values belong to $\{0, 1\}$ and H has total mass one. The definition of $\mathfrak{v}$ in Section A.1.2 thus gives

$$0 \le \mathfrak{v} \le \sum_{k=0}^{K^\star} (t_{k+1}^\star - t_k^\star - 1) = n - K^\star - 1.$$

For the fixed sample, choose ς uniformly from the finite group $\mathfrak{S}$, independently of all sample coordinates. Write $\mathbb{E}_\varsigma$ and $\mathbb{P}_\varsigma$ for expectation and probability in this additional finite randomization. This is equivalent to choosing independent uniform permutations in the different true segments. Reveal their images one true segment at a time. In a true segment of length $t_{k+1}^\star - t_k^\star$, reveal all but the last image, since the last one is determined by the others. A segment of length one requires zero random reveals.

We justify the range bound for the conditional expectations at each reveal. Fix its past and compare two possible next labels. For any completion with the first label as the next image, exchange the two compared labels in this completion. The resulting completion has the second label as its next image. This transformation is a bijection between the two sets of completions, and all the completions within each set have the same probability. The two fully permuted samples differ by an interchange of two positions in the same true segment. Their fitted slopes therefore differ by at most the true segment bound just proved. Averaging over the matched completions shows that the corresponding conditional expectations differ by at most the same quantity. Consequently, the conditional expectation of the final fitted slope, viewed as a function of the next revealed label, has range length at most this quantity.

Subtracting its conditional mean gives the Doob martingale difference for this reveal and preserves the range length. The centered bounded interval exponential estimate proved in Lemma 1 applies for every real parameter η. At each reveal in true segment k, it supplies the exponential factor

$$\exp\left\{\frac{\eta^2(1-2/n)^2\|\mathcal{W}_j\|_1^2}{2n^2}\left(\max_{t_k^\star<i,l\le t_{k+1}^\star}\|Y_i-Y_l\|_{\mathcal{H}}^2\right)^2\right\}.$$

There are $t_{k+1}^\star-t_k^\star-1$ reveals in that segment. Iterating conditional expectations over all the reveals multiplies their exponential factors. The sum of the centered differences telescopes to the fitted slope of the permuted sample minus its uniform orbit average. The latter is $\overline{\beta}_j$ by its definition in Section A.1.2. The weights $\mathcal{W}_{jK}$ are fixed throughout this permutation calculation, so interchanging the two finite sums gives

$$\begin{aligned}\mathbb{E}_\varsigma\widehat{\beta}_j^\varsigma &= \frac{1}{|\mathfrak{S}|}\sum_{\varsigma\in\mathfrak{S}}\sum_{K=K_0}^{K_{\max}}\mathcal{W}_{jK}\widehat{\mathcal{R}}^\varsigma(K+1)=\sum_{K=K_0}^{K_{\max}}\mathcal{W}_{jK}\left\{\frac{1}{|\mathfrak{S}|}\sum_{\varsigma\in\mathfrak{S}}\widehat{\mathcal{R}}^\varsigma(K+1)\right\}\\ &=\sum_{K=K_0}^{K_{\max}}\mathcal{W}_{jK}\overline{\mathcal{R}}(K+1)=\overline{\beta}_j.\end{aligned}$$

Before any images are revealed, the conditional expectation is this uniform average. After all the stated reveals, every permutation image is determined, so the final conditional expectation equals $\widehat{\beta}_j^\varsigma$. The intermediate conditional expectations cancel in the sum of consecutive differences, leaving $\widehat{\beta}_j^\varsigma-\overline{\beta}_j$. Summing the exponents therefore gives

$$\mathbb{E}_\varsigma\exp\{\eta(\widehat{\beta}_j^\varsigma-\overline{\beta}_j)\}\le\exp\left\{\frac{\eta^2(1-2/n)^2\|\mathcal{W}_j\|_1^2\mathfrak{v}}{2n^2}\right\},\qquad\eta\in\mathbb{R},\tag{28}$$

where $\widehat{\beta}_j^\varsigma$ denotes the original OLS slope computed from the permuted sample, in accordance with the orbit averages of Section A.1.2.

Suppose first that $\|\mathcal{W}_j\|_1^2\mathfrak{v}>0$. For a positive threshold u and $\eta>0$, exponential Markov inequality and (28) give

$$\mathbb{P}_\varsigma\{\widehat{\beta}_j^\varsigma-\overline{\beta}_j>u\}\le\exp\left\{-\eta u+\frac{\eta^2(1-2/n)^2\|\mathcal{W}_j\|_1^2\mathfrak{v}}{2n^2}\right\}.$$

The coefficient of η^2 is positive. Differentiating this quadratic exponent gives

$$-u+\frac{\eta(1-2/n)^2\|\mathcal{W}_j\|_1^2\mathfrak{v}}{n^2}.$$

Its zero and the corresponding minimum exponent are

$$\eta=\frac{n^2u}{(1-2/n)^2\|\mathcal{W}_j\|_1^2\mathfrak{v}},\qquad-\frac{n^2u^2}{2(1-2/n)^2\|\mathcal{W}_j\|_1^2\mathfrak{v}}.$$

For $u=\mathfrak{e}_j(x)$ from Section A.1.2, the squared threshold is

$$u^2=\frac{(1-2/n)^2\|\mathcal{W}_j\|_1^2}{n^2}2\mathfrak{v}(x+\log 2).$$

Substitution into the minimum exponent cancels these factors and gives $-(x + \log 2)$. Thus

$$\mathbb{P}_\varsigma\{\widehat{\beta}_j^\varsigma > \overline{\beta}_j + \mathfrak{e}_j(x)\} \le e^{-(x+\log 2)} = \frac{1}{2}e^{-x}.$$

If $\|\mathcal{W}_j\|_1 = 0$, its fitted slope and its orbit average are both zero. If $\mathfrak{v} = 0$, every true segment of length at least two has zero feature diameter. Every permitted interchange then leaves the slope unchanged. Every permutation of a finite set is a product of interchanges: one may place its first desired image by an interchange, then its second image, and continue until all positions agree. Applied inside each true segment, this establishes that the slope is constant on the entire orbit. Hence in both zero cases the slope equals its orbit average, $\mathfrak{e}_j(x) = 0$, and the event in the preceding display is empty. A union bound for the two slopes now gives, for every fixed sample in the domain,

$$\mathbb{P}_\varsigma\Big\{\widehat{\beta}_j^\varsigma > \overline{\beta}_j + \mathfrak{e}_j(x) \text{ for some } j \in \{1,2\}\Big\} \le e^{-x}.$$

It remains to transfer this finite permutation statement to the original sample law. Assumption 1 makes the coordinate laws identical within each true segment and independent across all indices. Thus reindexing their product law inside any true segment preserves the joint law. Applying this successively to the true segments shows that

$$(X_{\varsigma(1)}, \dots, X_{\varsigma(n)}) \quad \text{has the same law as} \quad (X_1, \dots, X_n)$$

for every deterministic $\varsigma \in \mathfrak{S}$. The group average $\overline{\beta}_j$ is invariant because composition with any fixed permutation simply reindexes the finite sum over $\mathfrak{S}$. The already established invariance of the empirical measure and the within true segment maxima proves invariance of $\mathfrak{v}$, and therefore of $\mathfrak{e}_j(x)$. In **(H2)**, the domain event is invariant as well.

Under **(H1)**, apply the preceding equality of laws to the indicator of the slope event in (27) and average over all deterministic permutations. Finite summation may be interchanged with expectation. Using the invariance of the orbit average and the radius then gives

$$\begin{aligned}
&\mathbb{P}\Big\{\widehat{\beta}_j > \overline{\beta}_j + \mathfrak{e}_j(x) \text{ for some } j \in \{1,2\}\Big\} \\
&= \mathbb{E}\left[\frac{1}{|\mathfrak{S}|}\sum_{\varsigma\in\mathfrak{S}} \mathbb{I}\left\{\widehat{\beta}_j^\varsigma > \overline{\beta}_j + \mathfrak{e}_j(x) \text{ for some } j \in \{1,2\}\right\}\right] \\
&\le \mathbb{E}[e^{-x}] = e^{-x}.
\end{aligned}$$

The finite average in the middle line is the probability over the additional uniform permutation, bounded above in the preceding calculation. Under **(H2)**, the same equality includes the invariant indicator $\mathbb{I}\{t_{\max} > t_{\min}\}$ inside the expectation. On its domain the finite average is at most e^{-x}, so the result is at most $e^{-x}\mathbb{P}(t_{\max} > t_{\min}) \le e^{-x}$. This proves the claimed result for the original observations and their original full sample weights.

Proof of Theorem 4. By Lemma 5, the matrix $\mathcal{W}$ and all fitted coefficients are well defined on the scheme's domain. At any sample in this domain for which

$$\widehat{\beta}_j \le \overline{\beta}_j + \mathfrak{e}_j(x), \qquad j = 1, 2,$$

multiplication by $-\alpha < 0$ and monotonicity of the positive part give

$$\widehat{c}_j = (-\alpha\widehat{\beta}_j)_+ \ge \big(-\alpha\{\overline{\beta}_j + \mathfrak{e}_j(x)\}\big)_+ = \underline{c}_j(x), \qquad j = 1, 2.$$

For each feasible K, the integer $\binom{n-1}{K}$ is at least one, so $x_1(K) \ge 0$; also $x_2(K) = (K+1)/n > 0$. Consequently,

$$\begin{aligned}
\mathfrak{p}(T_K) &= \widehat{c}_1 x_1(K) + (\widehat{c}_2 - \mathfrak{t})x_2(K) \\
&\ge \underline{c}_1(x)x_1(K) + (\underline{c}_2(x) - \mathfrak{t})x_2(K).
\end{aligned}$$

On $\mathcal{C}_x$, (25) bounds the last line below by $4\mathfrak{c}(T_K)/n$ for every feasible configuration. Thus (10) holds there whenever both displayed slope bounds hold. Theorem 3 applies to these finite measurable fitted penalties on its common event, and gives (11) at every sample in this intersection.

We now calculate the probability of this intersection. The quantities $\overline{\beta}_j$ are invariant under every permutation in $\mathfrak{S}$, since composition with a fixed permutation reindexes their defining finite average. The empirical measure H and the sets of feature distances within the true segments are invariant by Lemma 6, so $\mathfrak{v}$ and $\mathfrak{e}_j(x)$ are invariant. Both **(H2)** endpoints and $\mathfrak{t}$ are invariant as well. The deterministic design columns and the complexity $\mathfrak{c}(T_K)$ depend only on the model size. Every inequality defining $\mathcal{C}_x$ in (25) consequently has the same truth value before and after an allowed permutation, proving invariance of $\mathcal{C}_x$.

Apply the finite averaging calculation in the proof of Lemma 6 after multiplying each event indicator by $\mathbb{I}\{\mathcal{C}_x\}$, the indicator that the sample belongs to $\mathcal{C}_x$. Invariance of the joint law under the true segment permutations, and then invariance of $\mathcal{C}_x$, give

$$\begin{aligned}
&\mathbb{P}\Big(\mathcal{C}_x \cap \Big\{\widehat{\beta}_j > \overline{\beta}_j + \mathfrak{e}_j(x) \text{ for some } j \in \{1,2\}\Big\}\Big) \\
&= \mathbb{E}\left[\mathbb{I}\{\mathcal{C}_x\}\frac{1}{|\mathfrak{S}|}\sum_{\varsigma\in\mathfrak{S}}\mathbb{I}\left\{\widehat{\beta}^{\varsigma}_j > \overline{\beta}_j + \mathfrak{e}_j(x) \text{ for some } j \in \{1,2\}\right\}\right] \\
&\leq \mathbb{E}[\mathbb{I}\{\mathcal{C}_x\}e^{-x}] = e^{-x}\mathbb{P}(\mathcal{C}_x).
\end{aligned}$$

The inequality uses the bound for every fixed sample and its uniform random permutation established in the proof of (27). Under **(H2)**, membership in $\mathcal{C}_x$ includes $t_{\max} > t_{\min}$, so all quantities inside its indicator have their defined values; the product with that indicator is assigned zero on the complement of this domain. Subtracting the displayed probability from $\mathbb{P}(\mathcal{C}_x)$ shows that the calibration condition and both slope bounds hold together with probability at least $(1-e^{-x})\mathbb{P}(\mathcal{C}_x)$. The exceptional event of Theorem 3 has probability at most e^{-y}, so their intersection with its common event has probability at least

$$(1-e^{-x})\mathbb{P}(\mathcal{C}_x) - e^{-y}.$$

Every sample in this intersection satisfies (10) and (11), proving (26). Finally, substituting $\mathbb{P}(\mathcal{C}_x) \geq 1-\epsilon$ gives $(1-\epsilon)(1-e^{-x}) - e^{-y}$ because $1-e^{-x} > 0$.

A.2 Simulation Data Generation

A.2.1 Simulation Design: Common Settings

Across all experiments, we generated independent sequences of length $n = 400$ with a fixed number of change points $K^\star = 3$ located at equally spaced positions $t_1^\star = \lfloor n/4 \rfloor$, $t_2^\star = \lfloor n/2 \rfloor$, and $t_3^\star = \lfloor 3n/4 \rfloor$ (hence four segments of equal length). For each scenario and each signal level, results were computed from 500 Monte Carlo replications, where in each replication a fresh sequence was generated with an independent random seed.

The signal-to-noise ratio (SNR) was controlled by a scalar parameter $\rho \in [0,1]$ that modulates the magnitude of distributional differences across segments. In all scenarios, ρ enters through $\theta^{(k)}(\rho) = (1-\rho)\,\theta^{(1)} + \rho\,\theta^{(k),\mathrm{tar}}$, $k \in \{2,3,4\}$, where $\theta^{(1)}$ denotes the baseline (segment 1) parameter and $\theta^{(k),\mathrm{tar}}$ denotes the target parameter for segment k under a strong-change regime. In addition, in the density-valued and spherical scenarios, ρ also controls the variability within segments, so that smaller values of ρ correspond to weaker contrasts between segments and greater heterogeneity within segments. For each generated sequence, we construct the pairwise distance matrix $\mathbf{D} = (D_{ij})_{1\leq i,j\leq n}$, where $D_{ij} = d(X_i, X_j)$ and the metric d is chosen according to the data type, as detailed in Section A.2. This distance matrix was the common input used by all procedures compared in the simulation study.

A.2.2 Tuning parameters

For MCPD, we adopt the same default parameters as in their Algorithm 1 and set the number of permutations to 1000. For KCP, we use the Gaussian kernel by default. In the spherical setting of Section A.2.6, we employ the Laplacian kernel, motivated by the fact that Gaussian kernels constructed from geodesic distances on spheres are generally not positive definite, whereas Laplacian kernels based on geodesic distances can retain positive definiteness under suitable conditions Feragen et al. (2015). For the kernel bandwidth in KCP, we follow the recommendation for the multiple change-point setting $h = \mathrm{median}\{d(X_i, X_{i+1}) : 1 \leq i \leq n-1\}$. The penalty coefficients c_1

and c_2 are selected by the slope fit displayed in Algorithm 3. For this calibration, we set $\alpha = 2$ following the doubling rule of the slope heuristics described by Baudry et al. (2012), and $\mathrm{frac} = 0.6$ following the regression range used by Letzgus (2020, Section 4.2). Furthermore, since the MCPD procedure does not require specifying an upper bound on the number of change points, we likewise set, for both DPDCP and KCP, $K_{\max}$ to the largest value permitted by the minimal segment-length constraint.

A.2.3 Density-Valued Data: Univariate Distributions

In the density-valued setting, each observation X_i is a univariate probability distribution on $\mathbb{R}_+$, taken to be Gamma with a time-varying shape parameter. Specifically, within segment $S_k = \{t^\star_{k-1} + 1, \ldots, t^\star_k\}$, we generated a shape parameter $\alpha_i > 0$ according to a truncated Gaussian perturbation

$$\alpha_i = \max\{\alpha_{\min}, \widetilde{\alpha}_i\}, \qquad \widetilde{\alpha}_i \sim \mathcal{N}\big(\bar{\alpha}_k(\rho),\, \sigma^2_\alpha(\rho)\big),$$

where $\alpha_{\min} = 0.2$. The segment-specific mean shape is interpolated between a baseline value $\alpha_0 = 10$ and a "strong-change" target pattern

$$\alpha^{\mathrm{tar}} = (8, 12, 8, 12),\ \bar{\alpha}_k(\rho) = (1-\rho)\alpha_0 + \rho\, \alpha^{\mathrm{tar}}_k,\ k = 1, 2, 3, 4,$$

so that segments become increasingly distinguishable as ρ increases. To prevent changes in the first moment from dominating the experiment, the mean of the Gamma distribution was kept constant at $\mu_0 = 5$ by setting the rate parameter $\beta_i = \alpha_i/\mu_0$, so that $\mathbb{E}[X_i] = \alpha_i/\beta_i = \mu_0$ for all i.

The within-segment dispersion of the shape parameters was also modulated by ρ via

$$\sigma_\alpha(\rho) = (1-\rho)\sigma_{\mathrm{weak}} + \rho\, \sigma_{\mathrm{strong}},\ \sigma_{\mathrm{weak}} = 0.40,\ \sigma_{\mathrm{strong}} = 0.02,$$

so that weak-signal regimes combine small between-segment contrasts with substantial heterogeneity within segment.

Distances between distributions were computed using the 2-Wasserstein distance on $\mathbb{R}$, expressed through quantile functions. Let $Q_i(p)$ denote the quantile function of the Gamma distribution at time i. We evaluated Q_i on an equispaced grid $\{p_j\}_{j=1}^m$ over $(0, 1)$ with $m = 300$ (excluding a small neighborhood of 0 and 1 to avoid numerical instabilities), and approximated

$$d(X_i, X_j) = \left(\int_0^1 \big(Q_i(p) - Q_j(p)\big)^2\, dp \right)^{1/2}$$

by a trapezoidal rule on $\{p_j\}_{j=1}^m$.

A.2.4 Network Data: Stochastic Block Models

In the network setting, each observation X_i is an undirected simple graph with $N = 300$ nodes, represented by its (combinatorial) graph Laplacian $L_i = D_i - A_i$, where A_i is the adjacency matrix and D_i is the diagonal matrix of degrees. Graphs were generated from a stochastic block model (SBM) with segment-dependent connectivity parameters.

Segments 1–3 used $K_c = 3$ communities. The baseline (segment 1) SBM had community sizes $\pi^{(1)} = (100, 100, 100)$ and connectivity matrix

$$B^{(1)} = \begin{pmatrix} 0.2 & 0.001 & 0.001 \\ 0.001 & 0.2 & 0.001 \\ 0.001 & 0.001 & 0.2 \end{pmatrix}.$$

To control signal strength, we defined a strong-change target for the connectivity matrix,

$$B^{(2),\mathrm{tar}} = \begin{pmatrix} 0.8 & 0.001 & 0.001 \\ 0.001 & 0.2 & 0.001 \\ 0.001 & 0.001 & 0.8 \end{pmatrix},$$

and set, for segment 2,

$$B^{(2)}(\rho) = (1-\rho)B^{(1)} + \rho\, B^{(2),\mathrm{tar}},\ \pi^{(2)}(\rho) = \pi^{(1)}.$$

Segment 3 continued along the same direction by pushing $B^{(3)}(\rho)$ further toward $B^{(2),\mathrm{tar}}$ and simultaneously shrinking the community size vector toward the target $\pi^{(3),\mathrm{tar}} = (200, 50, 50)$, with $B^{(3)}(\rho) = (1-\rho)B^{(2)}(\rho) + \rho\, B^{(2),\mathrm{tar}}$, $\pi^{(3)}(\rho) \approx (1-\rho)\pi^{(1)} + \rho\, \pi^{(3),\mathrm{tar}}$, where the approximation is implemented by rounding to integer sizes and enforcing that the sizes sum to N.

Segment 4 reduced the number of communities to $K_c = 2$ by merging communities 2 and 3 of segment 3, yielding a two-block connectivity matrix $\widetilde{B}^{(4)}(\rho)$ and merged sizes $\widetilde{\pi}^{(4)}(\rho) = (\pi_1^{(3)}(\rho), \pi_2^{(3)}(\rho) + \pi_3^{(3)}(\rho))$. The merged connectivity matrix was obtained by the standard SBM aggregation formula: letting $w_2 = \pi_2^{(3)}(\rho)/(\pi_2^{(3)}(\rho) + \pi_3^{(3)}(\rho))$, $w_3 = 1 - w_2$, $\widetilde{B}_{11}^{(4)}(\rho) = B_{11}^{(3)}(\rho)$, $\widetilde{B}_{12}^{(4)}(\rho) = w_2 B_{12}^{(3)}(\rho) + w_3 B_{13}^{(3)}(\rho)$, and $\widetilde{B}_{22}^{(4)}(\rho) = w_2^2 B_{22}^{(3)}(\rho) + 2w_2 w_3 B_{23}^{(3)}(\rho) + w_3^2 B_{33}^{(3)}(\rho)$. We then interpolated toward a strong-change two-block target

$$B^{(4),\mathrm{tar}} = \begin{pmatrix} 0.5 & 0.01 \\ 0.01 & 0.5 \end{pmatrix}, \qquad \pi^{(4),\mathrm{tar}} = (200, 100),$$

by setting $B^{(4)}(\rho) = (1-\rho)\widetilde{B}^{(4)}(\rho) + \rho\, B^{(4),\mathrm{tar}}$, $\pi^{(4)}(\rho) \approx (1-\rho)\widetilde{\pi}^{(4)}(\rho) + \rho\, \pi^{(4),\mathrm{tar}}$, again with rounding and a sum constraint.

Conditional on community labels, edges were sampled independently as $A_{uv} \sim \mathrm{Bernoulli}(B^{(k)}_{g_u,g_v}(\rho))$ for $u < v$, with $A_{uu} = 0$ and $A_{vu} = A_{uv}$. In the main network experiments, node labels were kept aligned across time (i.e., community assignments were not randomly permuted across graphs). Moreover, when $\rho = 1$, the resulting data-generating mechanism is identical to that in Section 5.1 of Dubey & Zheng (2026).

Distances between networks were computed as Frobenius distances between Laplacians:

$$d(X_i, X_j) = \|L_i - L_j\|_F.$$

A.2.5 SPD Matrix Data: Wishart-type Covariance Models

In the SPD setting, each observation X_i is a $p \times p$ symmetric positive definite matrix with $p = 300$. Within segment S_k, we generated X_i as a ridge-regularized sample covariance matrix from a Gaussian model:

$$X_i = \frac{1}{\nu} Z_i^\top Z_i + \lambda I_p,\ Z_i \in \mathbb{R}^{\nu \times p} \text{ has i.i.d. rows } \mathcal{N}(0, \Sigma^{(k)}(\rho)),$$

with degrees of freedom $\nu = 240$ and ridge parameter $\lambda = 10^{-3}$. Equivalently, $\nu(X_i - \lambda I_p)$ follows a Wishart distribution with scale $\Sigma^{(k)}(\rho)$.

The baseline covariance was $\Sigma^{(1)} = I_p$. For segments $k \in \{2, 3, 4\}$, the covariance matrix was interpolated between I_p and a segment-specific target:

$$\Sigma^{(k)}(\rho) = (1-\rho) I_p + \rho\, \Sigma^{(k),\mathrm{tar}}.$$

The targets were chosen to encode qualitatively different correlation structures: (i) an AR(1)-type correlation with parameter -0.7 for segment 2, (ii) an AR(1)-type correlation with parameter $+0.7$ for segment 3, and (iii) an anisotropic diagonal target for segment 4. Concretely, letting $\left[\Sigma_{\mathrm{AR}}(\varrho)\right]_{ij} = \varrho^{|i-j|}$, we set $\Sigma^{(2),\mathrm{tar}} = \Sigma_{\mathrm{AR}}(-0.7)$ and $\Sigma^{(3),\mathrm{tar}} = \Sigma_{\mathrm{AR}}(0.7)$. For segment 4, we set $\Sigma^{(4),\mathrm{tar}} = \mathrm{diag}(d_1, \ldots, d_p)$, where

$$d_j = \exp(-a + 2a(j-1)/(p-1)),\ j = 1, \ldots, p,$$

with $a = 1.2$, and we rescale $(d_j)_{j=1}^p$ by its empirical mean so that $p^{-1}\sum_{j=1}^p d_j = 1$.

Distances between SPD matrices were computed using a Cholesky–Frobenius metric: writing $X_i = U_i^\top U_i$ for the upper-triangular Cholesky factor U_i,

$$d(X_i, X_j) = \|U_i - U_j\|_F.$$

A.2.6 Spherical Data without Temporal Dependence: Von Mises-Fisher Models

In the spherical setting, each observation X_i is a unit vector on the sphere $\mathbb{S}^{p-1} \subset \mathbb{R}^p$ with $p = 300$. Within each segment, observations were drawn i.i.d. from a von Mises–Fisher distribution $\mathrm{vMF}(\mu^{(k)}(\rho), \kappa(\rho))$.

Let e_1, e_2, e_3 denote the first three canonical basis vectors in $\mathbb{R}^p$ and let $\mu_0 = e_1$ be the baseline direction. Segment-specific target directions were taken as $\widetilde{\mu}^{(1)} = e_1$, $\widetilde{\mu}^{(2)} = e_2$, $\widetilde{\mu}^{(3)} = e_3$ and $\widetilde{\mu}^{(4)} = -e_1$. The mean direction in segment k was obtained by a normalized interpolation

$$\mu^{(k)}(\rho) = \frac{(1-\rho)\mu_0 + \rho\,\widetilde{\mu}^{(k)}}{\|(1-\rho)\mu_0 + \rho\,\widetilde{\mu}^{(k)}\|_2}, \qquad k = 1, 2, 3, 4.$$

The concentration parameter was simultaneously increased with ρ, $\kappa(\rho) = (1-\rho)\kappa_0 + \rho\,\kappa_1$, $\kappa_0 = 5$ and $\kappa_1 = 400$, so that weak-signal regimes correspond to both smaller directional changes between segment and larger dispersion within segment.

Distances between spherical observations were computed using the geodesic distance

$$d(X_i, X_j) = \arccos\big(X_i^\top X_j\big).$$

A.2.7 Spherical Data with Temporal Dependence: Geodesic AR(1) von Mises-Fisher Models

To assess the empirical robustness of DPDCP beyond the independence condition used in the theoretical analysis, we additionally considered a temporally dependent spherical change-point model. This simulation is designed as a stress test under serial dependence rather than as a setting covered by the oracle inequality.

Each observation X_i is a unit vector on the sphere $\mathbb{S}^{p-1} \subset \mathbb{R}^p$. We used $p = 300$ and generated sequences of length $n = 400$ with $K^\star = 3$ change points located at

$$t_1^\star = \lfloor n/4 \rfloor, \qquad t_2^\star = \lfloor n/2 \rfloor, \qquad t_3^\star = \lfloor 3n/4 \rfloor.$$

Let $t_0^\star = 0$ and $t_4^\star = n$, and define the four true segments by

$$S_k = \{t_{k-1}^\star + 1, \ldots, t_k^\star\}, \qquad k = 1, \ldots, 4.$$

Let e_1, e_2, e_3 denote the first three canonical basis vectors in $\mathbb{R}^p$, and let $\mu_0 = e_1$ be the baseline direction. The segment-specific target directions are

$$\widetilde{\mu}^{(1)} = e_1, \qquad \widetilde{\mu}^{(2)} = e_2, \qquad \widetilde{\mu}^{(3)} = e_3, \qquad \widetilde{\mu}^{(4)} = -e_1.$$

The signal strength is controlled by $\rho \in [0, 1]$. Since the fourth target direction is antipodal to the baseline direction, we use spherical linear interpolation along great-circle geodesics rather than Euclidean linear interpolation followed by normalization. For two non-antipodal unit vectors $a, b \in \mathbb{S}^{p-1}$, define

$$\mathrm{slerp}_\rho(a, b) = \frac{\sin\{(1-\rho)\theta\}}{\sin\theta} a + \frac{\sin(\rho\theta)}{\sin\theta} b, \qquad \theta = \arccos(a^\top b).$$

For the antipodal pair $(e_1, -e_1)$, we choose the deterministic orthogonal direction e_2 and set

$$\mathrm{slerp}_\rho(e_1, -e_1; e_2) = \cos(\pi\rho) e_1 + \sin(\pi\rho) e_2.$$

The segment-specific mean directions are then

$$\mu^{(1)}(\rho) = e_1, \qquad \mu^{(2)}(\rho) = \mathrm{slerp}_\rho(e_1, e_2),$$

$$\mu^{(3)}(\rho) = \mathrm{slerp}_\rho(e_1, e_3), \qquad \mu^{(4)}(\rho) = \mathrm{slerp}_\rho(e_1, -e_1; e_2).$$

Equivalently, since e_1, e_2, e_3 are mutually orthogonal,

$$\mu^{(2)}(\rho) = \cos\Big(\frac{\pi\rho}{2}\Big)\, e_1 + \sin\Big(\frac{\pi\rho}{2}\Big)\, e_2,$$

$$\mu^{(3)}(\rho) = \cos\Big(\frac{\pi\rho}{2}\Big)\, e_1 + \sin\Big(\frac{\pi\rho}{2}\Big)\, e_3,$$

and

$$\mu^{(4)}(\rho) = \cos(\pi\rho)e_1 + \sin(\pi\rho)e_2.$$

The von Mises–Fisher concentration parameter is also controlled by ρ:

$$\kappa(\rho) = (1-\rho)\kappa_0 + \rho\kappa_1, \qquad \kappa_0 = 5, \quad \kappa_1 = 400.$$

Thus, smaller values of ρ correspond to weaker directional separation and larger within-segment dispersion, whereas larger values of ρ correspond to stronger directional separation and tighter concentration around the segment-specific directions.

To introduce temporal dependence within each segment, we use a geodesic AR(1)-type von Mises–Fisher mechanism. Let $\phi_{\mathrm{AR}} \in [0, 1)$ denote the dependence parameter, which was set to $\phi_{\mathrm{AR}} = 0.5$ in our experiments. For each segment k, the first observation is generated as

$$X_{k,1} \sim \mathrm{vMF}_p\Big(\mu^{(k)}(\rho), \kappa(\rho)\Big),$$

where $\mathrm{vMF}_p(\mu, \kappa)$ denotes the von Mises–Fisher distribution on $\mathbb{S}^{p-1}$ with density proportional to $\exp\{\kappa\mu^\top x\}$.

For $s = 2, \ldots, |S_k|$, define the conditional center

$$c_{k,s} = \mathrm{Exp}_{\mu^{(k)}(\rho)}\left\{\phi_{\mathrm{AR}}\ \mathrm{Log}_{\mu^{(k)}(\rho)}(X_{k,s-1})\right\},$$

where Exp_μ and Log_μ are the Riemannian exponential and logarithm maps on the unit sphere. Conditional on the previous observation, we then generate

$$X_{k,s} \mid X_{k,s-1} \sim \mathrm{vMF}_p(c_{k,s}, \kappa(\rho)).$$

The full sequence is obtained by concatenating the four segmentwise sequences:

$$X_{t^\star_{k-1}+s} = X_{k,s}, \qquad s = 1, \ldots, |S_k|, \quad k = 1, \ldots, 4.$$

This construction induces Markovian temporal dependence within each segment, while the process is restarted at the beginning of each new segment.

For completeness, the exponential and logarithm maps used in the simulation are

$$\mathrm{Exp}_\mu(v) = \cos(\|v\|_2)\mu + \sin(\|v\|_2)\frac{v}{\|v\|_2}, \qquad v \in T_\mu\mathbb{S}^{p-1},$$

with $\mathrm{Exp}_\mu(0) = \mu$, and

$$\mathrm{Log}_\mu(x) = \frac{\theta}{\sin\theta}\left(x - \cos\theta\,\mu\right), \qquad \theta = \arccos(\mu^\top x),$$

with the usual limiting convention $\mathrm{Log}_\mu(\mu) = 0$.

Distances between spherical observations were computed using the geodesic distance

$$d(X_i, X_j) = \arccos(X_i^\top X_j).$$

A.3 Additional Simulation

Figure 4 presents density estimates of the empirical distribution of the estimated change-point locations for spherical data without temporal dependence at $\rho = 0.428$. Both DPDCP variants (**(H1)** and **(H2)**) produce sharply concentrated distributions with three well-separated modes that align with the true change-point locations, indicating accurate localization and low variability across replications.

Figure 5 and Figure 6 summarizes the change point detection performance under spherical data with and without temporal dependence, respectively. Although the latter setting introduces serial dependence within each segment and is not covered by the independence condition used in the oracle inequality, DPDCP still performs well across the considered signal levels. These results indicate that the proposed distance profile criterion remains robust for both independent and weakly dependent non Euclidean data.

For spherical data with temporal dependence, Table 2 shows that runtime increases with n, as expected, while DPDCP remains practically usable at $n = 2000$ and $n = 5000$. The MCPD result at $n = 5000$ is not reported because MCPD was already very slow at $n = 2000$, and therefore we did not further run it at $n = 5000$.

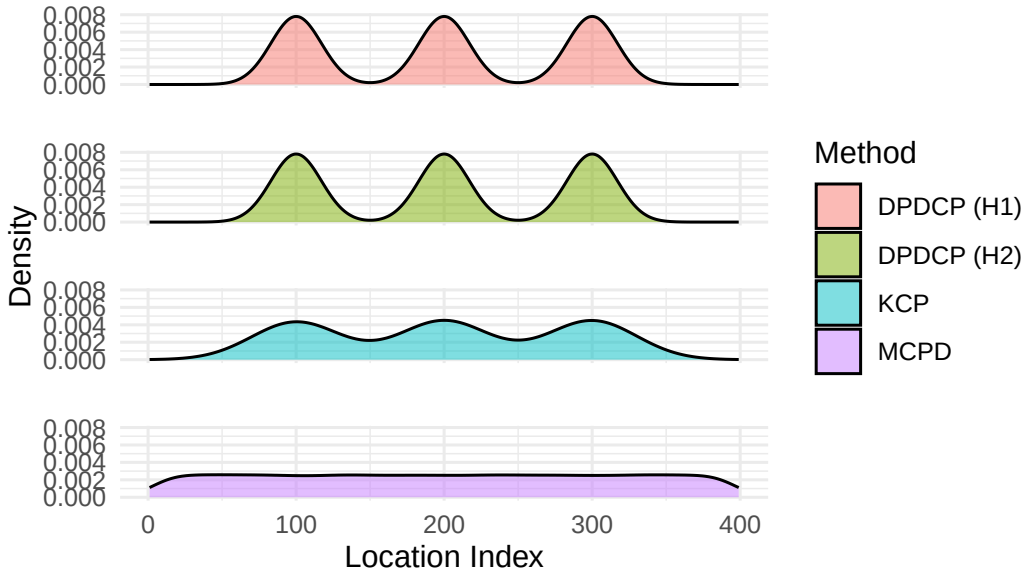

Figure 4: Estimated density functions of change-point locations for spherical data without temporal dependence at $\rho = 0.428$. For each method, the density summarizes the empirical distribution of estimated change-point locations obtained from 500 repeated experiments.

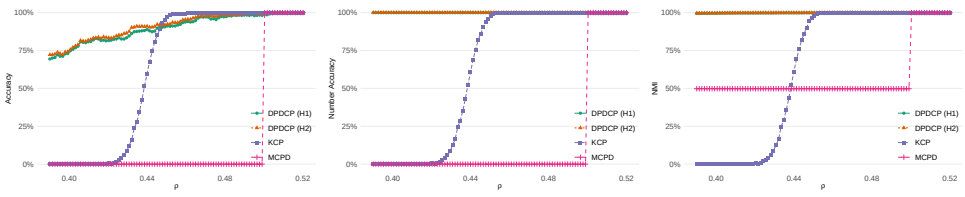

Figure 5: Performance comparison for spherical data without temporal dependence. The panels report, from left to right, the percentage of correctly estimating both the number and locations of change-points, the accuracy of estimating the number of change points, and NMI of the estimated segmentation as functions of ρ for different methods.

A.4 Additional Application

A.4.1 U.S. Stock Market Correlation Matrix

Figure 7 provides a two-dimensional MDS embedding computed from Δ, where the temporal trajectory shows clear regime separation and turning behavior around the detected change points.

Table 2: Running time (seconds) for spherical data with temporal dependence across different sample sizes. Each entry reports the mean runtime, with the median in parentheses.

Method	$n = 400$	$n = 2000$	$n = 5000$
DPDCP(**H1**)	0.04 (0.05)	17.86 (17.36)	506.31 (512.45)
DPDCP(**H2**)	0.04 (0.03)	5.38 (4.78)	158.07 (151.41)
KCP	0.02 (0.02)	1.88 (1.81)	36.78 (36.65)
MCPD	5.91 (5.92)	231.36 (231.88)	N/A (N/A)

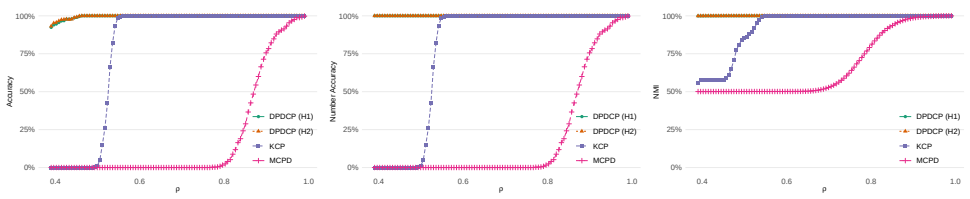

Figure 6: Performance comparison for spherical data with temporal dependence. The panels report, from left to right, the percentage of correctly estimating both the number and locations of change-points, the accuracy of estimating the number of change points, and NMI of the estimated segmentation as functions of ρ for different methods.

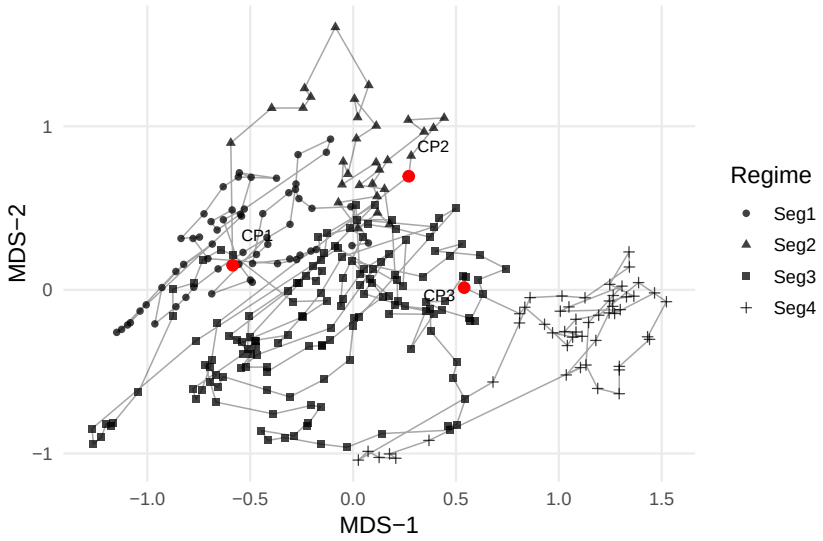

Figure 7: MDS embedding of rolling correlation matrices based on the SPD distance matrix Δ (window $= 60$, step $= 10$). Points are connected in temporal order, with red markers indicating the detected change points mapped to calendar dates.

A.4.2 U.S. Electricity Generation Data

We consider a monthly time series describing the evolution of the U.S. electricity generation mix, compiled from publicly available statistics released by the U.S. Energy Information Administration (EIA).[3] The dataset records the relative contributions of major energy sources to total net electricity generation over time and spans from 2001-01 to 2025-10, resulting in $n = 298$ observations. We classify the resources into seven categories: coal, petroleum-based fuels, natural gas, nuclear, conventional hydroelectric power, renewables (including wind, biomass, geothermal), and solar. Each monthly observation is represented as a compositional vector $x_t \in \Delta^6 = \{x \in \mathbb{R}^7_+ : \mathbf{1}^\top x = 1\}$. To account for the simplex geometry, we apply the square-root transformation $\phi(x) = (\sqrt{x_1}, \ldots, \sqrt{x_7})$, which maps compositions onto the positive orthant of the unit sphere. Pairwise dissimilarities are then measured via the induced geodesic distance $d(x, y) = \arccos\big(\phi(x)^\top \phi(y)\big)$, yielding a distance matrix that serves as input to the proposed distance-based change-point detection method.

Applying the proposed DPDCP procedure to the full monthly sequence identifies three prominent change points located at 2009-02, 2015-02, and 2019-03. The locations are marked by vertical lines in Figure 8, which display the temporal evolution of the seven generation shares. The estimated change points correspond to periods when multiple components exhibit concurrent shifts in level or trend, indicating substantive reconfigurations of the electricity generation mix rather than isolated

[3] https://www.eia.gov/electricity/data/browser/

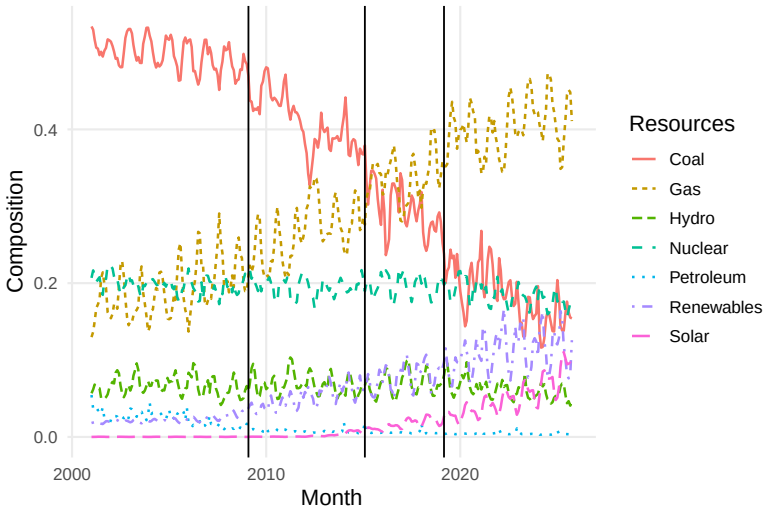

Figure 8: Temporal evolution of the U.S. electricity generation composition from 2001-01 to 2025-10. Each curve represents the share of a generation source, and the vertical solid lines indicate the change points detected by the proposed method.

fluctuations. The change around early 2015 aligns with a well-documented acceleration in renewable electricity deployment in the United States, during which wind and solar capacity expanded rapidly and began to materially affect the overall generation structure. An official report released by the U.S. administration highlights 2015 as a milestone year for renewable electricity progress, providing an external explanation for the observed structural break.[4] The change detected in early 2019 coincides with a further shift in the balance between fossil-based and renewable generation, during which renewable sources increasingly displaced coal in monthly generation totals. This transition is visible in the sustained decline of coal's share and the continued rise of renewable and natural gas contributions. The earlier change point in 2009 marks the onset of a longer-term transition phase, during which traditional coal-dominated generation began to weaken and alternative sources gradually gained prominence.

A.5 Algorithms and Implementation Details

In this section, we describe the algorithmic implementation of our DPDCP method. Algorithms 1 and 2 provide the computational details for constructing the segment cost matrices C^{H1} and C^{H2} under the **(H1)** and **(H2)** losses, respectively. Algorithm 3 then presents the complete dynamic-programming procedure for change-point estimation given a precomputed cost matrix.

A.6 Hardware

All results are obtained using a AMD Ryzen 9 9950X 16-Core Processor (4.30 GHz) CPU.

A.7 Limitations and Future Work

One limitation concerns computational cost. The current implementation constructs an aggregated matrix for the distance profile loss, which leads to an overall complexity of

$$O\{n^3 + n^2(K_{\max} + 1)\}.$$

[4] https://obamawhitehouse.archives.gov/blog/2016/01/13/renewable-electricity-progress-accelerated-2015

Algorithm 1 Construction of the **(H1)** segment cost matrix

Require: Distance matrix $\mathbf{D} \in \mathbb{R}^{n\times n}$ with $D_{m,i} = d(X_m, X_i)$; minimal segment length $\ell_{\min} \geq 2$.
Ensure: Cost matrix $C^{\mathrm{H1}} \in \mathbb{R}^{n\times n}$ with $C^{\mathrm{H1}}(a,b) = +\infty$ for $b - a + 1 < \ell_{\min}$; scaling constant $\kappa_{\mathrm{H1}} = n^{-3}$.

Initialize matrix $M \in \mathbb{R}^{n\times n}$ with $M_{ij} \leftarrow 0$ for all $1 \leq i, j \leq n$.
for $m = 1, 2, \ldots, n$ **do**
  Sort $(D_{m,1}, \ldots, D_{m,n})$ to obtain order statistics $(D_{m,(1)}, \ldots, D_{m,(n)})$.
  for $i = 1, 2, \ldots, n$ **do**
    Compute the reverse-rank count $w_m(i) \leftarrow \#\{r \in \{1, \ldots, n\} : D_{m,r} \geq D_{m,i}\}$.
  end for
  for $i = 1, 2, \ldots, n$ **do**
    for $j = i, i+1, \ldots, n$ **do**
      Update $M_{ij} \leftarrow M_{ij} + \min\{w_m(i), w_m(j)\}$ and set $M_{ji} \leftarrow M_{ij}$.
    end for
  end for
end for
Initialize vector $S_{\mathrm{diag}} \in \mathbb{R}^{n+1}$ and matrix $S_{\mathrm{col}} \in \mathbb{R}^{(n+1)\times n}$.
Set $S_{\mathrm{diag}}(1) \leftarrow 0$ and, for $t = 1, 2, \ldots, n$, set $S_{\mathrm{diag}}(t+1) \leftarrow S_{\mathrm{diag}}(t) + M_{tt}$.
for $j = 1, 2, \ldots, n$ **do**
  Set $S_{\mathrm{col}}(1, j) \leftarrow 0$ and, for $t = 1, 2, \ldots, n$, set $S_{\mathrm{col}}(t+1, j) \leftarrow S_{\mathrm{col}}(t, j) + M_{tj}$.
end for
Initialize $C^{\mathrm{H1}}(a,b) \leftarrow +\infty$ for all $1 \leq a, b \leq n$.
for $a = 1, 2, \ldots, n$ **do**
  Initialize $Q \leftarrow 0$.
  for $b = a, a+1, \ldots, n$ **do**
    Set $L \leftarrow b - a + 1$.
    if $b = a$ **then**
      Set $Q \leftarrow M_{aa}$.
    else
      Set $c \leftarrow S_{\mathrm{col}}(b, b) - S_{\mathrm{col}}(a, b)$.
      Update $Q \leftarrow Q + 2c + M_{bb}$.
    end if
    if $L \geq \ell_{\min}$ **then**
      Set $s \leftarrow S_{\mathrm{diag}}(b+1) - S_{\mathrm{diag}}(a)$.
      Set $C^{\mathrm{H1}}(a,b) \leftarrow s - \frac{1}{L}Q$.
    end if
  end for
end for
Output $(C^{\mathrm{H1}}, \kappa_{\mathrm{H1}})$ with $\kappa_{\mathrm{H1}} = n^{-3}$.

This cost can become challenging when the sample size grows large. One possible remedy is to use a Monte Carlo approximation for the integral in Equation (3):

$$\int \left[I\{d(w, X_i) \leq t\} - \widehat{F}_{[t_k, t_{k+1}]}(w, t) \right]^2 dH(w, t)$$
$$\approx \frac{1}{N} \sum_{j=1}^{N} \left[I\{d(w_j, X_i) \leq t_j\} - \widehat{F}_{[t_k, t_{k+1}]}(w_j, t_j) \right]^2,$$

where (w_j, t_j), $j = 1, \ldots, N$, are sampled independently from H conditional on the observations. As in the randomized feature approach of Rahimi & Recht (2007), the sampled features are shared across candidate segments. Their construction requires $O(Nn)$ operations; segment cost computation and dynamic programming remain necessary.

Figure 9 compares DPDCP with its Monte Carlo approximation using $N = 1024$ features for spherical data with temporal dependence, with $n = 400$ and 500 replications at each of 100 equally spaced signal levels $\rho \in [0.39, 0.99]$. The accuracy measure remains close, with a modest loss in localization accuracy at weaker signals. At $n = 2000$ and $\rho = 0.52$, the Monte Carlo implementations

Algorithm 2 Construction of the **(H2)** segment cost matrix

Require: Distance matrix $\mathbf{D} \in \mathbb{R}^{n\times n}$ with $D_{m,i} = d(X_m, X_i)$; minimal segment length $\ell_{\min} \geq 2$.
Ensure: Cost matrix $C^{\mathrm{H2}} \in \mathbb{R}^{n\times n}$ with $C^{\mathrm{H2}}(a,b) = +\infty$ for $b - a + 1 < \ell_{\min}$; scaling constant $\kappa_{\mathrm{H2}} = \big(n^2(t_{\max} - t_{\min})\big)^{-1}$.

Compute $t_{\min} \leftarrow \min_{1\leq i<j\leq n} D_{ij}$ and $t_{\max} \leftarrow \max_{1\leq i<j\leq n} D_{ij}$, $t_{\min} < t_{\max}$.
Initialize matrix $A \in \mathbb{R}^{n\times n}$ with $A_{ij} \leftarrow 0$ for all $1 \leq i, j \leq n$.
for $i = 1, 2, \ldots, n-1$ **do**
  for $j = i+1, i+2, \ldots, n$ **do**
    Compute $A_{ij} \leftarrow \sum_{m=1}^{n} \big|D_{m,i} - D_{m,j}\big|$ and set $A_{ji} \leftarrow A_{ij}$.
  end for
end for
Initialize matrix $S_{\mathrm{col}} \in \mathbb{R}^{(n+1)\times n}$.
for $j = 1, 2, \ldots, n$ **do**
  Set $S_{\mathrm{col}}(1, j) \leftarrow 0$ and, for $t = 1, 2, \ldots, n$, set $S_{\mathrm{col}}(t+1, j) \leftarrow S_{\mathrm{col}}(t, j) + A_{tj}$.
end for
Initialize $C^{\mathrm{H2}}(a, b) \leftarrow +\infty$ for all $1 \leq a, b \leq n$.
for $a = 1, 2, \ldots, n$ **do**
  Initialize $R \leftarrow 0$.
  for $b = a, a+1, \ldots, n$ **do**
    Set $L \leftarrow b - a + 1$.
    if $b > a$ **then**
      Set $c \leftarrow S_{\mathrm{col}}(b, b) - S_{\mathrm{col}}(a, b)$.
      Update $R \leftarrow R + c$.
    end if
    if $L \geq \ell_{\min}$ **then**
      Set $C^{\mathrm{H2}}(a, b) \leftarrow \frac{1}{L} R$.
    end if
  end for
end for
Output $(C^{\mathrm{H2}}, \kappa_{\mathrm{H2}})$ with $\kappa_{\mathrm{H2}} = \big(n^2(t_{\max} - t_{\min})\big)^{-1}$.

achieve speedup factors of 15.13 and 10.01 for **(H1)** and **(H2)**, respectively, based on average total runtime over 500 replications.

Futhermore, this computational issue should be considered alongside a practical advantage of DPDCP in specifying the similarity structure. KCP requires the practitioner to select a positive semidefinite kernel together with associated tuning choices, such as its bandwidth. For complex random objects, specifying a meaningful and computationally tractable metric d is often easier than selecting an appropriate kernel, since the metric can be guided by the geometry of the data, domain knowledge, or existing studies. A common kernel construction is $k(x, y) = \psi(d(x, y))$, where ψ is a scalar kernel profile. Although computationally convenient, this construction requires compatibility between ψ and d to ensure positive semidefiniteness, as discussed in Feragen et al. (2015).

Bandwidth selection introduces additional statistical and computational considerations. Kernel performance can be sensitive to the bandwidth, and a single global bandwidth may be adversely affected by heterogeneous distance scales, including those arising from the accumulation of jump magnitudes across multiple changes. Adapting bandwidths across observations or regions, or repeatedly evaluating many candidate values, may substantially increase computational cost.

For DPDCP, **(H1)** and **(H2)** specify the loss directly through distance profiles and an explicit weighting measure H, yielding a principled segmentation criterion for random objects in general metric spaces. This formulation avoids separate kernel design and bandwidth tuning, thereby reducing their associated practical and computational burden. The metric can be selected using the sources of knowledge described above; when no reliable distance is available in advance, metric learning offers a further approach to selecting distances from data (Xing et al., 2002; Weinberger & Saul, 2009; Kulis, 2013; Bellet et al., 2013).

Algorithm 3 Penalized dynamic programming

Require: Cost matrix $C \in \mathbb{R}^{n\times n}$ with infeasible entries set to $+\infty$; minimal segment length $\ell_{\min} \geq 2$; maximal change point number $K_{\max} \geq 1$; scaling constant $\kappa > 0$; slope parameters $\mathrm{frac} \in (0,1)$ and $\alpha > 0$. The inputs must satisfy $M \geq 2$ and $M - K_0 + 1 \geq 3$, where $M = \min\{K_{\max}, \lfloor n/\ell_{\min}\rfloor - 1\}$ and $K_0 = \max\{0, \lceil \mathrm{frac}(M+1)\rceil - 1\}$.

Ensure: Estimated number of change points $\widehat{K}$ and estimated change points $\widehat{T}_{\widehat{K}}$.

Set $K_{\max} \leftarrow \min\{K_{\max}, \lfloor n/\ell_{\min}\rfloor - 1\}$.
Initialize matrix $\mathrm{DP} \in \mathbb{R}^{(K_{\max}+1)\times n}$ and matrix $\mathrm{Prev} \in \mathbb{Z}^{(K_{\max}+1)\times n}$.
Set $\mathrm{DP}(K+1,t) \leftarrow +\infty$ and $\mathrm{Prev}(K+1,t) \leftarrow 0$ for all $K = 0,1,\ldots,K_{\max}$ and $t = 1,\ldots,n$.
for $t = \ell_{\min}, \ell_{\min}+1, \ldots, n$ **do**
    Set $\mathrm{DP}(1,t) \leftarrow C(1,t)$.
end for
for $K = 1,2,\ldots,K_{\max}$ **do**
    for $t = (K+1)\ell_{\min}, (K+1)\ell_{\min}+1, \ldots, n$ **do**
        Initialize $\mathrm{best} \leftarrow +\infty$ and $\mathrm{best_s} \leftarrow 0$.
        for $s = K\ell_{\min}, K\ell_{\min}+1, \ldots, t-\ell_{\min}$ **do**
            Set $\mathrm{val} \leftarrow \mathrm{DP}(K,s) + C(s+1,t)$.
            if $\mathrm{val} < \mathrm{best}$ **then**
                Set $\mathrm{best} \leftarrow \mathrm{val}$ and $\mathrm{best_s} \leftarrow s$.
            end if
        end for
        Set $\mathrm{DP}(K+1,t) \leftarrow \mathrm{best}$ and $\mathrm{Prev}(K+1,t) \leftarrow \mathrm{best_s}$.
    end for
end for
Initialize vector $\mathrm{dp}_{\mathrm{end}} \in \mathbb{R}^{K_{\max}+1}$ and vector $\widehat{\mathcal{R}} \in \mathbb{R}^{K_{\max}+1}$.
for $K = 0,1,\ldots,K_{\max}$ **do**
    Set $\mathrm{dp}_{\mathrm{end}}(K+1) \leftarrow \mathrm{DP}(K+1,n)$ and $\widehat{\mathcal{R}}(K+1) \leftarrow \kappa\,\mathrm{dp}_{\mathrm{end}}(K+1)$.
end for
Set $K_0 \leftarrow \max\{0, \lceil \mathrm{frac}\cdot(K_{\max}+1)\rceil - 1\}$.
Define $x_1(K) \leftarrow \frac{1}{n}\log\binom{n-1}{K}$ and $x_2(K) \leftarrow \frac{K+1}{n}$ for $K = 0,1,\ldots,K_{\max}$.
Compute the ordinary least squares estimates $(\widehat{\beta}_0, \widehat{\beta}_1, \widehat{\beta}_2)$ from the regression

$$\widehat{\mathcal{R}}(K+1) = \beta_0 + \beta_1 x_1(K) + \beta_2 x_2(K) + \varepsilon_K$$

over $K = K_0, K_0+1, \ldots, K_{\max}$.
Set $\widehat{c}_1 \leftarrow \max\{0, -\alpha\,\widehat{\beta}_1\}$ and $\widehat{c}_2 \leftarrow \max\{0, -\alpha\,\widehat{\beta}_2\}$.
Initialize vector $\mathrm{crit} \in \mathbb{R}^{K_{\max}+1}$.
for $K = 0,1,\ldots,K_{\max}$ **do**
    Set $\mathrm{pen}(K) \leftarrow \frac{\widehat{c}_1 \log\binom{n-1}{K} + \widehat{c}_2(K+1)}{n}$.
    Set $\mathrm{crit}(K+1) \leftarrow \widehat{\mathcal{R}}(K+1) + \mathrm{pen}(K)$.
end for
Set $\widehat{K} \leftarrow \arg\min_{0\leq K\leq K_{\max}} \mathrm{crit}(K+1)$.
Set $t \leftarrow n$ and initialize $\widehat{T}_{\widehat{K}} \leftarrow \emptyset$.
for $K = \widehat{K}, \widehat{K}-1, \ldots, 1$ **do**
    Set $s \leftarrow \mathrm{Prev}(K+1,t)$, update $\widehat{T}_{\widehat{K}} \leftarrow \widehat{T}_{\widehat{K}} \cup \{s\}$, and set $t \leftarrow s$.
end for
Sort $\widehat{T}_{\widehat{K}}$ increasingly.

In conclusion, the simulations indicate that the Monte Carlo approximation substantially reduces computational cost with little loss of estimation accuracy. The evidence here is empirical; establishing theoretical guarantees for the approximation remains a direction for future work.

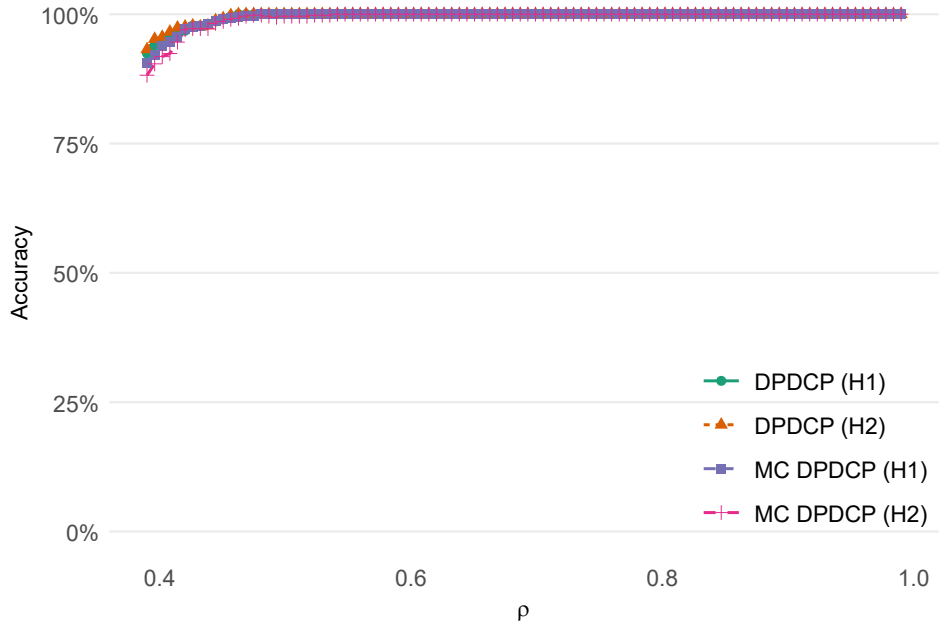

Figure 9: Performance comparison between DPDCP and its Monte Carlo approximation for spherical data with temporal dependence. The percentage of correctly estimating both the number and locations of change points is a function of ρ.