Bootstrapping Persistent Betti Numbers and Other Stabilizing Statistics

Benjamin Roycraft [1], **Johannes Krebs** [2] **and Wolfgang Polonik** [1]

[1] *Department of Statistics, University of California, Davis, One Shields Avenue, 95616, USA*
e-mail: [*] btroycraft@ucdavis.edu; [†] wpolonik@ucdavis.edu

[2] *Institute for Applied Mathematics, Heidelberg University, Im Neuenheimer Feld 205, 69120 Heidelberg, Germany*
e-mail: [**] krebs@uni-heidelberg.de

Abstract: The present contribution investigates multivariate bootstrap procedures for general stabilizing statistics, with specific application to topological data analysis. Existing limit theorems for topological statistics prove difficult to use in practice for the construction of confidence intervals, motivating the use of the bootstrap in this capacity. However, the standard nonparametric bootstrap does not directly provide for asymptotically valid confidence intervals in some situations. A smoothed bootstrap procedure, instead, is shown to give consistent estimation in these settings. The present work relates to other general results in the area of stabilizing statistics, including central limit theorems for functionals of Poisson and Binomial processes in the critical regime. Specific statistics considered include the persistent Betti numbers of Čech and Vietoris-Rips complexes over point sets in $\mathbb{R}^d$, along with Euler characteristics, and the total edge length of the k-nearest neighbor graph. Special emphasis is made throughout to weakening the necessary conditions needed to establish bootstrap consistency. In particular, the assumption of a continuous underlying density is not required. A simulation study is provided to assess the performance of the smoothed bootstrap for finite sample sizes, and the method is further applied to the cosmic web dataset from the Sloan Digital Sky Survey (SDSS). Source code is available at github.com/btroycraft/stabilizing_statistics_bootstrap.

MSC 2010 subject classifications: Primary 62F40; secondary 62H10, 62G05.
Keywords and phrases: Betti numbers, Bootstrap, Euler characteristic, Random geometric complexes, Stabilizing statistics, Stochastic geometry, Topological data analysis, Persistent homology.

1. Introduction

In recent years, a multitude of topological statistics have been developed to describe and analyze the structure of data, achieving notable success. These methods have seen application in astrophysics [1, 41, 42, 43], cancer genomics [3, 21, 11], medical imaging [18], materials science [29], fluid dynamics [30] and chemistry [52], and other wide ranging fields.

The use of simplicial complexes to summarize the geometric and topological properties of data culminates in the techniques of persistent homology. Summary statistics based on persistent homology, persistent Betti numbers, persistence diagrams, and derivatives thereof effectively extract essential topological properties from point cloud data. A broad introduction to the methods of topological data analysis can be found in [51, 15].

While the use of such statistics has seen wide success, very little is currently known about the statistical properties of these topological summaries. An initial attempt at statistical analysis using persistent homology can be seen in [10], with the later introduction of persistence landscapes in [9]. Likewise, central limit theorems have been developed for persistence

landscapes [13], Betti numbers [54] and persistent Betti numbers [27, 31] under a variety of asymptotic settings. However, the form of these results is insufficient to provide for valid confidence intervals.

In the construction of asymptotically valid confidence intervals, subsampling and bootstrap estimation have proven successful. In [23], various techniques are given for constructing confidence sets for persistence diagrams and derived statistics, including persistence diagrams generated from sublevel sets of the density function, as well as for the Čech and Vietoris-Rips complexes of data constrained to a manifold embedded in $\mathbb{R}^d$. In [13, 14], bootstrap consistency is established very generally for persistence landscapes drawn from independently generated point clouds in $\mathbb{R}^d$, assuming that the number of independent samples is allowed to grow.

However, even with these recent developments, the available techniques for constructing confidence sets using topological statistics remain severely limited. The bootstrap has proven one of the only effective tools, however the theoretical properties of bootstrap estimation applied to topological statistics are not well understood. For the large-sample asymptotic regime in particular, results are largely nonexistent.

The goal of this work is to provide the foundational theory for the bootstrap in this area. Here the validity of the bootstrap in the multivariate setting is established, a key step towards an eventual process-level result. However, the latter remains a significant technical hurdle. While motivated primarily by application to topological data analysis, the results presented here apply much more generally over a class of *stabilizing* statistics. For an additional application, we show convergence for the bootstrap applied to the total edge length of the k-nearest neighbor graph.

We also analyze the large-sample asymptotic properties of the bootstrap applied to the Čech and Vietoris-Rips complexes directly, where the underlying point cloud is a sample drawn from a common distribution on $\mathbb{R}^d$. In particular, we will show that the standard nonparametric bootstrap can fail to provide asymptotically valid confidence intervals directly in some cases. Via a smoothed bootstrap, however, we will construct multivariate confidence intervals for the mean persistent Betti number, which lie in bijection with the corresponding persistence diagram.

As defined in [38], a statistic *stabilizes* if the change in the function value induced by addition of new points to the underlying sample is at most locally determined. Applications of stabilization have allowed for the development of central limit theorems for several topological statistics. [54] show that Betti numbers exhibit the stabilization property, and provide a central limit theorem for Betti numbers derived from a homogenous Poisson process with unit intensity. [27] considers persistent Betti numbers in the homogenous Poisson process case with arbitrary intensity. Most recently [31] established multivariate central limit theorems for persistent Betti numbers with an underlying point cloud coming from either a nonhomogenous Poisson or binomial process. For the results in the present contribution, we draw significant inspiration from this most recent work.

An application of our general consistency result is made to the persistent Betti numbers of a class of distance-based simplicial complexes, including the Čech and Vietoris-Rips complexes. Throughout this work, a special focus is given towards weakening the necessary assumptions compared to previous results. Specifically, the theorems presented here apply for distributions with unbounded support, unbounded density, and possible discontinuities. We assume only a bound for the L_p-norm of the underlying sampling density.

For the first half of this paper, we focus on the theory of bootstrap estimation applied to stabilizing statistics. In Section 2 we will introduce the concept of stabilization and estab-

lish intermediate technical results in this context. We then present our general bootstrap consistency theorem.

In the second half, we introduce the main topological and geometric statistics of interest, applying the theory presented in the previous sections. In Section 3 we connect the general theory to the specific case of persistent homology and related statistics. Towards this end, we give a short introduction to simplicial complexes and persistent homology. In Section 4, the stabilization properties of persistent Betti numbers are analyzed, along with the Euler characteristic, for general classes of distance-based simplicial complexes. We establish bootstrap consistency in the large-sample limit for each of these statistics, as well as for the total edge length of the k-nearest neighbor graph. In Section 5 we provide several simulations demonstrating the finite-sample properties of the smoothed bootstrap applied to persistent Betti numbers. Finally, Section 6 illustrates the utility of the smoothed bootstrap with an application to a cosmic web dataset from the Sloan Digital Sky Survey (SDSS) [5]. Source code for the computational sections is available at github.com/btroycraft/stabilizing_statistics_bootstrap [44].

Appendix A gives an investigation of several altered problem settings in which precise stabilization properties may be derived. The proofs for all results can be found in Appendix B. Functionals considered include the "B-bounded" persistent Betti numbers and the "q-truncated" Euler characteristic.

2. Stabilizing Statistics

2.1. Central Limit Theorems for Stabilizing Statistics

Before proving bootstrap convergence, we give a brief overview of the existing work regarding stabilizing statistics. For the precise definitions used throughout this paper, see Section 2.2.

In the seminal work of [38], the chief objects of study are real valued functionals applied over point sets in $\mathbb{R}^d$. It is here that a stabilization property was first defined, and used to show central limit theorems for certain types of geometric functionals, including the length of the k-nearest neighbor graph and the number of edges in the sphere of influence graph. This initial work distilled two properties key to showing central limit theorems for geometric functionals. First is the *stabilization* property, and second is a moment bound. In short, we say that a functional ψ *stabilizes* if the cost of adding an additional point, or a set of points, to the point cloud varies only on a bounded region. Specific definitions differ by context.

In [38], the authors distinguish between two data generating regimes. First, results are shown for a homogenous Poisson process over $\mathbb{R}^d$. Alternatively, a binomial process is considered, being equivalent to a sample of fixed size from an appropriate probability distribution. Here, the functional under consideration is restricted to a bounded domain B_n of volume n, where n is allowed to increase. In this initial work, only homogenous Poisson processes and uniform binomial sampling are considered. In [39], a similar framework is used to establish laws of large numbers for graph-based functionals, including the number of connected components in the minimum spanning tree. Further quantitative refinements on the general central limit theorems for stabilizing statistics are shown in [32], [33], and [34].

As pertains to topological statistics, an initial central limit theorem for Betti numbers (see Section 3.2 for definitions) was shown in [54], establishing so-called *weak stabilization* for Betti numbers in the homogenous Poisson and uniform Binomial sampling settings. There an alternative set-up is being used where the domain is kept fixed, while the filtration

parameter is decreasing to zero. A similar result for persistent Betti numbers is given in [27].

Finally, [31] establishes multivariate central limit theorems for persistent Betti numbers under a flexible sampling setting. Here, a nonhomogeneous Poisson or binomial process is generated again over a growing domain with fixed filtration radii.

With these central limit theorem results, the stabilization property plays a central role in understanding the asymptotic behavior for wide classes of geometric and topological functionals. Unfortunately, as a reoccurring trend, explicit forms for the asymptotic normal distributions are unavailable or computationally intractable. In this work it is shown how a smoothed bootstrap procedure allows for consistent estimation of these inaccessible limiting distributions, and thus for any subsequent inference derived therefrom.

Further, the bootstrap convergence results shown in this paper apply even more broadly, given that the necessary assumptions are much weaker than normally used to establish central limit theorems. To the best of our knowledge, it is not known whether there exist stabilizing statistics which exhibit a non-normal limit, but our convergence results apply equally for any distributional limit.

2.2. Stabilization

Here, we extend and rephrase existing definitions found in [38], [39], [54], and [31] to provide a more general and consistent statistical framework. Let $\mathcal{X}\left(\mathbb{R}^d\right)$ denote the space consisting of multisets drawn from $\mathbb{R}^d$ with no accumulation points, with the further restriction that no point in a given multiset may be counted more than finitely often. Any locally-finite point process on $\mathbb{R}^d$ can be represented as a random element of $\mathcal{X}\left(\mathbb{R}^d\right)$. Let $\tilde{\mathcal{X}}\left(\mathbb{R}^d\right) \subset \mathcal{X}\left(\mathbb{R}^d\right)$ contain the finite multisets drawn from $\mathbb{R}^d$ and $\psi\colon \tilde{\mathcal{X}}\left(\mathbb{R}^d\right) \to \mathbb{R}$ be a measurable function. Furthermore, for $S, T \in \tilde{\mathcal{X}}\left(\mathbb{R}^d\right)$ define the *addition cost* of T to S as $D\left(S; \psi, T\right) := \psi\left(S \cup T\right) - \psi\left(S\right)$. When $T = \{z\}$ consists of a single point, we call $D_z\left(S; \psi\right) := D\left(S; \psi, \{z\}\right)$ an *add-one cost* or the *add-z cost*.

Broadly, we say that ψ *stabilizes* if the addition cost of a given T varies only on a bounded region. In the preceding literature, the terms "strong" and "weak" stabilization are very often used, with precise definitions changing based on circumstance. In the interest of providing more explanatory and specific terminology, we propose the following definitions.

Seen below, almost-sure and locally-determined almost-sure stabilization (see Definitions 2.4 and 2.5) correspond, respectively, to Definitions 3.1 and 2.1 in [38]. Here we have generalized by accounting for possible measurability issues, however the definitions are essentially equivalent. Let $B_z\left(r\right)$ denote the closed Euclidean ball centered at $z \in \mathbb{R}^d$ with radius r. For convenience, the dependence on ψ and T is implicit in each of the following.

Definition 2.1 (Terminal Addition Cost)**.** $D^\infty\colon \mathcal{X}\left(\mathbb{R}^d\right) \to \mathbb{R}$ is a *terminal addition cost* centered at $z \in \mathbb{R}^d$ if $D^\infty\left(S\right) = \lim_{l\to\infty} D\left(S \cap B_z\left(l\right)\right)$ for any $S \in \mathcal{X}\left(\mathbb{R}^d\right)$ such that the limit exists.

For a finite multiset $S \in \tilde{\mathcal{X}}\left(\mathbb{R}^d\right)$, the terminal addition cost centered at $z \in \mathbb{R}^d$ is $D^\infty\left(S\right) = D\left(S\right)$, because no further changes to the addition cost may occur once $S \cap B_z\left(a\right)$ contains all of S. This does not hold for infinite multisets, motivating a separate definition. In the special case where $T = \{z\}$ is a singleton at the centerpoint, the notation $D^\infty = D_z^\infty$ may be used, and will be seen throughout the remaining sections of the paper.

Definition 2.2 (Stabilization in Probability)**.** For $\mathbf{S}$ a point process taking value in $\mathcal{X}\left(\mathbb{R}^d\right)$, ψ *stabilizes* on $\mathbf{S}$ *in probability* if there exists a center point $z \in \mathbb{R}^d$ and a terminal addition cost D^∞ for ψ such that

$$\lim_{l\to\infty} \mathbb{P}^* \left[D\left(\mathbf{S} \cap B_z\left(l\right)\right) \neq D^\infty\left(\mathbf{S}\right)\right] = 0. \tag{2.1}$$

Here $\mathbb{P}^*$ denotes the outer probability of a set. Stabilization is said to occur *in probability* because, for any sequence of non-negative radii $\left(l_i\right)_{i\in\mathbb{N}}$ such that $l_i \to \infty$, $D\left(\mathbf{S} \cap B_z\left(l_i\right)\right) \xrightarrow{p} D^\infty\left(\mathbf{S}\right)$ whenever both quantities are measurable. D^∞ is unique up to a null set in this case. Stabilization in probability is difficult to show directly for many functions of interest. As such, we have the following:

Definition 2.3 (Radius of Stabilization)**.** $\rho\colon \mathcal{X}\left(\mathbb{R}^d\right) \to [0,\infty]$ is a *radius of stabilization* for ψ centered at $z \in \mathbb{R}^d$ if, for any $S \in \mathcal{X}\left(\mathbb{R}^d\right)$ and $l \in \mathbb{R}$ such that $\rho\left(S\right) \leq l < \infty$,

$$D\left(S \cap B_z\left(l\right)\right) = D\left(S \cap B_z\left(\rho\left(S\right)\right)\right). \tag{2.2}$$

$D^\infty\left(S\right) := D\left(S \cap B_z\left(\rho\left(S\right)\right)\right)$ is a valid terminal addition cost. In the case where $\lim_{l\to\infty} D\left(S \cap B_z\left(l\right)\right)$ does not exist, $\rho\left(S\right) = \infty$ necessarily, with the stabilization criterion satisfied vacuously. As with the terminal addition cost, when $T = \{z\}$ we denote $\rho = \rho_z$.

In general, for any ψ there exists a unique minimal radius of stabilization, defined as the pointwise minimum over all such radii sharing the same centerpoint. This minimum exists because $\psi\left(S \cap B_z\left(l\right)\right)$ is piecewise constant in $0 \leq l < \infty$, changing value only when a new point of S is added, and because S has no accumulation points.

Definition 2.4 (Stabilization Almost Surely)**.** For $\mathbf{S}$ a point process taking value in $\mathcal{X}\left(\mathbb{R}^d\right)$, ψ stabilizes on $\mathbf{S}$ *almost surely* if there exists a radius of stabilization $\rho\colon \mathcal{X}\left(\mathbb{R}^d\right) \to [0,\infty]$ for ψ centered at $z \in \mathbb{R}^d$ such that

$$\lim_{L\to\infty} \mathbb{P}^* \left[\rho\left(\mathbf{S}\right) > L\right] = 0. \tag{2.3}$$

Mirroring our previous terminology, we say stabilization occurs *almost surely* because, for any sequence of nonnegative radii $\left(l_i\right)_{i\in\mathbb{N}}$ such that $l_i \to \infty$, $D\left(\mathbf{S} \cap B_z\left(l_i\right)\right) \xrightarrow{a.s.} D^\infty\left(\mathbf{S}\right) = D\left(\mathbf{S} \cap B_z\left(\rho\left(\mathbf{S}\right)\right)\right)$ whenever both quantities are measurable. Here we use outer probability, because a radius of stabilization may not be a measurable function, specifically considering the unique minimal radius. Almost sure stabilization implies stabilization in probability, as shown in the following.

Proposition 2.1. *For* $\mathbf{S}$ *a simple point process taking values in* $\mathcal{X}\left(\mathbb{R}^d\right)$*, let* ψ *stabilize on* $\mathbf{S}$ *almost surely. Then* ψ *stabilizes on* $\mathbf{S}$ *in probability.*

For our proof techniques, it is often necessary to compare the stabilization properties of a function over a range of related point processes. For example, corresponding binomial, Poisson, and Cox processes can be shown to have essentially equivalent local properties, while differing globally. As defined in Definition 2.3, a given radius of stabilization could feasibly show completely different behavior on each process type. This motivates the following:

Definition 2.5 (Locally Determined Radius of Stabilization)**.** The radius of stabilization ρ centered at $z \in \mathbb{R}^d$ is *locally determined* if for any $S, T \in \mathcal{X}\left(\mathbb{R}^d\right)$

$$T \cap B_z\left(\rho\left(S\right)\right) = S \cap B_z\left(\rho\left(S\right)\right) \implies \rho\left(T\right) = \rho\left(S\right).$$

With the local-determination criterion from Definition 2.5, we can assure that stabilization must occur simultaneously on any two point processes which are locally equivalent. As in the non-locally-determined case, there exists a unique minimal locally-determined radius of stabilization:

Proposition 2.2. *For $\mathcal{R}$ the space of locally-determined radii of stabilization for ψ centered at $z \in \mathbb{R}^d$, let $\rho^*\colon \mathcal{X}\left(\mathbb{R}^d\right) \to [0,\infty]$ such that $\rho^*(S) = \inf_{\rho\in\mathcal{R}} \rho(S)$. Then ρ^* is a locally determined radius of stabilization for ψ centered at z.*

2.3. Technical Results

In all of the following, $\mathcal{P}\left(\mathbb{R}^d\right)$ denotes the set of probability distributions over $\mathbb{R}^d$. $Y_1, ..., Y_n \overset{\text{iid}}{\sim} G$ is a sample from $G \in \mathcal{P}\left(\mathbb{R}^d\right)$ and $Y' \sim G$ an independent copy. Let $\mathbf{Y}_n = \{Y_i\}_{i=1}^n$ be the induced multiset. This definition may be simply denoted by $\mathbf{Y}_n := \{Y_i\}_{i=1}^n \overset{\text{iid}}{\sim} G$. For a measurable function $\psi\colon \tilde{\mathcal{X}}\left(\mathbb{R}^d\right) \to \mathbb{R}$, define the following conditions:

(E1) For a given $\mathcal{C} \subseteq \mathcal{P}\left(\mathbb{R}^d\right)$ and some $a > 2$, there exists $E_a < \infty$ such that

$$\sup_{G\in\mathcal{C}} \sup_{n\in\mathbb{N}} \mathbb{E}\left[\left|\psi\left(\sqrt[d]{n}\left(\mathbf{Y}_n \cup \{Y'\}\right)\right) - \psi\left(\sqrt[d]{n}\mathbf{Y}_n\right)\right|^a\right] \leq E_a. \tag{2.4}$$

(E2) For some $a > 2$ and $R > 0$, there exist $U_a > 0$ and $u_a > 1$ satisfying the following property: For any $S \in \tilde{\mathcal{X}}\left(\mathbb{R}^d\right)$ and $y \in \mathbb{R}^d$,

$$|\psi(S \cup \{y\}) - \psi(S)|^a \leq U_a\left(1 + \#\left\{S \cap B_y(R)\right\}^{u_a}\right). \tag{2.5}$$

(E1) requires a moment bound that holds uniformly in the sample size and distribution $G \in \mathcal{C}$. Clearly, if (E1) is satisfied for $\mathcal{C}$, it is also satisfied for any subset of $\mathcal{C}$. In the context of the topological statistics considered in this work, (E1) is primarily useful for proof purposes, and is mainly established via (E2) (See Lemma 2.3). However, as will be seen with the case of the k-nearest neighbor graph, Corollary 4.6, there exist useful statistics which do not conform to (E2), and the more general condition must be used. (E1) is related to the "uniform bounded moments" condition, Definition 2.2 in [38]. Our version has been suitably generalized, the original definition considering only $a = 4$. Let $\mathcal{C}_{p,M}\left(\mathbb{R}^d\right)$ denote the class of probability distributions $G \in \mathcal{P}\left(\mathbb{R}^d\right)$ admitting a density g such that $\|g\|_p \leq M$. We have the following:

Lemma 2.3. *For $p > 2$, let ψ satisfy (E2) with $u_a \leq p - 1$ for some $a > 2$. Then for any $M < \infty$, ψ satisfies (E1) for $\mathcal{C}_{p,M}\left(\mathbb{R}^d\right)$.*

For d_{TV} the total variation distance between probability distributions and $B_F(\epsilon, d_{\text{TV}})$ the closed ϵ-neighborhood of F under d_{TV}, we have the following stabilization conditions:

(S1) For a given $\mathcal{C} \subseteq \mathcal{P}\left(\mathbb{R}^d\right)$, $F \in \mathcal{C}$, $b > 0$, and some $(l_\epsilon)_{\epsilon>0}$ such that $\lim_{\epsilon\to 0} l_\epsilon \epsilon^b = 0$, as $\epsilon \to 0$,

$$\sup_{G\in\mathcal{C}\cap B_F(\epsilon;d_{\text{TV}})} \sup_{n\in\mathbb{N}} \mathbb{P}\left[D_{\sqrt[d]{n}Y'}\left(\left(\sqrt[d]{n}\mathbf{Y}_n\right) \cap B_{\sqrt[d]{n}Y'}(l_\epsilon)\right) \neq D_{\sqrt[d]{n}Y'}\left(\sqrt[d]{n}\mathbf{Y}_n\right)\right] \to 0.$$

(S2) For $G \in \mathcal{P}\left(\mathbb{R}^d\right)$, there exist locally-determined radii of stabilization $(\rho_z)_{z\in\mathbb{R}^d}$ for ψ satisfying

$$\lim_{L\to\infty} \sup_{n\in\mathbb{N}} \mathbb{P}^*\left[\rho_{\sqrt[d]{n}Y'}\left(\sqrt[d]{n}\mathbf{Y}_n\right) > L\right] = 0. \tag{2.6}$$

(S1) and (S2) can be summarized as uniform stabilization conditions, either in probability or almost surely. (S1) as stated is a technical condition mainly serving to weaken the necessary conditions providing for bootstrap consistency. As such, we have the following lemma linking (S1) and (S2).

Lemma 2.4. *Let ψ satisfy (S2) for $F \in C_{p,M}\left(\mathbb{R}^d\right)$. Then ψ satisfies (S1) for $C_{p,M}\left(\mathbb{R}^d\right)$, F, $b = (p-2)/(d(p-1))$, and any $(l_\epsilon)_{\epsilon>0}$ such that $\lim_{\epsilon\to 0} l_\epsilon \epsilon^{(p-2)/(d(p-1))} = 0$ and $\lim_{\epsilon\to 0} l_\epsilon = \infty$.*

We can often greatly simplify the addition costs and radii of stabilization required in (S1) and (S2). For example, given a translation-invariant function ψ and any D_0, ρ_0 for ψ centered at 0, corresponding quantities can be constructed for any other center point. For $z \in \mathbb{R}^d$, $D_z \colon \mathcal{X}\left(\mathbb{R}^d\right) \to \mathbb{R}$ where $D_z(S) = D_0(S - z)$ is an add-z cost for ψ centered at z. Likewise $\rho_z \colon \mathcal{X}\left(\mathbb{R}^d\right) \to [0,\infty]$ where $\rho_z(S) = \rho_0(S - z)$ is a radius of stabilization for ψ centered at z. In the following, $\mathbf{P}_\lambda$ denotes a homogeneous Poisson process on $\mathbb{R}^d$ with intensity λ.

Lemma 2.5. *Let $F \in C_{p,M}$ with $p > 2$ and $M < \infty$. Let ρ_0 be a locally-determined radius of stabilization for ψ centered at 0. Suppose that for any given $a, b \in (0,\infty)$, and $\delta > 0$, there exists an $L_{a,b,\delta} < \infty$ and a measurable set $A_{a,b,\delta}$ with $\rho_0^{-1}((L_{a,b,\delta},\infty]) \subseteq A_{a,b,\delta}$ such that*

$$\sup_{\lambda\in[a,b]} \mathbb{P}^*\left[\rho_0(\mathbf{P}_\lambda) > L_{a,b,\delta}\right] \leq \sup_{\lambda\in[a,b]} \mathbb{P}\left[\mathbf{P}_\lambda \in A_{a,b,\delta}\right] \leq \delta. \tag{2.7}$$

Then for any $\delta > 0$ there exists an $n_\delta < \infty$ and $L_\delta < \infty$ such that

$$\sup_{n\geq n_\delta} \mathbb{P}^*\left[\rho_0(\mathbf{X}_n - X') > L_\delta\right] \leq \delta. \tag{2.8}$$

Lemma 2.5 provides a convenient tool for "de-Poissonizing" a locally-determined radius of stabilization. Often it is easier to show stabilization properties for a homogeneous Poisson process than for a binomial process directly. Lemma 2.5 allows for the extension of homogeneous Poisson results to the binomial setting, as is required for Lemma 4.1 and Corollary 4.6. Note that the conclusion is not the same as the statement of (S1), only applying for $n \geq n_\delta$. Some extra effort is required for the conclusion to hold for all $n \in \mathbb{N}$, depending on the specifics of the function ψ considered. We come to the following important proposition, the main supporting result for our general bootstrap consistency theorem, Theorem 2.7.

Proposition 2.6. *For $p > 2$ and $M < \infty$, let ψ satisfy (E1) and (S1) for $\mathcal{C}_{p,M}\left(\mathbb{R}^d\right)$, $F \in \mathcal{C}_{p,M}\left(\mathbb{R}^d\right)$, and some $a > 2$. Then for any $G \in \mathcal{C}_{p,M}\left(\mathbb{R}^d\right) \cap B_F(\epsilon, d_{TV})$, there exist iid coupled random variables $((X_i, Y_i))_{i\in\mathbb{N}}$ such that $\mathbf{X}_n = \{X_i\}_{i=1}^n \overset{iid}{\sim} F$, $\mathbf{Y}_n = \{Y_i\}_{i=1}^n \overset{iid}{\sim} G$, and*

$$\sup_{n\in\mathbb{N}} \operatorname{Var}\left[\frac{1}{\sqrt{n}}\left(\psi\left(\sqrt[d]{n}\mathbf{X}_n\right) - \psi\left(\sqrt[d]{n}\mathbf{Y}_n\right)\right)\right] \leq \gamma_\epsilon. \tag{2.9}$$

The value γ_ϵ does not depend on G and satisfies $\lim_{\epsilon\to 0}\gamma_\epsilon = 0$.

For any two distributions $\mathcal{L}_1$ and $\mathcal{L}_2$ over $\mathbb{R}$, we may define the 2-Wasserstein distance between $\mathcal{L}_1$ and $\mathcal{L}_2$ as

$$W_2(\mathcal{L}_1, \mathcal{L}_2) := \sqrt{\inf_{U\sim\mathcal{L}_1, V\sim\mathcal{L}_2} \mathbb{E}\left[(U-V)^2\right]} \tag{2.10}$$

where it is assumed that U and V follow a joint distribution with marginals $\mathcal{L}_1$ and $\mathcal{L}_2$. For $\mathcal{L}$ denoting the law or distribution of a random variable, the variance given in the conclusion of Proposition 2.6 bounds above

$$W_2^2\Bigg(\mathcal{L}\left\{\frac{1}{\sqrt{n}}\left(\psi\left(\sqrt[d]{n}\mathbf{X}_n\right)-\mathbb{E}\left[\psi\left(\sqrt[d]{n}\mathbf{X}_n\right)\right]\right)\right\}, \tag{2.11}$$
$$\mathcal{L}\left\{\frac{1}{\sqrt{n}}\left(\psi\left(\sqrt[d]{n}\mathbf{Y}_n\right)-\mathbb{E}\left[\psi\left(\sqrt[d]{n}\mathbf{Y}_n\right)\right]\right)\right\}\Bigg).$$

Consequently, Proposition 2.6 shows that this W_2 distance can be made arbitrarily small uniformly over a neighborhood of distributions around F. An appropriately smoothed empirical distribution falls within such a small neighborhood with high probability, given sufficiently large sample sizes.

Furthermore, it can be seen that Proposition 2.6 extends directly to finite sums. Given any $(A_i)_{i=1}^k$ and $(B_i)_{i=1}^k$, we have that $\mathrm{Var}\left[\sum_{i=1}^k A_i - \sum_{i=1}^k B_i\right] \leq k\sum_{i=1}^k \mathrm{Var}\left[A_i - B_i\right]$. Thus, if the conclusion of Proposition 2.6 holds for any finite set of functions, $(\psi_i)_{i=1}^k$, it also holds for $\sum_{i=1}^k \psi_i$, with rate depending on the worst case ψ_i.

It should be noted that (S1) is slightly stronger than necessary to establish Proposition 2.6. As stated, $D_{\sqrt[d]{n}Y'}\left(\left(\sqrt[d]{n}\mathbf{Y}_n\right)\cap B_{\sqrt[d]{n}Y'}(l_\epsilon)\right)$ itself is compared to the terminal add-one cost $D_{\sqrt[d]{n}Y'}\left(\sqrt[d]{n}\mathbf{Y}_n\right)$. As could be useful for some statistics, it is only required that an appropriate bound displays the desired stabilization property, see the provided proof for details.

2.4. Smoothed Bootstrap

The bootstrap is an estimation technique used to construct approximate confidence intervals for a given population parameter. In cases where asymptotic approximations for the sampling distribution of a statistic are inconvenient or unavailable, bootstrap estimation provides a general tool for constructing approximate confidence intervals. Bootstrap estimation is well-studied in the statistical literature, an introduction being provided in [40]. In this section, we will show consistency for a smoothed bootstrap in estimating the limiting distribution of a standardized stabilizing statistic, ψ, in the multivariate setting. We describe the general procedure below:

Let $\mathbf{X}_n = \{X_i\}_{i=1}^n \overset{\text{iid}}{\sim} F$. We estimate the sampling distribution of

$$\frac{1}{\sqrt{n}}\left(\psi\left(\sqrt[d]{n}\mathbf{X}_n\right)-\mathbb{E}\left[\psi\left(\sqrt[d]{n}\mathbf{X}_n\right)\right]\right) \tag{2.12}$$

using a plug-in estimator $\hat{F}_n$ for the underlying data distribution F. In the standard non-parametric bootstrap, we estimate F by the empirical distribution, giving probability to each unique value of $(X_i)_{i=1}^n$, proportional to the number of repetitions within $\mathbf{X}_n$. We have the bootstrap statistic

$$\frac{1}{\sqrt{m}}\left(\psi\left(\sqrt[d]{m}\mathbf{X}_m^*\right)-\mathbb{E}\left[\psi\left(\sqrt[d]{m}\mathbf{X}_m^*\right)\middle|\mathbf{X}_n\right]\right), \tag{2.13}$$

where $\mathbf{X}_m^* = \{X_i^*\}_{i=1}^m \overset{\text{iid}}{\sim} \hat{F}_n|\mathbf{X}_n$, conditional on $\mathbf{X}_n$. The sampling distribution of the bootstrap version provides an estimate for the distribution of the original statistic, which

in the ideal case converges to the truth in the large-sample limit. Confidence intervals for $\mathbb{E}\left[\psi\left(\sqrt[d]{n}\mathbf{X}_n\right)\right]$ are then constructed from the bootstrap distribution and $\psi\left(\sqrt[d]{n}\mathbf{X}_n\right)$.

However, as will be seen in Section 4.1, for some classes of topological statistics the standard bootstrap may not directly replicate the correct sampling distribution asymptotically. Consequently, we instead estimate F by a smoothed distribution approximation. Such a smoothed bootstrap procedure can be shown to provide consistent estimation, even when the standard nonparametric bootstrap may fail.

For the smoothed bootstrap sampling procedure outlined here, we require that F has a density f. Let $\hat{f}_n$ be an estimator for the true density with corresponding distribution $\hat{F}_n$, each a function of the sample $\mathbf{X}_n$. Conditional on $\mathbf{X}_n$, we draw bootstrap samples $\mathbf{X}_m^*$ independently from $\hat{F}_n|\mathbf{X}_n$. A particular choice of $\hat{f}_n$ is given via kernel density estimation. For a kernel function Q and bandwidth $h > 0$, the kernel density estimator of $f(x)$ based on the sample $(X_i)_{i=1}^n$ is $\hat{f}_{n,h}(x) := 1/\left(nh^d\right)\sum_{i=1}^n Q\left((x - X_i)/h\right)$.

In practice, when Q corresponds to a probability density, the kernel density estimator allows for convenient sampling, as is required for implementation. Generating a sample from $\hat{f}_{n,h}$ is equivalent to first drawing from the empirical distribution on $\mathbf{X}_n$, then adding independent noise following the distribution defined by Q, scaled by the bandwidth h. Other density estimators, including those using higher-order kernels, may not facilitate efficient sampling. However, the theory established here supports the use of any density estimator which meets the required convergence criteria, computational factors aside. More complicated data-dependent estimators are also possible, falling under a similar sampling framework. See Sections 5 and 6 for specifics on density estimation as pertains to this work from a practical perspective.

We now present our main result. The following theorem establishes consistency for the smoothed bootstrap in the multivariate setting. We give the result for a vector of stabilizing statistics. In the context of the topological statistics introduced in Section 3, this can be the persistent Betti numbers or Euler characteristic evaluated at different filtration parameters.

Theorem 2.7. *Let $F \in \mathcal{P}\left(\mathbb{R}^d\right)$ with density f such that $\|f\|_p < \infty$ for some $p > 2$. Furthermore, let F and $\hat{f}_n$ be such that $\|\hat{f}_n - f\|_1 \to 0$ and $\|\hat{f}_n - f\|_p \to 0$ in probability (resp. a.s.). Suppose $\vec{\psi}\colon \tilde{\mathcal{X}}\left(\mathbb{R}^d\right) \to \mathbb{R}^k$ has component functions $\psi_j\colon \tilde{\mathcal{X}}\left(\mathbb{R}^d\right) \to \mathbb{R}$, $1 \le j \le k$ satisfying (E1) and (S1) for $\mathcal{C}_{p,M}\left(\mathbb{R}^d\right)$, $M > \|f\|_p$, F, and $b = (p-2)/(d(p-1))$. Then for a sample $\mathbf{X}_n = \{X_i\}_{i=1}^n \overset{iid}{\sim} F$, $(m_n)_{n\in\mathbb{N}}$ such that $\lim_{n\to\infty} m_n = \infty$, a bootstrap sample $\mathbf{X}_{m_n}^* = \{X_i^*\}_{i=1}^{m_n} \overset{iid}{\sim} \hat{F}_n|\mathbf{X}_n$, and a multivariate distribution G,*

$$\frac{1}{\sqrt{n}}\left(\vec{\psi}\left(\sqrt[d]{n}\mathbf{X}_n\right) - \mathbb{E}\left[\vec{\psi}\left(\sqrt[d]{n}\mathbf{X}_n\right)\right]\right) \xrightarrow{d} G$$

if and only if

$$\frac{1}{\sqrt{m_n}}\left(\vec{\psi}\left(\sqrt[d]{m_n}\mathbf{X}_{m_n}^*\right) - \mathbb{E}\left[\vec{\psi}\left(\sqrt[d]{m_n}\mathbf{X}_{m_n}^*\right)\middle|\mathbf{X}_n\right]\right) \xrightarrow{d} G \text{ in probability (resp. a.s.).}$$

Theorem 2.7 establishes the asymptotic validity of bootstrap estimation for a range of stabilizing statistics under fairly mild conditions on the underlying density. However, it should be noted that further restrictions on the density and density estimate may be required to satisfy (E1) and (S1), see Corollary 4.6 for example. The conditions under which $\|\hat{f}_{n,h_n} - f\|_1 \to 0$ in probability or a.s. can be found in [20]. Proposition C.1 considers the convergence of $\|\hat{f}_{n,h_n} - f\|_p$, either in probability or almost surely. This result is outside the main

contribution of this paper, but is interesting in its own right. Notably, no conditions are placed on the density f except $\|f\|_p < \infty$.

As a point of caution, it is known that kernel density estimators suffer from a curse of dimensionality. The convergence properties of the density estimator $\hat{f}_n$ appear implicitly within the necessary assumptions for Theorem 2.7. In particular, diminishing performance can be expected in higher dimensions, as shown by the provided simulations of Section 5.

The above result holds for any choice of m_n such that $\lim_{n\to\infty} m_n = \infty$, and is stated as such for the sake of generality. In practical application, $m_n = n$ is standard, and will be used throughout the simulation and data analysis sections of this paper. However, given that the computational complexity of ψ often grows quickly with n, using a smaller m_n could prove more feasible from a computational perspective.

Strictly speaking, convergence to a limiting distribution is not required for the bootstrap to provide asymptotically valid confidence intervals. Proposition 2.6 gives that, with high probability, the smoothed bootstrap and true sampling distributions become close in 2-Wasserstein distance. Provided that the cumulative distribution function $F_{\vec{\psi}_n}$ of $(\vec{\psi}(\sqrt[d]{n}\mathbf{X}_n) - \mathbb{E}[\vec{\psi}(\sqrt[d]{n}\mathbf{X}_n)])/\sqrt{n}$ has the property

$$\lim_{\delta\to 0} \limsup_{n\to\infty} \sup_{x\in\mathbb{R}^d} \left| F_{\vec{\psi}_n}(x+\delta) - F_{\vec{\psi}_n}(x) \right| \to 0, \tag{2.14}$$

it can be shown that confidence intervals constructed from the bootstrap statistic still achieve the stated confidence level with high probability, given a sufficiently large sample. Convergence to a continuous limiting CDF is just one way of satisfying this condition. However, this extension is unavailable for the topological statistics considered here, as the behavior of the finite sample statistics is currently very poorly understood.

In the later sections, we will show that the necessary moment and stabilization conditions for Theorem 2.7 are satisfied for several specific statistics of interest, chiefly the Euler characteristic and persistent Betti numbers for a class of simplicial complexes.

3. Simplicial Complexes and Persistence Homology

3.1. Simplicial Complexes

Let $\mathcal{K} = \{K^r\}_{r\in\mathbb{R}}$ be a filtration of simplicial complexes, with $K^r \subseteq K^t$ for $r < t$. Each complex is a collection of *simplices*, subsets of the vertex multiset, V. Here any repeated vertices are considered distinct. For a collection of simplices K to be a simplicial complex, for any two simplices $S \subset V$ and $T \subset S$, $S \in K$ only if $T \in K$. Here a simplex is only included along with all of its subsets. For a given simplicial complex K, K_q denotes the subset of K consisting of all q-simplices. q-simplices are those simplices consisting of $q+1$ vertices. Each q-simplex is said to have dimension q. A graph or network refers to a simplicial complex consisting of only 1-simplices (edges) and 0-simplices (vertices).

We will be looking at simplicial complexes constructed over point clouds in $\mathbb{R}^d$. The two major examples are the Čech and Vietoris-Rips complexes:

$$K^r_{\mathrm{C}}(S) = \left\{\sigma \subseteq S \colon \exists z \in \mathbb{R}^d \text{ s.t. } \|z - x\| \le r \ \forall x \in \sigma\right\} \tag{3.1}$$

$$K^r_{\mathrm{VR}}(S) = \left\{\sigma \subseteq S \colon \|x - y\| \le 2r \ \forall x, y \in \sigma\right\}. \tag{3.2}$$

Each of these complexes summarizes the geometric and topological properties within a given point cloud. The Vietoris-Rips complex can be considered a "completion" of the Čech

complex, in so much that the Vietoris-Rips complex is the largest simplicial complex with the same edge set as the Čech complex. While the primary motivation for the results given here is application to the Čech and Vietoris-Rips complexes, our main results apply for a range of possible complexes. For example, for computational reasons it is often convenient to limit the number of simplices present within the final complex. As such, we have two approximations, the alpha complex and its completion

$$K_{\alpha}^{r}(S) = \left\{\sigma \subseteq S \colon \exists z \in \mathbb{R}^d \text{ s.t. } \|z-x\| \leq r \text{ and } \|z-x\| \leq \|z-y\| \ \forall x \in \sigma \ \forall y \in S\right\}$$
$$K_{\alpha^*}^{r}(S) = \left\{\sigma \subseteq S \colon \{x,y\} \in K_{\alpha}^{r}(S) \ \forall x,y \in \sigma\right\}.$$

These complexes avoid adding simplices between disparate points, controlling the total size of the complex. It has been shown that the alpha and Čech complexes are both homotopy equivalent to a union of closed balls around the underlying point set, thus sharing equivalent homology groups. However, for the completion, denoted here as the alpha* complex, there is no such relationship. The alpha complex is a subcomplex of the Čech complex as well as the Delaunay complex

$$K_{\mathrm{D}}(S) = \left\{\sigma \subseteq S \colon \exists z \in \mathbb{R}^d \text{ s.t. } \|z-x\| \leq \|z-y\| \ \forall x \in \sigma \ \forall y \in S\right\}. \tag{3.3}$$

3.2. Persistent Homology

Now, of chief interest are the topological properties for a given simplicial complex. Both the Čech and Vietoris-Rips complexes reflect the structure present within an underlying point cloud. As such the topology of each provides an effective summary statistic for describing the structural properties of a dataset in $\mathbb{R}^d$. We provide below a short introduction to homology and persistence homology as used in topological data analysis.

Define $C(K)$ to be the free abelian group generated by the simplices in K. Elements of $C(K)$ are sums of the form $\sum_{i\in I} a_i\sigma_i$, where $\sigma_i \in K$ for a_i an appropriate group element. If we further allow the coefficients to come from a field, then $C(K)$ is a vector space. For the purposes of this paper, coefficients are drawn from the two-element field $\mathbb{F}_2 = \{0,1\}$. $C(K)$ is equipped with a linear boundary operator $\partial \colon C(K) \to C(K)$ where $\partial(\{x_1,...,x_{q+1}\}) = \sum_{i=1}^{q}(-1)^i\{x_1,...,x_{i-1},x_{i+1},...,x_{q+1}\}$. As a fundamental property, $\partial \circ \partial = 0$. With coefficients in $\mathbb{F}_2$, the boundary of a simplex reduces to the sum of all its faces. $C_q(K) = C(K_q)$ is the subspace spanned by the q-simplices of K, with the image of $C_q(K)$ under ∂ lying in $C_{q-1}(K)$. $\partial_q \colon C_q(K) \to C_{q-1}(K)$ denotes the restriction of ∂ to $C_q(K)$.

We now construct the homology groups of K. Let $Z(K) = \ker(\partial)$ be the subspace of $C(K)$ containing the cycles, those elements whose boundary under ∂ is 0. $Z_q(K) = Z(K_q) = \ker(\partial_q)$ is the restriction of $Z(K)$ to dimension q. Let $B(K) = \mathrm{im}(\partial)$ denote the subspace of boundaries in $C(K)$. $B_q(K) = B(K_q) = \mathrm{im}(\partial_{q+1})$ is the subspace consisting of the boundaries of elements in $C_{q+1}(K)$, lying in $C_q(K)$.

The *homology groups* are given by $H_q(K) := Z_q(K)/B_q(K)$, the cycles Z_q in dimension q modulo the boundaries B_q. In words, the elements of the homology groups represent "holes" within the simplicial complex, shown by closed loops whose interior is not filled by other elements in the complex. These homology groups provide a topological summary of the structure in the simplicial complex K. As stated previously, because we assume field coefficients for $C(K)$, each homology group is also a vector space. The *Betti numbers* of

the complex represent the degree or dimension of each homology space. We denote the q-th Betti number of K by $\beta_q(K) = \dim(Z_q(K)/B_q(K)) = \dim(Z_q(K)) - \dim(B_q(K))$. Moving forward, Betti numbers and their like will be of primary interest.

Homology provides a topological invariant constructed from a single simplicial complex. For a filtration of nested simplicial complexes, *persistent homology* provides more detail. Given a filtration $\mathcal{K} = \{K^r\}_{r\in\mathbb{R}}$, the homology groups for each complex, $H_q(K^r)$, are defined. However, due to the nested structure of the filtration, simplices are shared across complexes, and thus there exists a natural inclusion map between homology spaces. Cycles in $Z_q(K^r)$ are also cycles in $Z_q(K^t)$ if $r < t$. The boundary spaces behave similarly. For a given equivalence class $x + B_q(K^r) \in H_q(K^r)$, $x + B_q(K^r) \to x + B_q(K^t)$ specifies the inclusion map from $H_q(K^r)$ to $H_q(K^t)$.

If a given element $\tilde{x} \in H_q(K^r)$ maps to $\tilde{y} \in H_q(K^t)$ upon inclusion, with $\tilde{y} \neq B_q(K^t)$, we say that $\tilde{x}$ represents a persistent cycle across the filtration. Essentially the same underlying element is reflected in the homology groups over a range of simplicial complexes. The collection of homology groups and inclusion maps form a *persistence module*. A wide body of work exists on the properties of these persistence modules, see [55] for an introduction. For any cycle feature in the filtration, there is a well defined death time, being the smallest parameter level for which the given element lies in the kernel. The Betti numbers of a filtration form a function in the filtration parameter, r. We use the notation $\beta_q^r(\mathcal{K}) := \beta_q(K^r)$. The Betti numbers in this context count the number of persistent features extant at r.

It is a fundamental theorem of persistent homology that a sufficiently well-behaved persistence module can be represented by a *persistence diagram*. A diagram $\mathcal{D}(\mathcal{K})$ is a multiset in $\mathbb{R}^2 \times \mathbb{Z}$ of points (b, d, q). Each point represents a single persistent feature in the module. b denotes the birth time of the feature, being the smallest parameter level for which that feature is represented in the homology groups. Likewise d gives the death time, and q the dimension of the feature. The collection of persistent features represented by the diagram are a basis for the corresponding persistence module.

The persistence diagram is a simple summary statistic which condenses the complex topological information present within a filtration. An example of a persistence diagram is shown in Figure 1.

3.3. Persistent Betti Numbers

We arrive at the main focus of this section. For $r \leq s$, define the *persistent homology groups* of a filtration $\mathcal{K} = \{K^r\}_{r\in\mathbb{R}}$ as

$$H_q^{r,s}(\mathcal{K}) := Z_q(K^r) / (B_q(K^s) \cap Z_q(K^r)). \tag{3.4}$$

Nonzero elements in this group represent features born at or before time r which persist until at least time s. The dimension of these spaces gives the *persistent Betti numbers*

$$\beta_q^{r,s}(K) := \dim(Z_q(K^r)/B_q(K^s) \cap Z_q(K^r)) \tag{3.5}$$

$$= \dim(Z_q(K^r)) - \dim(B_q(K^s) \cap Z_q(K^r)). \tag{3.6}$$

Persistent Betti numbers are in one-to-one correspondence with the respective persistence diagram. Here $\beta_q^{r,s}(\mathcal{K})$ counts the number of points in $\mathcal{D}(\mathcal{K})$ of feature dimension q falling within $(-\infty, r] \times (s, \infty]$. When $s = r$, we recover the regular Betti numbers, $\beta_q^{r,r}(\mathcal{K}) = \beta_q(K^r)$. An important result for persistent Betti numbers is given in the following lemma.

Lemma 3.1 (Geometric Lemma)**.** *[Lemma 2.11 in [27]] Let* $\mathcal{J} = \{J^r\}_{r\in\mathbb{R}}$ *and* $\mathcal{K} = \{K^r\}_{r\in\mathbb{R}}$ *be filtrations of simplicial complexes with with* $J^r \subseteq K^r$ *for all* $r \in \mathbb{R}$*. Then*

$$\left|\beta_q^{r,s}(\mathcal{K}) - \beta_q^{r,s}(\mathcal{J})\right| \leq \max\left\{\#\left\{K_q^r \setminus J_q^r\right\}, \#\left\{K_{q+1}^s \setminus J_{q+1}^s\right\}\right\} \tag{3.7}$$

$$\leq \#\left\{K_q^r \setminus J_q^r\right\} + \#\left\{K_{q+1}^s \setminus J_{q+1}^s\right\}. \tag{3.8}$$

The Geometric Lemma 3.1 relates the change in persistent Betti numbers between two filtrations to the additional simplices gained moving between them. As a brief explanation of the lemma, simplices can be divided into two classes, positive and negative. For two simplicial complexes $J \subset K$, if we imagine adding the additional q-simplices in K to J one by one, a positive q-simplex will increase the dimension of Z_q by one, and a negative q-simplex will increase the dimension of B_{q-1} by one. Either change can affect the persistent Betti numbers. This dichotomy is a basic result from persistent homology, see [7]. The bound given in the Geometric Lemma describes a worst case, when all q-simplices at time r are positive or all $(q+1)$-simplices at time s are negative. The Geometric Lemma will be critical moving forward, as it allows us to control the change in persistent Betti numbers by counting appropriate simplices.

3.4. Euler Characteristic

For a given simplicial complex K, the *Euler characteristic* is defined as

$$\chi(K) := \sum_{k=0}^{\infty} (-1)^k \#\{K_k\}. \tag{3.9}$$

Provided there is an $m \in \mathbb{N}$ such that the Betti numbers $\beta_q(K)$ are 0 for all $q > m$ (as in (D4) holds), it can be shown that the Euler characteristic has the following identity with the Betti numbers:

$$\chi(K) = \sum_{k=0}^{\infty} (-1)^k \beta_k(K). \tag{3.10}$$

This relationship with the Betti numbers makes the Euler characteristic an important topological invariant in its own right. Applications of the Euler characteristic and derivatives may be found in [41, 43, 49].

3.5. k-Nearest Neighbor Graph

The k-nearest neighbor graph $\mathcal{K}_{\mathrm{NN},k}$ of a vertex set S connects each point $x \in S$ with the k closest vertices to x within $\mathbf{S} \setminus x$. This graph may either be directed or undirected. $\mathcal{K}_{\mathrm{NN},k}$ is commonly used to analyze the clustering structure of a point cloud. Let the total length of the edges in this graph be denoted by $l_{\mathrm{NN},k}$. The total length of the k-nearest neighbor graph, when suitably scaled, provides a measure of the average local "density", or concentration of the points in S. In Section 4.5, we will show bootstrap consistency for $l_{\mathrm{NN},k}$ within the stabilization framework.

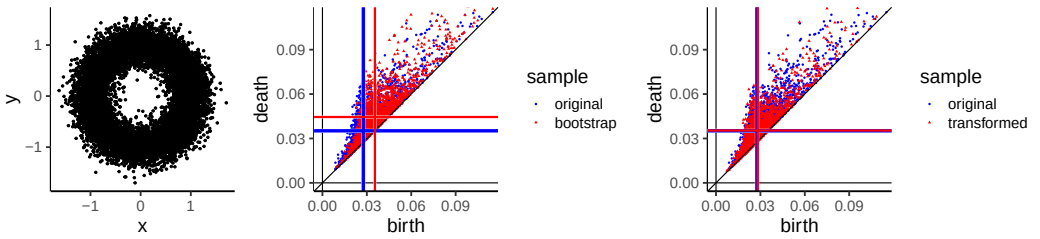

FIG 1. Left: *The original data set of size* $n = 10,000$*, from which a single standard bootstrap sample is drawn.* Middle: *Persistence diagrams for both the original and bootstrap samples, along with lines denoting the median birth and death in each diagram. The asymptotic bias discussed in Section 4.1 can be clearly seen.* Right: *Persistence diagrams after application of a multiplicative correction factor of* $\sqrt{1-e^{-1}} \approx 0.795$ *to the bootstrap sample. Note that the median birth/death times correspond after transformation.*

4. Bootstrapping Topological Statistics

4.1. Nonparametric Bootstrap

In this section, we will argue that the standard nonparametric bootstrap may fail to reproduce the correct sampling distribution asymptotically when applied to common topological statistics.

For a wide class of simplicial complexes built over point sets in $\mathbb{R}^d$, the corresponding persistence diagram is unaffected by the inclusion of repeated points within the vertex set. This behavior holds for both the Vietoris-Rips and Čech complexes, defined in Section 3.1. In the case of the Čech complex, this phenomenon is seen most directly. The Čech complex under the Euclidean metric is homologically equivalent to a union of closed balls centered on the vertex points in $\mathbb{R}^d$. Additional repetitions of vertex points leave both this union and the derived persistence diagram unchanged.

In these cases where repetitions may be ignored in the calculation of statistics, the standard bootstrap behaves effectively like a subsampling technique. The size of a given subsample is random, equal to the number of unique points present in the corresponding bootstrap sample.

Given a random sample $\mathbf{X}_n = \{X_1, ..., X_n\}$, it can be shown using elementary arguments that a given bootstrap sample $\mathbf{X}_n^*$ of size n from the empirical distribution over $\mathbf{X}_n$ is expected to contain $n(1-(1-1/n)^n) \approx \left(1-e^{-1}\right) n \approx 0.632n$ unique points. As such, $\mathbf{X}_n^*$ behaves similarly to a sample of size $0.632n$, but is not scaled accordingly within the statistic $\left(\beta_q^{r,s}\left(\sqrt[d]{n}\mathbf{X}_n^*\right) - \mathbb{E}\left[\beta_q^{r,s}\left(\sqrt[d]{n}\mathbf{X}_n^*\right)\middle|\mathbf{X}_n\right]\right) / \sqrt[d]{n}$. This discrepancy in scaling introduces a non-negligible asymptotic bias. The effect is illustrated in Figure 1 for the Vietoris-Rips complex.

Furthermore, the standard nonparametric bootstrap results in a fundamentally different point process limit at small scales when compared to the original sample. For the original sample, when $\mathbf{X}_n$ is drawn from a distribution with density f, the shifted and rescaled sample $\sqrt[d]{n}\left(\mathbf{X}_n - z\right)$ approaches a homogeneous Poisson process $\mathbf{P}_z$ with intensity $f(z)$. From the preceeding stabilization literature ([38], [31]), this limiting local point process drives the asymptotic sampling distribution of $\left(\beta_q^{r,s}\left(\sqrt[d]{n}\mathbf{X}_n\right) - \mathbb{E}\left[\beta_q^{r,s}\left(\sqrt[d]{n}\mathbf{X}_n\right)\right]\right) / \sqrt[d]{n}$. Considering the large-sample behavior of $\sqrt[d]{n}\left(\mathbf{X}_n^* - z\right)|\mathbf{X}_n$, the smoothed bootstrap sampling procedure described in Section 2.4 can be shown to reproduce the same local Poisson process

$\mathbf{P}_z$ asymptotically.

However, the same is not true for the standard bootstrap when repeated points are ignored. In this case, $\sqrt[d]{n}\,(\mathbf{X}_n^* - z)\,|\mathbf{X}_n$ is restricted to the discrete set $\sqrt[d]{n}\,(\mathbf{X}_n - z)$, and thus cannot reproduce $\mathbf{P}_z$, whose domain is $\mathbb{R}^d$. For this case, we describe the resulting point process limit $\mathbf{Q}_z$ in two steps. First, a homogenous Poisson process $\mathbf{P}_z$ is generated, representing $\sqrt[d]{n}\,(\mathbf{X}_n - z)$. Defined conditionally, $\mathbf{Q}_z|\mathbf{P}_z$ is a random subset of $\mathbf{P}_z$ such that $\mathbb{P}\left[x \in \mathbf{Q}_z|\mathbf{P}_z\right] = 1 - e^{-1} \approx .632$, considering each point $x \in \mathbf{P}_z$ independently. We have $\sqrt[d]{n}\,(\mathbf{X}_n^* - z) \to \mathbf{Q}_z$.

This difference in local behavior, combined with the asymptotic bias effect illustrated earlier, are strong indicators that $\left(\beta_q^{r,s}\left(\sqrt[d]{n}\mathbf{X}_n^*\right) - \mathbb{E}\left[\beta_q^{r,s}\left(\sqrt[d]{n}\mathbf{X}_n^*\right)\right]\big|\mathbf{X}_n\right) / \sqrt[d]{n}$ and $\left(\beta_q^{r,s}\left(\sqrt[d]{n}\mathbf{X}_n\right) - \mathbb{E}\left[\beta_q^{r,s}\left(\sqrt[d]{n}\mathbf{X}_n\right)\right]\right) / \sqrt[d]{n}$ likely do not share a weak limit. A technical treatment is omitted here, and is outlined merely to justify the use of our smoothed bootstrap procedure in place of the standard nonparametric bootstrap. The smoothed bootstrap procedure provides for bootstrap consistency (Corollaries 4.2 and 4.3), and in the following sections we consider only this approach.

4.2. General Conditions for Simplicial Complexes

The results presented in the following sections apply for a range of simplicial complexes constructed over point clouds in $\mathbb{R}^d$. Here we will explain the specific conditions used, and for which common simplicial complexes they apply. Let K be a function taking as input $S \in \tilde{\mathcal{X}}\left(\mathbb{R}^d\right)$, giving as output a simplicial complex with vertices in S. For a given simplex σ, let the set diameter be $\operatorname{diam}(\sigma)$. We have the following conditions:

(K1) For any $S \in \tilde{\mathcal{X}}\left(\mathbb{R}^d\right)$ and $z \notin S$, $K(S) \subseteq K(S \cup \{z\})$. Furthermore, $\sigma \in K(S \cup \{z\}) \setminus K(S)$ only if $z \in \sigma$.
(K2) For any $S \in \tilde{\mathcal{X}}\left(\mathbb{R}^d\right)$ and $z \in \mathbb{R}^d$, $\sigma \in K(S)$ only if $\sigma - z \in K(S - z)$.

(D1) There exists $\phi < \infty$ such that for any $S \in \tilde{\mathcal{X}}\left(\mathbb{R}^d\right)$, $\sigma \in K^r(S)$ only if $\operatorname{diam}(\sigma) \leq \phi$.
(D2) There exists $\phi < \infty$ such that for any $S \in \tilde{\mathcal{X}}\left(\mathbb{R}^d\right)$ and $z \in \mathbb{R}^d$, $\sigma \in K(S \cup \{z\}) \triangle K(S)$ only if $\sigma \subset B_z(\phi)$.
(D3) There exists an $\eta > 0$ such that for any $S \in \tilde{\mathcal{X}}\left(\mathbb{R}^d\right)$ and $x \in Z(K(S))$, $\operatorname{diam}(x) \leq \eta$ only if $x \in B(K(S))$.
(D4) There exists an $m \in \mathbb{N}$ such that for any $k > m$ and $S \in \tilde{\mathcal{X}}\left(\mathbb{R}^d\right)$, $Z_k(K(S)) = B_k(K(S))$.

(K1) means that the addition of a new point will not change the existing complex, only add new simplices. Furthermore, any new simplices gained must contain the added point as a vertex. (K2) gives that the complex is essentially translation invariant. (D1) sets a maximum diameter for any simplex in the complex. (D2) gives that the influence of a new point on the complex is confined to a local region around that point, within a fixed diameter. This condition allows for both the addition and removal of simplices from the complex, but only within the prescribed radius. It can be easily shown that if (D2) holds for ϕ, (D1) holds for 2ϕ. Conversely if both (K1) and (D1) hold for ϕ, (D2) also holds for ϕ. Finally, (D3) gives that no small loops can exist with unfilled interiors, and (D4) gives that all Betti numbers are 0 in sufficiently high feature dimensions.

Now, let $\mathcal{K} = \left(K^r\right)_{r \in \mathbb{R}}$ be a function taking as input $S \in \tilde{\mathcal{X}}\left(\mathbb{R}^d\right)$, giving as output a filtration of simplicial complexes with vertices in S. As a slight abuse, we will often refer to the function $\mathcal{K}$ as a filtration of simplicial complexes, even though it is a function defining

more than a single filtration, depending on the underlying point cloud. We say that a given condition is satisfied for $\mathcal{K}$ if it is satisfied by K^r for any $r \in \mathbb{R}$. In the cases of (D1), (D2), and (D3), ϕ and η may depend on r as increasing functions $\phi\colon \mathbb{R} \to [0,\infty)$ and $\eta\colon \mathbb{R} \to [0,\infty)$.

It can be shown that all of (K1)-(D3) are satisfied for both the Vietoris-Rips and Čech complexes in $\mathbb{R}^d$ using $\phi(r) = \eta(r) = 2r$. The same functions apply for the alpha complex in $\mathbb{R}^d$ and its completion $\mathcal{K}_{\alpha^*}$, with the notable exception that (K1) is violated. Finally, it is known that (D4) is satisfied by the alpha, Čech, and Delauney complexes in $\mathbb{R}^d$ for $m = d-1$.

While covering a wide class of distance-based simplicial complexes, there are several complexes used in practice that may fail to satisfy any or all of these. For example, the addition of a new point to the Delaunay complex, Gabriel graph, witness complex, or k-nearest neighbor graph can both add and remove simplices, violating (K1). Furthermore, there is not any limit on the simplex diameter within any of these complexes, violating (D1). Likewise, the addition of a single point can alter simplices at arbitrarily large distances, violating (D2). As a special note, it is common in practice to consider the intersection of the Vietoris-Rips and Delaunay complexes, which unfortunately may violate all the assumptions here. It is unclear if an extension or special consideration could be made to incorporate these complexes.

4.3. Stabilization of Persistent Betti Numbers

To apply the general bootstrap theorem, we first require a technical lemma establishing a locally-determined radius of stabilization for persistent Betti numbers. The result given applies for general classes of simplicial complexes constructed over subsets of $\mathbb{R}^d$, using the conditions listed previously. Reiterating, $\mathcal{C}_{p,M}\left(\mathbb{R}^d\right)$ is the class of distributions G on $\mathbb{R}^d$ with densities g such that $\|g\|_p \leq M$. We have the following:

Lemma 4.1. *Let $F \in \mathcal{C}_{p,M}\left(\mathbb{R}^d\right)$ for some $p > 2$ and $M < \infty$, and let $\mathcal{K} = \{K^r\}_{r\in\mathbb{R}}$ be a filtration of simplicial complexes satisfying (K2), (D2), and (D3). Then for any $r \in \mathbb{R}$, $s \in \mathbb{R}$, and $q \geq 0$, $\beta_q^{r,s}(\mathcal{K})$ satisfies (S2) for F.*

4.4. Bootstrap Results for Persistence Homology

Here we present the main applied results of this paper. Each is derived from Theorem 2.7 and the stabilization lemma for persistent Betti numbers (Lemma 4.1). For given vectors of birth and death times, $\vec{r} = (r_i)_{i=1}^k$ and $\vec{s} = (s_i)_{i=1}^k$, let $\beta_q^{\vec{r},\vec{s}} = \left(\beta_q^{r_i,s_i}\right)_{i=1}^k$ denote the multivariate function whose components are the persistent Betti numbers evaluated at each pair of birth and death times. For a vector of filtration times $\vec{r} = (r_i)_{i=1}^k$, let $\chi^{\vec{r}}$ denote the function giving the Euler characteristic at each time r_i, with $\chi^{\vec{r}} := (\chi(K^{r_i}))_{i=1}^k$.

The following apply for $F \in \mathcal{P}\left(\mathbb{R}^d\right)$ with density f such that $\|f\|_p < \infty$ for some $p > 2$, as specified. F and $\hat{F}_n$ are such that $\hat{F}_n$ has density $\hat{f}_n$, $\|\hat{f}_n - f\|_1 \to 0$, and $\|\hat{f}_n - f\|_p \to 0$ in probability (resp. *a.s.*). Let $\mathbf{X}_n = \{X_i\}_{i=1}^n \overset{\text{iid}}{\sim} F$ and $(m_n)_{n\in\mathbb{N}}$ such that $\lim_{n\to\infty} m_n = \infty$. $\mathbf{X}_{m_n}^* = \{X_i^*\}_{i=1}^{m_n} \overset{\text{iid}}{\sim} \hat{F}_n \big| \mathbf{X}_n$ is a bootstrap sample and G a multivariate distribution. Recalling the conclusion of Theorem 2.7, for a multivariate statistic $\vec{\psi}$:

Statement 4.1.

$$\frac{1}{\sqrt{n}}\left(\vec{\psi}\left(\sqrt[d]{n}\mathbf{X}_n\right) - \mathbb{E}\left[\vec{\psi}\left(\sqrt[d]{n}\mathbf{X}_n\right)\right]\right) \xrightarrow{d} G$$

if and only if

$$\frac{1}{\sqrt{m_n}}\left(\vec{\psi}\left(\sqrt[d]{m_n}\mathbf{X}^*_{m_n}\right) - \mathbb{E}\left[\vec{\psi}\left(\sqrt[d]{m_n}\mathbf{X}^*_{m_n}\right)\middle|\mathbf{X}_n\right]\right) \xrightarrow{d} G \text{ in probability (resp. a.s.).}$$

For cases with a corresponding central limit theorem, G is the limiting normal distribution of the original standardized statistic.

Corollary 4.2 (**Persistent Betti Numbers**). *Let $q \geq 0$ and $p > 2q+3$. Let $\mathcal{K}$ be a filtration of simplicial complexes satisfying* (K1), (K2), (D1), *and* (D3). *Then for any given $\vec{r}$, $\vec{s}$, Statement 4.1 holds for $\beta_q^{\vec{r},\vec{s}}$.*

Corollary 4.3 (**Persistent Betti Numbers - Alt.**). *Let $q \geq 0$ and $p > 2q+5$. Let $\mathcal{K}$ be a filtration of simplicial complexes satisfying* (K2), (D2), *and* (D3). *Then for any given $\vec{r}$, $\vec{s}$, Statement 4.1 holds for $\beta_q^{\vec{r},\vec{s}}$.*

The only differences between the above corollaries are the conditions satisfied by the underlying simplicial complex and the necessary norm bound on the density. The corresponding results for the Betti numbers follow as special cases of Corollaries 4.2 and 4.3, when the given birth and death parameters are equal ($\beta_q^{\vec{r}} = \beta_q^{\vec{r},\vec{r}}$). Also, although the statements of Corollaries 4.2 and 4.3 are given in terms of a fixed feature dimension q, a direct extension exists if $q = q_i$ is allowed to differ for each (r_i, s_i). The form as given shows the dependence of the density norm assumption on the chosen feature dimension.

The higher value of p required in Corollary 4.3 compared to Corollary 4.2 can be explained intuitively based on the assumptions used. For the persistent Betti numbers, the main quantity controlling convergence is the expected number of simplices altered or introduced when a new datapoint is added to the sample. (D2) ensures that these simplices fall within a small ball around the new data point. The stated density norm conditions control the expected number of points, and by extension possible simplices, that can lie within that small ball. Introducing (K1) further controls the number of possible simplices, and allows for a weakening of the necessary norm condition. (K1) requires that, as the sample grows by a single point, any additional simplices must contain the new point as a vertex, and no deletion of simplices is possible. This means that every added simplex has one less "free" vertex, and a weaker norm condition is required for control. The same intuition applies whenever (K1) is assumed.

In the specific case of the alpha complex, both of the above Corollaries 4.2 and 4.3 apply. While the alpha complex does not satisy (K1), it has equal persistent Betti numbers to the Čech complex, which does. Thus, the weaker conditions of Corollary 4.2 are sufficient in this unique case.

Corollary 4.4 (**Euler Characteristic**). *Let $m < \infty$ and $p > 2m+3$. Let $\mathcal{K}$ be a filtration of simplicial complexes satisfying* (K1), (K2), (D1), (D3), *and* (D4). *Then for any given $\vec{r}$, Statement 4.1 holds for $\chi^{\vec{r}}$.*

Corollary 4.5 (**Euler Characteristic - Alt.**). *Let $m < \infty$ and $p > 2m+5$. Let $\mathcal{K}$ be a filtration of simplicial complexes satisfying* (K2), (D2), (D3), *and* (D4). *Then for any given $\vec{r}$, Statement 4.1 holds for $\chi^{\vec{r}}$.*

It is suspected that some of the simplicial complex assumptions can be relaxed in the persistent Betti number and Euler characteristic cases, but the extent to which this is possible is still unknown. Specifically, Corollary 4.2 requires a translation-invariant simplicial complex (K2), along with the elimination of small loops via (D3). See Appendix A for altered "B-bounded persistent Betti number" and "q-truncated Euler characteristic" problem settings where these issues may be resolved.

To strengthen Corollaries 4.2-4.5 with rates, we require more specific knowledge about the convergence to G of the original statistic. For persistent Betti numbers in the multivariate setting, general central limit theorems have been shown in [31], but little is known at this time with regards to rates of convergence. Proposition 2.6 does allow for rates of convergence in 2-Wasserstein distance between the bootstrap and true sampling distributions for finite sample sizes, but is phrased in terms of a tail probability for the radius of stabilization. See the proofs of Corollaries 4.2-4.5 for details. For persistent Betti numbers the tail behavior of the radius of stabilization is poorly understood. Owing to these difficulties, we may only conclude consistency of the smoothed bootstrap for the functions considered.

4.5. Bootstrap Results for k-Nearest Neighbor Graphs

In the following, let $\mathcal{D}_{\gamma,r_0}(C)$ be the class of distributions G with support on a bounded $C \subset \mathbb{R}^d$ such that $\int_{B_x(r)} \mathrm{d}G \geq \gamma r^d$ for all $r \leq r_0$ and $x \in C$.

Corollary 4.6 (Total Edge Length of the k-Nearest Neighbor Graph). *Let $p > 2$. Furthermore, let $F \in \mathcal{D}_{\gamma,r_0}(C)$ and $\mathbb{1}\left\{\hat{F}_n \in \mathcal{D}_{\gamma,r_0}(C)\right\} \to 1$ in probability (resp. a.s.). Then Statement 4.1 holds for $l_{NN,k}$.*

The conditions of Corollary 4.6 are in particular satisfied when C is known and convex, with f bounded below on C by a constant, provided further that $\|\hat{f}_n - f\|_\infty \to 0$ in probability (resp. a.s.). We include this final result to demonstrate the utility of stabilization as a general tool for proving bootstrap convergence theorems outside of topological data analysis. The k-nearest neighbor graph does not fall under the general simplicial complex conditions provided in Section 4.2, thus special treatment is needed to show the required stabilization and moment conditions. Here we rely on previous results from the literature, see [38] for stabilization results and the corresponding central limit theorem.

5. Simulation Study

In this section we present the results of a series of simulations illustrating the finite-sample properties of the smoothed bootstrap applied to persistent Betti numbers $\beta_q^{r,s}$ of the Vietoris-Rips complex constructed over point sets in $\mathbb{R}^d$. Precise definitions and an introduction to the properties of these statistics may be found in Section 3. Source code for this section, as well as for the data analysis of Section 6 is available at
github.com/btroycraft/stabilizing_statistics_bootstrap [44].

We investigate the coverage probability of bootstrap confidence intervals on the expected persistent Betti numbers $\mathbb{E}\left[\beta_q^{r,s}\left(\sqrt[d]{n}\mathbf{X}_n\right)\right]$ for a variety of feature dimensions, sample sizes, data generating mechanisms, and bandwidth selectors. Table 1 lists brief descriptions of the data distributions considered. For more detailed explanations, see Appendix D. The results of the simulations are given in Table 2. For the persistent Betti numbers, a single choice of (r, s) was made for each combination of distribution and feature dimension, chosen to lie within

Label	Description
F_1	Rotationally symmetric in $\mathbb{R}^2$, finite L_8 norm
F_2	Rotationally symmetric in $\mathbb{R}^2$, finite L_2 norm, infinite L_8 norm
F_3	$\mathbb{S}^1$ embedded in $\mathbb{R}^2$, additive Gaussian noise
F_4	Uniformly distributed over $B_0(1)$ in $\mathbb{R}^3$, additive Gaussian noise
F_5	5 clusters in $\mathbb{R}^3$, additive exponential noise
F_6	$\mathbb{S}^2$ embedded in $\mathbb{R}^5$, additive Cauchy noise
F_7	Flat figure-8 embedded in $\mathbb{R}^{10}$, additive Gaussian noise

TABLE 1

Description of densities or distributions considered for the simulation study of Section 5. For the distributions based on manifolds, we first draw uniformly from the manifold, then apply the prescribed additive noise. Detailed explanations of the distributions considered, along with precise definitions are available in Appendix D.

the main body of features in the corresponding persistence diagram. For computational reasons, only feature dimensions $q = 1$ and $q = 2$ are considered.

We consider five data-driven bandwidth selectors. First are the "Hpi.diag" (plug-in), "Hlscv.diag" (least-squares cross-validation), and "Hscv.diag" (smoothed cross-validation) selectors from the *ks* package in R. Second, we include the adaptive bandwidth selector described in Section 6. While this selector is tailored for the specifics of astronomical data, we include it here for completeness. Each of these four selectors are available for data dimension up to $d = 6$. Last, we consider Silverman's rule of thumb (see [46]) via "bw.silv" from the *kernelboot* package in R, which accepts data in any dimension.

For the two cross-validation selectors, note that a bandwidth is not always selected, throwing errors on some datasets. To accommodate the automatic setting of this simulation study, any error-producing data sets were simply rejected for each of these cases.

There is a noticeable drop-off in coverage as the data dimension increases. This is expected, as the kernel density estimator is known to suffer from a "curse of dimensionality". For distribution F_6, which exhibits heavy tails, only the adaptive bandwidth selector performed well, because outliers are weighted much less heavily in this case. It is likely that performance will suffer generally in the presence of heavy tailed data when using one of the selectors with common bandwidth.

The coverage proportion is generally smaller than the nominal level of 95%. Therefore, it is recommended to use a larger than desired level, especially for limited sample sizes. In terms of general performance, we recommend any of "Hpi.diag", "Hlscv.diag", or "Hscv.diag". These selectors provide the most consistent coverage, and effectively replicate the nominal 95% level in many cases, especially for the largest sample size $n = 400$. Silverman's rule performs badly in several cases, and should only be used in the absence of better alternatives.

6. Data Analysis

In this section we show how smoothed bootstrap estimation performs on a real dataset. We consider a selection of galaxies from the Sloan Digital Sky Survey [5], chosen from a selection of sky with right ascension values between 100° and 270° and declination between −7° and 70°. Three slices of galaxies were considered, separated by redshift, a measure of radial distance from the solar system. The selections consist of galaxies with red-shift within $(0.025, 0.026)$, $(0.027, 0.028)$, and $(0.029, 0.030)$, respectively. These slices were chosen to investigate the topological properties of the cosmic web across time. In this case, due to the rough homogeneity of the web at large scales, few significant topological deviations are

Distr.	F_1	F_2	F_3	F_4	F_5	F_6	F_7	F_4	F_5	F_6	F_7
	$q = 1$							$q = 2$			
r	4.94	5.20	3.03	1.92	0.30	1.78	1.28	2.96	0.39	2.71	1.46
s	5.36	5.60	3.28	2.12	0.31	1.91	1.32	3.04	0.40	2.80	1.47
$n = 100$	0.896	0.965	0.921	0.859	0.954	0.19		0.908	0.705	0.038	
	0.931	0.959	0.914	0.809	0.941	0.133		0.903	0.604	0.045	
	0.903	0.97	0.91	0.859	0.927	0.049		0.902	0.363	0.002	
	0.922	0.898	0.899	0.71	0.725	0.736		0.837	0.048	0.051	
	0.359	0.931	0.942	0.864	0	0	0.656	0.902	0	0	0.045
$n = 200$	0.908	0.971	0.94	0.898	0.942	0.159		0.878	0.795	0.125	
	0.92	0.972	0.946	0.891	0.923	0.106		0.872	0.707	0.074	
	0.888	0.975	0.959	0.906	0.892	0.06		0.908	0.277	0.031	
	0.888	0.909	0.828	0.783	0.773	0.705		0.673	0.032	0.27	
	0.299	0.954	0.903	0.899	0	0	0.766	0.882	0	0	0.537
$n = 300$	0.9	0.971	0.926	0.921	0.94	0.183		0.854	0.906	0.225	
	0.94	0.971	0.938	0.896	0.94	0.087		0.854	0.917	0.072	
	0.913	0.971	0.94	0.896	0.922	0.054		0.855	0.964	0.074	
	0.93	0.923	0.864	0.786	0.771	0.735		0.712	0.551	0.575	
	0.283	0.956	0.925	0.906	0	0	0.835	0.856	0	0	0.508
$n = 400$	0.918	0.961	0.947	0.934	0.96	0.175		0.851	0.883	0.259	
	0.927	0.951	0.938	0.92	0.955	0.063		0.839	0.88	0.076	
	0.908	0.976	0.933	0.924	0.939	0.062		0.863	0.958	0.099	
	0.911	0.922	0.874	0.813	0.825	0.771		0.695	0.952	0.789	
	0.266	0.961	0.909	0.922	0.114	0	0.891	0.859	0	0	0.584

TABLE 2

Coverage proportions for 95% *smoothed bootstrap confidence intervals on the mean persistent Betti numbers; coverage is estimated using* $N = 1,000$ *independent base samples with* $B = 500$ *bootstrap samples each. True mean persistent Betti numbers are estimated using a large* ($N = 100,000$) *number of independent samples from the true distribution. For each case, the values from top to bottom: Coverage proportions using "Hpi.diag", "Hlscv.diag", "Hscv.diag", "adaptive", and "bw.silv" bandwidth selectors, respectively. (see Section 5)*

expected.

Subset limits were chosen to maintain computational feasibility and avoid measurement gaps. In an initial cleaning step, each slice was flattened using an area-preserving cylindrical projection and trimmed so that the slices share a common boundary with the same number of galaxies (2374) per slice. Angular units are converted to distances in Megaparsecs (Mpc) based on the redshift and Hubble's constant.

The distribution of galaxies in each dataset is modeled by a random sample from some bivariate probability distribution, where the location of each galaxy is drawn independently from the overall distribution. As a part of the model framework, the effect of gravitational interaction manifests via a macroscopic change in the matter distribution, rather than as dependency between individual galaxies.

Following the recommendation of [24], we estimate the density of the matter distribution using the adaptive bandwidth selector described in [8]. This adaptive bandwidth selector was chosen to accommodate for the large variations in density present within astronomy data. The selectors considered in Section 5 do not perform well in this context, often oversmoothing by a large margin. A pilot density estimator was constructed based on the "Hpi.diag" plug-in bandwidth selector and a Gaussian kernel.

Visualizations of the density estimates are provided in Figure 2. Generally, the fit adequately captures the filament structures present in the raw data. Within the persistence diagrams, the mass of features present close to the main diagonal represents small-scale holes between neighboring galaxies, whereas features farther from the diagonal represent the large-scale holes formed by relatively disparate galaxies.

We apply the Vietoris-Rips complex to each of the slices, and calculate a selection of persistent Betti numbers in dimensions $q = 0$ and $q = 1$. The 0-dimensional features summarize cluster and filament structure, whereas the 1-dimensional features describe voids and depressions. The transformed datasets and persistence diagrams in dimension $q = 1$ can be seen in Figure 2. We consider the Betti numbers β_0^r and β_1^r, as well as the persistent Betti numbers $\beta_1^{r,r+1}$ for $r = 3, ..., 30$ Mpc. Filtration parameters for the persistent Betti numbers were chosen to lie close to the diagonal $r = s$, excluding features with a lifetime less than 1 Mpc. We use bootstrap estimation to construct nominal 98% confidence intervals for the population mean values, both pointwise and simultaneous within each regime across $r = 3, ..., 30$ Mpc. The number of bootstrap replicates used was $B = 20,000$, with results seen in Figure 3.

In feature dimension $q = 0$, the curves show similar behavior across the slices. Consistent with our empirical results, similar Betti curves are expected when the within-filament matter distribution and overall frequency of filaments for each sample are equal. For feature dimension $q = 1$, more variation is present. However, as can be seen from the bootstrap confidence intervals, much of this variation is explained by random fluctuation. For example, while a notable depression around the scale of 8 Mpc exists for the third slice, it is still within the margins of error provided. From this analysis, we do not find significant differences in the topological properties of the three samples over the range of filtration parameters considered. The difference in topological structure seen within each pair of Betti curves is within the margin of error provided by the bootstrap confidence intervals, especially considering the wider simultaneous intervals.

The consistency shown in Section 4.4 for bootstrap estimation applies only for those features within the "body" of topological features, being those occurring at a local scale. Features with large persistence or ones that appear at large diameter are not accounted for in this, as their relative weight is small within the persistent Betti numbers. As such, our

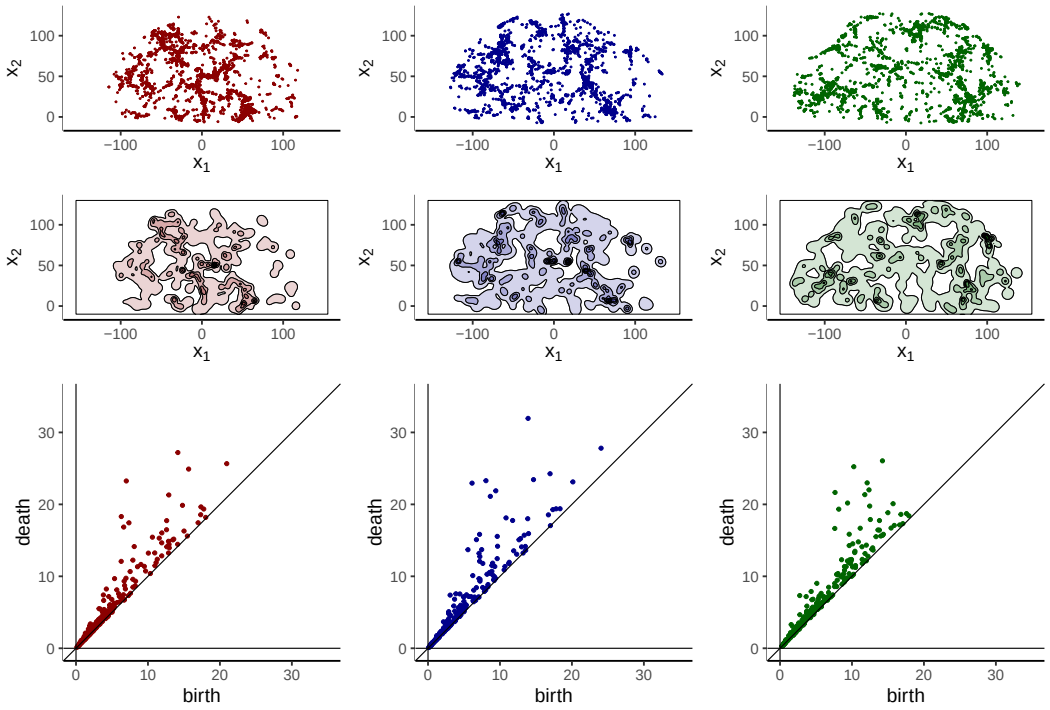

FIG 2. *Top row: Transformed point clouds. Middle row: Density estimates using adaptive bandwidth. Bottom row: Persistence diagrams in dimension $q = 1$ for the Vietoris-Rips complex. Columns from left to right: Galaxies with redshifts within* $(0.025, 0.026)$*,* $(0.027, 0.028)$*, and* $(0.029, 0.030)$*, respectively. Axis units are given in Megaparsecs (Mpc).*

analysis does not preclude differences in topology at a large relative scale, describing the largest galactic structures.

7. Discussion

In this work we have shown the large-sample consistency of multivariate bootstrap estimation for a range of stabilizing statistics. This includes the persistent Betti numbers, the Euler characteristic, and the total edge length of the k-nearest neighbor graph. However, many open questions still remain.

In Section 4.1 it was argued that the standard nonparametric bootstrap may fail to directly reproduce the correct sampling distribution asymptotically for topological statistics like the persistent Betti numbers. However, there remains the possibility that a corrected version of the standard bootstrap could provide for consistency. As discussed in Section 4.1, standard bootstrap sampling results in a fundamentally different point process limit at small scales. Previous stabilization results primarily consider Poisson and related processes, meaning a full theoretical treatment of the standard bootstrap would likely require reconstructing much of the previous stabilization and central limit theorem results for the alternative limiting process.

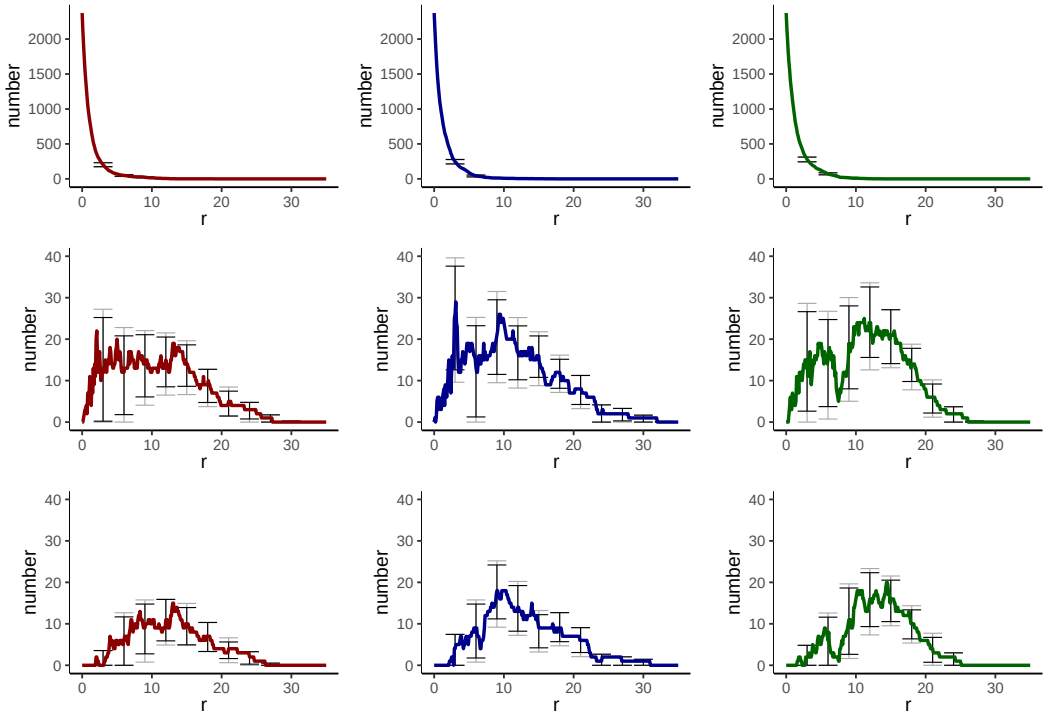

FIG 3. *Betti curves for the Vietoris-Rips complex. Top row: Betti numbers* β_0^r. *Middle row: Betti numbers* β_1^r. *Bottom Row: persistent Betti numbers* $\beta_1^{r,r+1}$. *Columns correspond with those of Figure 2. Axis units are given in Megaparsecs (Mpc). For each of* $r = 3, ..., 30$ *Mpc, simultaneous bootstrap confidence bands are given in gray, drawn from bootstrap samples of size* $B = 20,000$. *Likewise, pointwise intervals are given in black.*

The results for the smoothed bootstrap presented here apply only in the multivariate setting, the obvious extension being to stochastic processes. Essential to a process-level result concerning the persistent Betti numbers would be a convenient tail bound for the radius of stabilization, which is yet unavailable. In the case of persistent Betti numbers, there is a strong relationship between the persistent Betti function and an empirical CDF in two dimensions. As such, there is much established theory in that regard which may be applied once stochastic equicontinuity is established.

In practice it is common that data comes not from a density in $\mathbb{R}^d$, but instead from a manifold. It is suspected that a version of the results in this paper could apply in the manifold setting. However, this requires a bootstrap that adapts to a possibly unknown manifold structure, similar to that found in [28]. Combined with the inherent challenges of working with manifolds, this extension presents many technical hurdles.

Furthermore, in this work we have shown only consistency for bootstrap estimation to a common limiting distribution. The rates of convergence in the 2-Wasserstein distance regarding the persistent Betti numbers rely on the unknown tail properties of the corresponding radius of stabilization. Quantifying these tail properties is a challenging open problem, and seems to be a key step towards an eventual rate calculation, as well as the previously mentioned process-level result.

Finally, there are several statistics of interest, including those based on the Delaunay complex, which do not fit into the specific frameworks provided here. It may be that these statistics may still satisfy Theorem 2.7 in the general case, by techniques others than those provided here.

Acknowledgements

Thank you to the reviewers for their helpful comments and thorough examination of this work.

Benjamin Roycraft was partially supported by the National Science Foundation (NSF), grant number DMS-1148643. Johannes Krebs was partially supported by the German Research Foundation (DFG), grant number KR-4977/2-1. Wolfgang Polonik was partially supported by the National Science Foundation (NSF), grant number DMS-2015575.

Funding for the Sloan Digital Sky Survey IV has been provided by the Alfred P. Sloan Foundation, the U.S. Department of Energy Office of Science, and the Participating Institutions. SDSS-IV acknowledges support and resources from the Center for High-Performance Computing at the University of Utah. The SDSS web site is www.sdss.org.

References

[1] Adler, R. J., Agami, S. and Pranav, P. (2017). Modeling and Replicating Statistical Topology and Evidence for CMB Nonhomogeneity. *Proc. Natl. Acad. Sci. USA* **114** 11878–11883. MR3725115

[2] Aldous, D. and Steele, J. M. (1992). Asymptotics for Euclidean Minimal Spanning Trees on Random Points. *Probab. Theory Related Fields* **92** 247–258. MR1161188

[3] Arsuaga, J., Borrman, T., Cavalcante, R., Gonzalez, G. and Park, C. (2015). Identification of Copy Number Aberrations in Breast Cancer Subtypes Using Persistence Topology. *Microarrays* **4** 339–369.

[4] Biscio, C. A. N., Chenavier, N., Hirsch, C. and Svane, A. M. (2020). Testing Goodness of Fit for Point Processes Via Topological Data Analysis. *Electron. J. Stat.* **14** 1024–1074. MR4067816

[5] Blanton, M. R., Bershady, M. A., Abolfathi, B., Albareti, F. D., Allende Prieto, C., Almeida, A., Alonso-García, J., Anders, F., Anderson, S. F., Andrews, B. and et al. (2017). Sloan Digital Sky Survey IV: Mapping the Milky Way, Nearby Galaxies, and the Distant Universe. *Astronomical Journal* **154** 28.

[6] Bobrowski, O. and Mukherjee, S. (2015). The Topology of Probability Distributions on Manifolds. *Probab. Theory Related Fields* **161** 651–686. MR3334278

[7] Boissonnat, J.-D., Chazal, F. and Yvinec, M. (2018). *Geometric and Topological Inference. Cambridge Texts in Applied Mathematics*. Cambridge University Press, Cambridge. MR3837127

[8] Breiman, L., Meisel, W. and Purcell, E. (1977). Variable Kernel Estimates of Multivariate Densities. *Technometrics* **19** 135–144.

[9] Bubenik, P. (2015). Statistical Topological Data Analysis Using Persistence Landscapes. *J. Mach. Learn. Res.* **16** 77–102. MR3317230

[10] Bubenik, P. and Kim, P. T. (2007). A Statistical Approach to Persistent Homology. *Homology Homotopy Appl.* **9** 337–362. MR2366953

[11] Camara, P. G., Rosenbloom, D. I. S., Emmett, K. J., Levine, A. J. and Rabadan, R. (2016). Topological Data Analysis Generates High-Resolution, Genome-wide Maps of Human Recombination. *Cell Systems* **3** 83–94.

[12] Chazal, F. and Divol, V. (2018). The Density of Expected Persistence Diagrams and Its Kernel Based Estimation. In *34th International Symposium on Computational Geometry. LIPIcs. Leibniz Int. Proc. Inform.* **99** Art. No. 26, 15. Schloss Dagstuhl. Leibniz-Zent. Inform., Wadern. MR3824270

[13] Chazal, F., Fasy, B. T., Lecci, F., Rinaldo, A., Singh, A. and Wasserman, L. (2015). On the Bootstrap for Persistence Diagrams and Landscapes. *Modeling and Analysis of Information Systems* **20** 111–120.

[14] Chazal, F., Fasy, B. T., Lecci, F., Rinaldo, A. and Wasserman, L. (2015). Stochastic Convergence of Persistence Landscapes and Silhouettes. *J. Comput. Geom.* **6** 140–161. MR3323391

[15] Chazal, F. and Michel, B. (2017). An Introduction to Topological Data Analysis: Fundamental and Practical Aspects for Data Scientists.

[16] Chen, Y.-C., Wang, D., Rinaldo, A. and Wasserman, L. (2015). Statistical Analysis of Persistence Intensity Functions.

[17] Chung, Y.-M. and Lawson, A. (2019). Persistence Curves: A Canonical Framework for Summarizing Persistence Diagrams.

[18] Crawford, L., Monod, A., Chen, A. X., Mukherjee, S. and Rabadán, R. (2019). Predicting Clinical Outcomes in Glioblastoma: An Application of Topological and Functional Data Analysis. *Journal of the American Statistical Association* 1–12.

[19] Devroye, L., Györfi, L., Lugosi, G. and Walk, H. (2017). On the Measure of Voronoi Cells. *J. Appl. Probab.* **54** 394–408. MR3668473

[20] Devroye, L. P. and Wagner, T. J. (1979). The L_1 Convergence of Kernel Density Estimates. *Ann. Statist.* **7** 1136–1139. MR536515

[21] DeWoskin, D., Climent, J., Cruz-White, I., Vazquez, M., Park, C. and Arsuaga, J. (2010). Applications of Computational Homology to the Analysis of Treatment Response in Breast Cancer Patients. *Topology Appl.* **157** 157–164. MR2556091

[22] Edelsbrunner, Letscher and Zomorodian (2002). Topological Persistence and

Simplification. *Discrete & Computational Geometry* **28** 511–533.
[23] Fasy, B. T., Lecci, F., Rinaldo, A., Wasserman, L., Balakrishnan, S. and Singh, A. (2014). Confidence Sets for Persistence Diagrams. *Ann. Statist.* **42** 2301–2339. MR3269981
[24] Ferdosi, B. J., Buddelmeijer, H., Trager, S. C., Wilkinson, M. H. F. and Roerdink, J. B. T. M. (2011). Comparison of Density Estimation Methods for Astronomical Datasets. *Astronomy & Astrophysics* **531** A114.
[25] Folland, G. B. (1999). *Real Analysis: Modern Techniques and Applications*. Wiley.
[26] Hansen, B. E. (2008). Uniform Convergence Rates for Kernel Estimation with Dependent Data. *Econometric Theory* **24** 726–748. MR2409261
[27] Hiraoka, Y., Shirai, T. and Trinh, K. D. (2018). Limit Theorems for Persistence Diagrams. *Ann. Appl. Probab.* **28** 2740–2780. MR3847972
[28] Kim, J., Shin, J., Rinaldo, A. and Wasserman, L. (2018). Uniform Convergence Rate of the Kernel Density Estimator Adaptive to Intrinsic Volume Dimension.
[29] Kramar, M., Goullet, A., Kondic, L. and Mischaikow, K. (2013). Persistence of Force Networks in Compressed Granular Media. *Physical Review E* **87**.
[30] Kramár, M., Levanger, R., Tithof, J., Suri, B., Xu, M., Paul, M., Schatz, M. F. and Mischaikow, K. (2016). Analysis of Kolmogorov Flow and Rayleigh–bénard Convection Using Persistent Homology. *Physica D: Nonlinear Phenomena* **334** 82–98.
[31] Krebs, J. T. N. and Polonik, W. (2019). On the Asymptotic Normality of Persistent Betti Numbers.
[32] Lachièze-Rey, R., Schulte, M. and Yukich, J. E. (2019). Normal Approximation for Stabilizing Functionals. *The Annals of Applied Probability* **29**.
[33] Lachièze-Rey, R., Peccati, G. and Yang, X. (2020). Quantitative Two-scale Stabilization on the Poisson Space.
[34] Last, G., Peccati, G. and Schulte, M. (2015). Normal Approximation on Poisson Spaces: Mehler's Formula, Second Order Poincaré Inequalities and Stabilization. *Probability Theory and Related Fields* **165** 667–723.
[35] Latała, R. (1997). Estimation of Moments of Sums of Independent Real Random Variables. *Ann. Probab.* **25** 1502–1513. MR1457628
[36] Owada, T. (2018). Limit Theorems for Betti Numbers of Extreme Sample Clouds with Application to Persistence Barcodes. *Ann. Appl. Probab.* **28** 2814–2854. MR3847974
[37] Owada, T. and Adler, R. J. (2017). Limit Theorems for Point Processes under Geometric Constraints (and Topological Crackle). *Ann. Probab.* **45** 2004–2055. MR3650420
[38] Penrose, M. D. and Yukich, J. E. (2001). Central Limit Theorems for Some Graphs in Computational Geometry. *Ann. Appl. Probab.* **11** 1005–1041. MR1878288
[39] Penrose, M. D. and Yukich, J. E. (2003). Weak Laws of Large Numbers in Geometric Probability. *Ann. Appl. Probab.* **13** 277–303. MR1952000
[40] Politis, D. N., Romano, J. P. and Wolf, M. (1999). *Subsampling. Springer Series in Statistics*. Springer-Verlag, New York. MR1707286
[41] Pranav, P., Adler, R. J., Buchert, T., Edelsbrunner, H., Jones, B. J. T., Schwartzman, A., Wagner, H. and van de Weygaert, R. (2019). Unexpected Topology of the Temperature Fluctuations in the Cosmic Microwave Background. *Astronomy & Astrophysics* **627** A163.
[42] Pranav, P., Edelsbrunner, H., van de Weygaert, R., Vegter, G., Kerber, M., Jones, B. J. T. and Wintraecken, M. (2016). The Topology of the Cosmic Web in Terms of Persistent Betti Numbers. *Monthly Notices of the Royal Astronomical*

Society **465** 4281–4310.
[43] Pranav, P., van de Weygaert, R., Vegter, G., Jones, B. J. T., Adler, R. J., Feldbrugge, J., Park, C., Buchert, T. and Kerber, M. (2019). Topology and Geometry of Gaussian Random Fields I: On Betti Numbers, Euler Characteristic, and Minkowski Functionals. *Monthly Notices of the Royal Astronomical Society* **485** 4167–4208.
[44] Roycraft, B. (2021). github.com/btroycraft/stabilizing_statistics_bootstrap.
[45] Roycraft, B., Krebs, J. and Polonik, W. (2021). Supplement to "Bootstrapping Persistent Betti Numbers and Other Stabilizing Statistics".
[46] Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis. Monographs on Statistics and Applied Probability*. Chapman & Hall, London. MR848134
[47] Singh, S. and Póczos, B. (2016). Analysis of k-Nearest Neighbor Distances with Application to Entropy Estimation.
[48] Trinh, K. D. (2019). On Central Limit Theorems in Stochastic Geometry for Add-one Cost Stabilizing Functionals. *Electron. Commun. Probab.* **24** Paper No. 76, 15. MR4049088
[49] Turner, K., Mukherjee, S. and Boyer, D. M. (2014). Persistent Homology Transform for Modeling Shapes and Surfaces. *Inf. Inference* **3** 310–344. MR3311455
[50] Ulmer, M., Ziegelmeier, L. and Topaz, C. M. (2019). A Topological Approach to Selecting Models of Biological Experiments. *PLOS ONE* **14** 1-18.
[51] Wasserman, L. (2018). Topological Data Analysis. *Annu. Rev. Stat. Appl.* **5** 501–535. MR3774757
[52] Xia, K., Feng, X., Tong, Y. and Wei, G. W. (2014). Persistent Homology for the Quantitative Prediction of Fullerene Stability. *Journal of Computational Chemistry* **36** 408–422.
[53] Yogeshwaran, D. and Adler, R. J. (2015). On the Topology of Random Complexes Built Over Stationary Point Processes. *Ann. Appl. Probab.* **25** 3338–3380. MR3404638
[54] Yogeshwaran, D., Subag, E. and Adler, R. J. (2017). Random Geometric Complexes in the Thermodynamic Regime. *Probab. Theory Related Fields* **167** 107–142. MR3602843
[55] Zomorodian, A. and Carlsson, G. (2005). Computing Persistent Homology. *Discrete Comput. Geom.* **33** 249–274. MR2121296

Appendix A: Altered Problem Settings

A.1. B-Bounded Persistent Betti Numbers

To effectively quantify the radius of stabilization for persistent Betti numbers, it is necessary to place controls on the size of possible cycles within a simplicial complex. Large loops extend the influence of a single point beyond the local region, and complicate statistical analysis. As such, we present the following definitions which eliminate any large loops. Note that the statistics of this appendix are presented for their convenient theoretical properties, not their practical significance. Let $S \in \tilde{\mathcal{X}}\left(\mathbb{R}^d\right)$ and $K = K\left(S\right)$ be a simplicial complex with vertices in S. For a given chain of simplices $\sum_{i=1}^m \sigma_i \in C\left(K\left(S\right)\right)$, we have the diameter given by $\operatorname{diam}\left(\sum_{i=1}^m \sigma_i\right) := \operatorname{diam}\left(\bigcup_{i=1}^m \sigma_i\right)$. Let the space of *B-bounded cycles* of the complex K be the vector space, denoted by $Z_{q,B}\left(K\right)$, spanned by cycles in K with diameter no larger than B. We have $Z_{q,B}\left(K\right) := \operatorname{span}\left\{x \in Z_q\left(K\right) \text{ s.t. } \operatorname{diam}\left(x\right) \leq B\right\}$. Likewise let the space of *B-bounded boundaries* be $B_{q,B}\left(K\right) := \operatorname{span}\left\{x \in B_q\left(K\right) \text{ s.t. } \operatorname{diam}\left(x\right) \leq B\right\}$.

The definitions presented here are directly inspired by a previous concept under the name "M-bounded persistence" found in [4], and may be viewed as a generalization thereof. In this previous work, it was shown that a diameter bound of this type is sufficient for establishing functional central limit theorems for persistent Betti numbers, and thus an extension is desireable. The original definition given in [4] is based on a correspondence between loops and connected components in the complement space, and does not apply to arbitrary simplicial complexes and feature dimensions.

The change in naming effected here is not meant to drawn a distinction between the two definitions, but purely to avoid overloading symbols within this paper. M is used in this work to denote an upper bound for a density norm.

The B-bounded spaces obey many of the same properties as their original counterparts. We have $B_{q,M}(K) \subseteq Z_{q,M}(K)$. Thus we can define the *B-bounded homology spaces* as $H_{q,B}(K) = Z_{q,B}(K)/B_{q,B}(K)$. It should be noted that these definitions allow for chains of unbounded diameter, so long as there exists a decomposition into a sum of bounded chains. Furthermore, for $B_{q,B}(K)$, the diameter control is on the chains $x \in B_q(K)$, not on a corresponding $y \in C_{q+1}(K)$ with $x = \partial y$. It is possible to have a chain with arbitrarily high diameter, whose boundary has diameter less than B.

We next define the analog of Betti numbers and persistent Betti numbers over a filtration of simplicial complexes in the bounded context. Given a filtration $\mathcal{K} = \{K^r\}_{r\in\mathbb{R}}$, we have B-bounded analogs for the Betti numbers, persistent homology spaces, and persistent Betti numbers given by

$$\beta^r_{q,B}(\mathcal{K}) := \dim\left(H_{q,B}(K^r)\right) \tag{A.1}$$

$$= \dim\left(Z_{q,B}(K^r)\right) - \dim\left(B_{q,B}(K^r)\right) \tag{A.2}$$

$$H^{r,s}_{q,B}(\mathcal{K}) := \frac{Z_{q,B}(K^r)}{Z_{q,B}(K^r) \cap B_{q,B}(K^s)} \tag{A.3}$$

$$\beta^{r,s}_{q,B}(\mathcal{K}) := \dim\left(H^{r,s}_{q,B}(\mathcal{K})\right) \tag{A.4}$$

$$= \dim\left(Z_{q,B}(K^r)\right) - \dim\left(Z_{q,B}(K^r) \cap B_{q,B}(K^s)\right). \tag{A.5}$$

Unfortunately, no direct analog of the Geometric Lemma 3.1 exists for B-bounded persistent Betti numbers. The addition of a positive simplex can add more than one dimension to $Z_{q,B}$. Consider Z_q consisting of a single cycle with diameter above B but below $2B$, meaning $Z_{q,B} = \{0\}$ initially. Now let the loop be split in two by a new simplex σ. Each piece may now be of diameter less than B, unlike the original. In this way a single simplex can increase the dimension of $Z_{q,B}$ by two or more. The same is true for the negative simplices. Consider the same setup, but now extend each simplex towards a distant point x in a cone. In this case we have $Z_{q,B} = B_{q,B} = \{0\}$ initially. The inclusion of the simplex $\sigma \cup \{x\}$ will split the boundary space just as before into two bounded pieces.

Thus it becomes clear that we must utilize slightly different techniques when considering B-bounded persistence. We have the following inequality, the analog of the Geometric Lemma for B-bounded persistent Betti numbers.

Lemma A.1. *Let $\mathcal{J} = \{J^r\}_{r\in\mathbb{R}}$ and $\mathcal{K} = \{K^r\}_{r\in\mathbb{R}}$ be filtrations of simplicial complexes with $J^r \subseteq K^r$ for all $r \in \mathbb{R}$. Then*

$$\left|\beta_{q,B}^{r,s}(\mathcal{K}) - \beta_{q,B}^{r,s}(\mathcal{J})\right| \leq \max\left\{\dim\left(\frac{Z_{q,B}(K^r)}{Z_{q,B}(J^r)}\right), \dim\left(\frac{B_{q,B}(K^s)}{B_{q,B}(J^s)}\right)\right\} \tag{A.6}$$

$$\leq \dim\left(\frac{Z_{q,B}(K^r)}{Z_{q,B}(J^r)}\right) + \dim\left(\frac{B_{q,B}(K^s)}{B_{q,B}(J^s)}\right). \tag{A.7}$$

Proof.

$$\begin{aligned}
&\left|\beta_{q,B}^{r,s}(\mathcal{K}) - \beta_{q,B}^{r,s}(\mathcal{J})\right| \\
&= \left|\dim\left(\frac{Z_{q,B}(K^r)}{Z_{q,B}(K^r)\cap B_{q,B}(K^s)}\right) - \dim\left(\frac{Z_{q,B}(J^r)}{Z_{q,B}(J^r)\cap B_{q,B}(J^s)}\right)\right| \\
&= \left|\dim\left(\frac{Z_{q,B}(K^r)+B_{q,B}(K^r)}{Z_{q,B}(J^r)+B_{q,B}(K^r)}\right) - \dim\left(\frac{Z_{q,B}(J^r)\cap B_{q,B}(K^s)}{Z_{q,B}(J^r)\cap B_{q,B}(J^s)}\right)\right| \\
&\leq \max\left\{\dim\left(\frac{Z_{q,B}(K^r)+B_{q,B}(K^r)}{Z_{q,B}(J^r)+B_{q,B}(K^r)}\right), \dim\left(\frac{Z_{q,B}(J^r)\cap B_{q,B}(K^s)}{Z_{q,B}(J^r)\cap B_{q,B}(J^s)}\right)\right\} \\
&\leq \max\left\{\dim\left(\frac{Z_{q,B}(K^r)}{Z_{q,B}(J^r)}\right), \dim\left(\frac{B_{q,B}(K^s)}{B_{q,B}(J^s)}\right)\right\} \\
&\leq \dim\left(\frac{Z_{q,B}(K^r)}{Z_{q,B}(J^r)}\right) + \dim\left(\frac{B_{q,B}(K^s)}{B_{q,B}(J^s)}\right).
\end{aligned}$$

□

We make a note here about the difference between Lemma A.1 and the Geometric Lemma 3.1. While drawn from the same fundamental inequality, in the persistent Betti number case, we reduce to counting the simplices that are added when moving from one complex to the other. This reduction cannot be made in the B-bounded case, and we must count the number of additional linearly independent loops and boundaries. Different combinatorial techniques will be needed when applying each lemma, as can be seen in the proofs of Corollaries 4.2, 4.3, and A.6.

A.2. Stabilization Results

We define the q-truncated Euler characteristics as

$$\chi_q(K) := \sum_{k=0}^{q} (-1)^k \#\{K_k\}. \tag{A.8}$$

We have the following stabilization lemmas for B-bounded persistent Betti numbers and q-truncated Euler characteristics. Since in both Lemmas A.2 and A.3 the radius of stabilization is a deterministic constant, (S2) is satisfied for any distribution G. The same is true in the following results for the truncated Euler characteristics.

Lemma A.2. *Let $\mathcal{K}$ satisfy (K1). Then for any $B \geq 0$, $r \in \mathbb{R}$, $s \in \mathbb{R}$, $q \in \mathbb{N}_0$, and $z \in \mathbb{R}^d$, $\rho_z = 2B$ is a locally determined radius of stabilization for $\beta_{q,B}^{r,s}(\mathcal{K})$ centered at z.*

Proof. Let $a \geq 2B$ and $S \in \mathcal{X}\left(\mathbb{R}^d\right)$. We decompose $Z_{q,B}\left(K^r\left(\left(S \cap B_z\left(a\right)\right) \cup \{z\}\right)\right)$ into three spaces. Let U_z be spanned by the generators of $Z_{q,B}\left(K^r\left(\left(S \cap B_z\left(a\right)\right) \cup \{z\}\right)\right)$ with z as a vertex. Let U_a be spanned by the generators with vertices within $B_z\left(a\right) \setminus B_z\left(2B\right)$. Finally, let U_* be spanned by the generators without z as a vertex and with no vertices within $B_z\left(a\right) \setminus B_z\left(2B\right)$. Since the generators of $Z_{q,B}\left(K^r\left(\left(S \cap B_z\left(a\right)\right) \cup \{z\}\right)\right)$ have diameter at most B, there are no generating cycles with vertices both at z and in $B_z\left(a\right) \setminus B_z\left(2B\right)$.

By (K1) we have $Z_{q,B}\left(K^r\left(\left(S \cap B_z\left(a\right)\right) \cup \{z\}\right)\right) = U_z + U_a + U_*$, $Z_{q,B}\left(K^r\left(S \cap B_z\left(a\right)\right)\right) = U_a + U_*$, $Z_{q,B}\left(K^r\left(\left(S \cap B_z\left(2B\right)\right) \cup \{z\}\right)\right) = U_z + U_*$, and $Z_{q,B}\left(K^r\left(S \cap B_z\left(2B\right)\right)\right) = U_*$.

Now for any cycle within U_z, the associated vertex set must lie within $B_z\left(B\right)$. Likewise, for any cycle in U_a, the associated vertex set must lie within $B_z\left(a\right) \setminus B_z\left(B\right)$. These vertex sets cannot intersect, thus $U_z \cap U_a = \{0\}$.

Now, consider any vector spaces X, Y, and Z such that $X \cap Y = \{0\}$. Because $X \cap Y \cap Z$ is a subspace of $X \cap Y$, it is also the trivial space $\{0\}$. We have

$$\begin{aligned} & \dim\left(X + Y + Z\right) - \dim\left(Y + Z\right) && \text{(A.9)} \\ = \ & \dim\left(X\right) - \dim\left(X \cap Y\right) - \dim\left(X \cap Z\right) + \dim\left(X \cap Y \cap Z\right) \\ = \ & \dim\left(X\right) - \dim\left(X \cap Z\right) \\ = \ & \dim\left(X + Z\right) - \dim\left(Z\right). && \text{(A.10)} \end{aligned}$$

We use this result in each of the following. We have

$$\begin{aligned} & \dim\left(Z_{q,B}\left(K^r\left(\left(S \cap B_z\left(a\right)\right) \cup \{z\}\right)\right)\right) - \dim\left(Z_{q,B}\left(K^r\left(S \cap B_z\left(a\right)\right)\right)\right) && \text{(A.11)} \\ = \ & \dim\left(U_z + U_a + U_*\right) - \dim\left(U_a + U_*\right) \\ = \ & \dim\left(U_z + U_*\right) - \dim\left(U_*\right) \\ = \ & \dim\left(Z_{q,B}\left(K^r\left(\left(S \cap B_z\left(2B\right)\right) \cup \{z\}\right)\right)\right) - \dim\left(Z_{q,B}\left(K^r\left(S \cap B_0\left(2B\right)\right)\right)\right). && \text{(A.12)} \end{aligned}$$

A similar result holds for the boundaries. Let V_z, V_a, and V_* be defined similarly to U_z, U_a, and U_*, respectively, instead using the generators of $B_{q,B}\left(K^s\left(\left(S \cap B_z\left(a\right)\right) \cup \{z\}\right)\right)$. Similarly $B_{q,B}\left(K^s\left(\left(S \cap B_z\left(a\right)\right) \cup \{z\}\right)\right) = V_z + V_a + V_*$, $B_{q,B}\left(K^s\left(\left(S \cap B_z\left(2B\right)\right) \cup \{z\}\right)\right) = V_z + V_*$, and we conclude $V_z \cap V_a = \{0\}$. Furthermore, we have $U_z \cap V_a = V_z \cap U_a = \{0\}$ by similar vertex-based arguments. Then $\left(U_z + V_z\right) \cap \left(U_a + V_a\right) = \{0\}$. We have

$$\begin{aligned} & \dim\left(Z_{q,B}\left(K^r\left(\left(S \cap B_z\left(a\right)\right) \cup \{z\}\right)\right) \cap B_{q,B}\left(K^s\left(\left(S \cap B_z\left(a\right)\right) \cup \{z\}\right)\right)\right) \\ & - \dim\left(Z_{q,B}\left(K^r\left(S \cap B_z\left(a\right)\right)\right) \cap B_{q,B}\left(K^s\left(S \cap B_z\left(a\right)\right)\right)\right) \\ \\ = \ & \dim\left(\left(U_z + U_a + U_*\right) \cap \left(V_z + V_a + V_*\right)\right) \\ & - \dim\left(\left(U_a + U_*\right) \cap \left(V_a + V_*\right)\right) \\ = \ & \dim\left(U_z + U_a + U_*\right) + \dim\left(V_z + V_a + V_*\right) - \dim\left(U_z + U_a + U_* + V_z + V_a + V_*\right) \\ & - \dim\left(U_a + U_*\right) - \dim\left(V_a + V_*\right) + \dim\left(U_a + U_* + V_a + V_*\right) \\ = \ & \dim\left(U_z + U_*\right) - \dim\left(U_*\right) + \dim\left(V_z + V_*\right) - \dim\left(V_*\right) \\ & - \dim\left(U_z + U_* + V_z + V_*\right) + \dim\left(U_* + V_*\right) \\ = \ & \dim\left(\left(U_z + U_*\right) \cap \left(V_z + V_*\right)\right) - \dim\left(U_* \cap V_*\right) \\ = \ & \dim\left(Z_{q,B}\left(K^r\left(\left(S \cap B_z\left(2B\right)\right) \cup \{z\}\right)\right) \cap B_{q,B}\left(K^s\left(\left(S \cap B_z\left(2B\right)\right) \cup \{z\}\right)\right)\right) \\ & - \dim\left(Z_{q,B}\left(K^r\left(S \cap B_z\left(2B\right)\right)\right) \cap B_{q,B}\left(K^s\left(S \cap B_z\left(2B\right)\right)\right)\right). \end{aligned}$$

Combining these pieces, the B-bounded persistent Betti numbers must stabilize after a constant radius of $\rho_z = 2B$. □

Lemma A.3. *Let $\mathcal{K}$ satisfy (D2). Then for any $B \geq 0$, $r \in \mathbb{R}$, $s \in \mathbb{R}$, $q \geq 0$, and $z \in \mathbb{R}^d$, $\rho_z = 2\max\{\phi(r), \phi(s)\} + 2B$ is a locally determined radius of stabilization for $\beta^{r,s}_{q,B}(\mathcal{K})$ centered at z.*

Proof. Denote by $\phi := \max(\phi(r), \phi(s))$. Let $S \in \mathcal{X}(\mathbb{R}^d)$. Furthermore, let T be any finite multiset of points in $\mathbb{R}^d$ with $S \cap B_z(2\phi + 2B) \subseteq T$ and $y \notin B_z(2\phi + 2B)$. We have the following partition. Let U_z, U_y, and U_*, respectively, be spanned by the generators of $Z_{q,B}(K^r(T))$ having: a simplex within $B_z(\phi)$, a simplex within $B_y(\phi)$, or neither. Let U^*_z be spanned by the generators of $Z_{q,B}(K^r(T \cup \{z\}))$ have a simplex in $B_z(\phi)$. Finally, let U^*_y be spanned by the generators of $Z_{q,B}(K^r(T \cup \{y\}))$ have a simplex in $B_y(\phi)$.

By (D2) we have $Z_{q,B}(K^r(T)) = U_z + U_* + U_y$, $Z_{q,B}(K^r(T \cup \{z\})) = U^*_z + U_* + U_y$, $Z_{q,B}(K^r(T \cup \{y\})) = U_z + U_* + U^*_y$, and $Z_{q,B}(K^r(T \cup \{z, y\})) = U^*_z + U_* + U^*_y$.

Now for any cycle within U_z, the associated vertex set must lie within $B_z(\phi + B)$. Likewise, for any cycle in U_y, the associated vertex set must lie within $B_y(\phi + B)$. Because $\|y - z\| > 2\phi + 2B$, these vertex sets cannot intersect, thus $U_z \cap U_y = \{0\}$. Likewise $U_z \cap U^*_y = U^*_z \cap U_y = U^*_z \cap U^*_y = \{0\}$.

Now, for any vector spaces X, X^*, Y, and Z such that $X \cap Y^* = X^* \cap Y^* = \{0\}$, we have

$$\begin{aligned}
&\dim(X^* + Y^* + Z) - \dim(X + Y^* + Z)\\
=\ &\dim(X) - \dim(X^*) + \dim(X \cap Y^*) + \dim(X \cap Z) - \dim(X^* \cap Y^*) - \dim(X^* \cap Z)\\
&+ \dim(X^* \cap Y^* \cap Z) - \dim(X \cap Y^* \cap Z)\\
=\ &\dim(X) - \dim(X^*) + \dim(X \cap Z) - \dim(X^* \cap Z)\\
=\ &\dim(X + Z) - \dim(X^* + Z).
\end{aligned}$$

Thus we have

$$\begin{aligned}
&\dim(Z_{q,B}(K^r(T \cup \{z, y\}))) - \dim(Z_{q,B}(K^r(T \cup \{y\})))\\
=\ &\dim\left(U^*_z + U_* + U^*_y\right) - \dim\left(U_z + U_* + U^*_y\right)\\
=\ &\dim(U^*_z + U_*) - \dim(U_z + U_*)\\
=\ &\dim(U^*_z + U_* + U_y) - \dim(U_z + U_* + U_y)\\
=\ &\dim(Z_{q,B}(K^r(T \cup \{z\}))) - \dim(Z_{q,B}(K^r(T))).
\end{aligned}$$

A similar result holds for the boundaries. Let V_z, V^*_z, V_y, V^*_y, and V_* be defined similarly to U_z, U^*_z, U_y, U^*_y, and U_*, respectively, instead using the generators of $B_{q,B}(K^r(T))$, $B_{q,B}(K^r(T \cup \{y\}))$, and $B_{q,B}(K^r(T \cup \{y\}))$. Similarly $B_{q,B}(K^s(T)) = V_z + V_* + V_y$, $B_{q,B}(K^s(T \cup \{z\})) = V^*_z + V_* + V_y$, $B_{q,B}(K^s(T \cup \{y\})) = V_z + V_* + V^*_y$, and $B_{q,B}(K^s(T \cup \{z, y\})) = V^*_z + V_* + V^*_y$. We conclude $V_z \cap V_y = V_z \cap V^*_y = V^*_z \cap V_y = V^*_z \cap V^*_y = \{0\}$. Furthermore, we have $U_z \cap V_y = U_z \cap V^*_y = U^*_z \cap V_y = U^*_z \cap V^*_y = \{0\}$ and $V_z \cap U_y = V_z \cap U^*_y = V^*_z \cap U_y = V^*_z \cap U^*_y = \{0\}$ by similar vertex-based arguments. Thus $(U_z + V_z) \cap (U_y + V_y) = (U_z + V_z) \cap \left(U^*_y + V^*_y\right) = (U^*_z + V^*_z) \cap (U_y + V_y) =$

$(U_z^* + V_z^*) \cap \left(U_y^* + V_y^*\right) = \{0\}$. We have

$$\begin{aligned}
& \dim\left(Z_{q,B}\left(K^r\left(T \cup \{z,y\}\right)\right) \cap B_{q,B}\left(K^s\left(T \cup \{z,y\}\right)\right)\right) \\
& - \dim\left(Z_{q,B}\left(K^r\left(T \cup \{y\}\right)\right) \cap B_{q,B}\left(K^s\left(T \cup \{y\}\right)\right)\right) \\
= \; & \dim\left(\left(U_z^* + U_* + U_y^*\right) \cap \left(V_z^* + V_* + V_y^*\right)\right) - \dim\left(\left(U_z + U_* + U_y^*\right) \cap \left(V_z + V_* + V_y^*\right)\right) \\
= \; & \dim\left(U_z^* + U_* + U_y^*\right) + \dim\left(V_z^* + V_* + V_y^*\right) - \dim\left(U_z^* + U_* + U_y^* + V_z^* + V_* + V_y^*\right) \\
& - \dim\left(U_z + U_* + U_y^*\right) - \dim\left(V_z + V_* + V_y^*\right) + \dim\left(U_z + U_* + U_y^* + V_z + V_* + V_y^*\right) \\
= \; & \dim\left(U_z^* + U_* + U_y\right) - \dim\left(U_z + U_* + U_y\right) + \dim\left(V_z^* + V_* + V_y\right) - \dim\left(V_z + V_* + V_y\right) \\
& - \dim\left(U_z^* + U_* + U_y + V_z^* + V_* + V_y\right) + \dim\left(U_z + U_* + U_y + V_z + V_* + V_y\right) \\
= \; & \dim\left(\left(U_z^* + U_* + U_y\right) \cap \left(V_z^* + V_* + V_y\right)\right) - \dim\left(\left(U_z + U_* + U_y\right) \cap \left(V_z + V_* + V_y\right)\right) \\
= \; & \dim\left(Z_{q,B}\left(K^r\left(T \cup \{z\}\right)\right) \cap B_{q,B}\left(K^s\left(T \cup \{z\}\right)\right)\right) \\
& - \dim\left(Z_{q,B}\left(K^r\left(T\right)\right) \cap B_{q,B}\left(K^s\left(T\right)\right)\right).
\end{aligned}$$

Thus, the addition of y to T does not change the add-z cost. We proceed inductively. Starting with $S \cap B_z\left(2\phi + 2B\right)$, for any $a > 2\phi + 2B$, the finitely many points of $\left(S \cap B_z\left(a\right)\right) \setminus \left(S \cap B_z\left(2\phi + 2B\right)\right)$ may be added one at a time, while leaving the add-z cost unchanged. Thus, the B-bounded persistent Betti numbers must stabilize after a constant radius of $\rho_z = 2\phi + 2B$. □

Lemma A.4. *Let K satisfy (K1) and (D1). Then for any $z \in \mathbb{R}^d$ and $q \geq 0$, $\rho_z = \phi$ is a locally determined radius of stabilization for $\chi_q\left(K\right)$ centered at z.*

Proof. Let $a \geq \phi$. By (K1) and (D1), we can partition $K\left(\left(S \cap B_z\left(a\right)\right) \cup \{z\}\right)$ into the sets

$$U := \left\{\sigma \in K\left(\left(S \cap B_z\left(a\right)\right) \cup \{z\}\right) \text{ s.t. } z \in \sigma\right\} \tag{A.13}$$

$$V := \left\{\sigma \in K\left(\left(S \cap B_z\left(a\right)\right) \cup \{z\}\right) \text{ s.t. } \sigma \subset B_z\left(\phi\right) \setminus \{z\}\right\} \tag{A.14}$$

$$W := \left\{\sigma \in K\left(\left(S \cap B_z\left(a\right)\right) \cup \{z\}\right) \text{ s.t. } \sigma \cap B_z\left(a\right) \setminus B_z\left(\phi\right) \neq \emptyset\right\} \tag{A.15}$$

Condition (D1) gives that no simplices may simultaneously have z as a vertex and intersect $B_z\left(a\right) \setminus B_z\left(\phi\right)$, thus U, V, and W indeed partition $K\left(\left(S \cap B_z\left(a\right)\right) \cup \{z\}\right)$. Condition (K1) gives that the addition of $\{z\}$ and $S \cap \left(B_z\left(a\right) \setminus B_z\left(\phi\right)\right)$ to $S \cap B_z\left(\phi\right)$ may only introduce simplices to $K\left(S \cap B_z\left(\phi\right)\right)$ with vertices somewhere within $S \cap \left(B_z\left(a\right) \setminus B_z\left(\phi\right)\right) \cup \{z\}$, and thus not included in V. Therefore, $V \subseteq K\left(S \cap B_z\left(\phi\right)\right)$. Furthermore, since $S \cap B_z\left(\phi\right) \subset \left(S \cap B_z\left(a\right)\right) \cup \{z\}$, $K\left(S \cap B_z\left(\phi\right)\right) \subseteq V$. Thus we have $V = K\left(S \cap B_z\left(\phi\right)\right)$. Using similar arguments, condition (K1) also gives $K\left(\left(S \cap B_z\left(\phi\right)\right) \cup \{z\}\right) = U \cup V$ and $K\left(S \cap B_z\left(a\right)\right) = V \cup W$.

For U_k, V_k, and W_k denoting the set of k-simplices contained in U, V, and W, respectively,

the add-z cost for the q-truncated Euler characteristic becomes

$$\chi_q\left(K\left(\left(S\cap B_z\left(a\right)\right)\cup\{z\}\right)\right)-\chi_q\left(K\left(S\cap B_z\left(a\right)\right)\right) \tag{A.16}$$

$$=\sum_{k=0}^{q}\left(-1\right)^k\#\left\{K_k\left(\left(S\cap B_z\left(a\right)\right)\cup\{z\}\right)\right\}-\sum_{k=0}^{q}\left(-1\right)^k\#\left\{K_k\left(S\cap B_z\left(a\right)\right)\right\} \tag{A.17}$$

$$=\sum_{k=0}^{q}\left(-1\right)^k\left(\#\left\{U_k\right\}+\#\left\{V_k\right\}+\#\left\{W_k\right\}\right)-\sum_{k=0}^{q}\left(-1\right)^k\left(\#\left\{V_k\right\}+\#\left\{W_k\right\}\right) \tag{A.18}$$

$$=\sum_{k=0}^{q}\left(-1\right)^k\left(\#\left\{U_k\right\}+\#\left\{V_k\right\}\right)-\sum_{k=0}^{q}\left(-1\right)^k\#\left\{V_k\right\} \tag{A.19}$$

$$=\sum_{k=0}^{q}\left(-1\right)^k\#\left\{K_k\left(\left(S\cap B_z\left(\phi\right)\right)\cup\{z\}\right)\right\}-\sum_{k=0}^{q}\left(-1\right)^k\#\left\{K_k\left(S\cap B_z\left(\phi\right)\right)\right\} \tag{A.20}$$

$$=\chi_q\left(K\left(\left(S\cap B_z\left(\phi\right)\right)\cup\{z\}\right)\right)-\chi_q\left(K\left(S\cap B_z\left(\phi\right)\right)\right). \tag{A.21}$$

We see that $\chi_q\left(K\right)$ stabilizes after a constant radius of $\rho_z=\phi$, thus the local-determination criterion is immediately satisfied. □

Lemma A.5. *Let K satisfy (D2). Then for any $z\in\mathbb{R}^d$ and $q\geq 0$, $\rho_z=2\phi$ is a locally determined radius of stabilization for $\chi_q\left(K\right)$ centered at z.*

Proof. Let $z\in\mathbb{R}^d$ and $S\in\mathcal{X}\left(\mathbb{R}^d\right)$. Furthermore, let T be a finite multiset of points in $\mathbb{R}^d$ such that $S\cap B_z\left(2\phi\right)\subseteq T$. Let $y\notin B_z\left(2\phi\right)$. Consider the partition

$$U:=\left\{\sigma\in K\left(T\right)\text{ s.t. }\sigma\subset B_z\left(\phi\right)\right\} \tag{A.22}$$

$$U^*:=\left\{\sigma\in K\left(T\cup\{z\}\right)\text{ s.t. }\sigma\subset B_z\left(\phi\right)\right\} \tag{A.23}$$

$$V:=\left\{\sigma\in K\left(T\right)\text{ s.t. }\sigma\subset B_y\left(\phi\right)\right\} \tag{A.24}$$

$$V^*:=\left\{\sigma\in K\left(T\cup\{y\}\right)\text{ s.t. }\sigma\subset B_y\left(\phi\right)\right\} \tag{A.25}$$

$$W:=\left\{\sigma\in K\left(T\right)\text{ s.t. }\sigma\nsubseteq B_z\left(\phi\right)\text{ and }\sigma\nsubseteq B_y\left(\phi\right)\right\}. \tag{A.26}$$

Condition (D2) limits the influence of a single additional point on the complex to the ball of radius ϕ around it. $B_z\left(\phi\right)\cap B_z\left(\phi\right)=\emptyset$ because $\left\|y-z\right\|>2\phi$. Thus we have $K\left(T\right)=U\cup W\cup V$, $K\left(T\cup\{z\}\right)=U^*\cup W\cup V$, $K\left(T\cup\{y\}\right)=U\cup W\cup V^*$, and $K\left(T\cup\{y,z\}\right)=U^*\cup W\cup V^*$.

For U_k, U_k^*, V_k, V_k^*, and W_k denoting the set of k-simplices contained in U, U^*, V, V^*,

and W, respectively, the add-z cost for the q-truncated Euler characteristic becomes

$$\begin{aligned}
&\chi_q\left(K\left(T\cup\{y,z\}\right)\right)-\chi_q\left(K\left(T\cup\{y\}\right)\right)\\
&=\sum_{k=0}^{q}(-1)^k\#\left\{K_k\left(T\cup\{y,z\}\right)\right\}-\sum_{k=0}^{q}(-1)^k\#\left\{K_k\left(T\cup\{y\}\right)\right\}\\
&=\sum_{k=0}^{q}(-1)^k\left(\#\left\{U_k^*\right\}+\#\left\{W_k\right\}+\#\left\{V_k^*\right\}\right)-\sum_{k=0}^{q}(-1)^k\left(\#\left\{V_k^*\right\}+\#\left\{W_k\right\}\right)\\
&=\sum_{k=0}^{q}(-1)^k\left(\#\left\{U_k^*\right\}+\#\left\{W_k\right\}+\#\left\{V_k\right\}\right)-\sum_{k=0}^{q}(-1)^k\left(\#\left\{V_k\right\}+\#\left\{W_k\right\}\right)\\
&=\sum_{k=0}^{q}(-1)^k\#\left\{K_k\left(T\cup\{z\}\right)\right\}-\sum_{k=0}^{q}(-1)^k\#\left\{K_k\left(T\right)\right\}\\
&=\chi_q\left(K\left(T\cup\{z\}\right)\right)-\chi_q\left(K\left(T\right)\right).
\end{aligned}$$

We see that the addition of $\{y\}$ does not change the add-z cost. Starting with $T=S\cap B_z(\phi)$, for any radius $a>2\phi$, $S\cap\left(B_z(a)\setminus B_z(\phi)\right)$ consists of finitely many points, which may be added to $S\cap B_z(\phi)$ in succession while leaving the add-z cost unchanged. We conclude that $\rho_z=2\phi$ is a radius of stabilization for $\chi_q(K)$, and is locally-determined by virtue of being constant. □

A.3. Bootstrap Results

Here we give bootstrap convergence results for the altered statistics defined in Appendix A.1. For given vectors of birth and death times, $\vec{r}=(r_i)_{i=1}^k$ and $\vec{s}=(s_i)_{i=1}^k$, let $\beta_{q,B}^{\vec{r},\vec{s}}=\left(\beta_{q,B}^{r_i,s_i}\right)_{i=1}^k$ denote the multivariate function whose components are the B-bounded persistent Betti numbers evaluated at each pair of birth and death times. Likewise, for a vector of filtration times $\vec{r}=(r_i)_{i=1}^k$, let $\chi_q^{\vec{r}}$ denote the multivariate function giving the q-truncated Euler characteristic at each time r_i, with $\chi_q^{\vec{r}}(\mathcal{K}):=\left(\chi_q\left(K^{r_i}\right)\right)_{i=1}^k$.

The following apply for $F\in\mathcal{P}\left(\mathbb{R}^d\right)$ with density f such that $\|f\|_p<\infty$ for some $p>2$, as specified. F and $\hat{F}_n$ are such that $\hat{F}_n$ has density $\hat{f}_n$, $\|\hat{f}_n-f\|_1\to 0$, and $\|\hat{f}_n-f\|_p\to 0$ in probability (resp. *a.s.*). Let $\mathbf{X}_n=\{X_i\}_{i=1}^n\overset{\text{iid}}{\sim}F$ and $(m_n)_{n\in\mathbb{N}}$ such that $\lim_{n\to\infty}m_n=\infty$. $\mathbf{X}_{m_n}^*=\{X_i^*\}_{i=1}^{m_n}\overset{\text{iid}}{\sim}\hat{F}_n\big|\mathbf{X}_n$ is a bootstrap sample and G a multivariate distribution. Recalling the conclusion of Theorem 2.7, for a multivariate statistic $\vec{\psi}$:

Statement A.1.

$$\frac{1}{\sqrt{n}}\left(\vec{\psi}\left(\sqrt[d]{n}\mathbf{X}_n\right)-\mathbb{E}\left[\vec{\psi}\left(\sqrt[d]{n}\mathbf{X}_n\right)\right]\right)\xrightarrow{d}G$$

if and only if

$$\frac{1}{\sqrt{m_n}}\left(\vec{\psi}\left(\sqrt[d]{m_n}\mathbf{X}_{m_n}^*\right)-\mathbb{E}\left[\vec{\psi}\left(\sqrt[d]{m_n}\mathbf{X}_{m_n}^*\right)\big|\mathbf{X}_n\right]\right)\xrightarrow{d}G \text{ in probability (resp. a.s.).}$$

Corollary A.6. *Let $q\geq 0$ and $p>2q+3$. Let $\mathcal{K}$ be a filtration of simplicial complexes satisfying (K1). Then for any given $\vec{r}$, $\vec{s}$, and $B>0$, Statement A.1 holds for $\beta_{q,B}^{\vec{r},\vec{s}}$.*

Note the difference in necessary conditions between Corollaries 4.2 and A.6. Corollary 4.2 notably requires a translation-invariant simplicial complex, along with the elimination of small loops via (D3). Corollary A.6 imposes relatively few assumptions on the underlying simplicial complex. As a general statement, it can be seen that the B-bounded persistent Betti numbers defined here are better behaved than the unbounded persistent Betti numbers. Furthermore, the B-bounded persistent Betti numbers allow for an explicit rate calculation for the 2-Wasserstein metric in Proposition 2.6, see Appendix B.6 for details. For the unbounded persistent Betti numbers, this rate is stated implicitly in terms of the unknown tail probability for the radius of stabilization.

Proof. Let $\mathbf{Y}_n = \{Y_i\}_{i=1}^n$ be iid and Y' an independent copy. For a given $q \geq 0$, $B \geq 0$, and $r, s \in \mathbb{R}$, we will show that $B_{q,B}^{r,s}(\mathcal{K})$ satisfies assumption (E2).

Applying Lemma A.1, we must bound above the number of linearly independent B-bounded q-cycles and q-boundaries added when $\{\sqrt[d]{n}Y'\}$ is included with the sample $\sqrt[d]{n}\mathbf{Y}_n$. We start by considering the cycles. By (K1) the addition of $\sqrt[d]{n}Y'$ will only introduce simplices to the complex having $\sqrt[d]{n}Y'$ as a vertex. As such, any B-bounded cycles in $Z_q\left(K^r\left(\sqrt[d]{n}\left(\mathbf{Y}_n \cup \{Y'\}\right)\right)\right)$ not having $\sqrt[d]{n}Y'$ as a vertex must already be in $Z_q\left(K^r\left(\sqrt[d]{n}\mathbf{Y}_n\right)\right)$, and thus in $Z_{q,B}\left(K^r\left(\sqrt[d]{n}\mathbf{Y}_n\right)\right)$. Thus, we must only bound the possible number of linearly independent B-bounded cycles within $K^r\left(\sqrt[d]{n}\left(\cup\{Y'\}\right)\right)$ which have $\sqrt[d]{n}Y'$ as a vertex.

Let $I_n := \sum_{i=1}^n \mathbb{1}\left\{\|Y_i - Y'\| \leq B/\sqrt[d]{n}\right\}$ be the number of sample points falling within B of $\sqrt[d]{n}$.

We will construct a worst-case scenario. For any simplicial complexes $J \subseteq K$, we have that $Z_{q,B}(J) \subseteq Z_{q,B}(K)$. The addition of more simplices to $K^r\left(\sqrt[d]{n}\left(\mathbf{Y}_n \cup \{Y'\}\right)\right)$ having $\sqrt[d]{n}Y'$ as a vertex may increase the dimension of $Z_{q,B}\left(K^r\left(\sqrt[d]{n}\left(\mathbf{Y}_n \cup \{Y'\}\right)\right)\right)$, but will not alter $Z_{q,B}\left(K^r\left(\sqrt[d]{n}\mathbf{Y}_n\right)\right)$. As a worst case, we assume K^r is such that all possible simplices containing $\sqrt[d]{n}Y'$ are included. Thus, for any simplex $\sigma \in K^r\left(\sqrt[d]{n}\mathbf{Y}_n\right)$ such that $\sigma \subseteq \left(\sqrt[d]{n}\mathbf{Y}_n\right) \cap B_{\sqrt[d]{n}Y'}(B)$ and $\operatorname{diam}(\sigma) \leq B$, $\partial\left(\sigma \cup \left\{\sqrt[d]{n}Y'\right\}\right)$ has diameter at most B, contains $\sqrt[d]{n}Y'$ as a vertex, and is a cycle within $Z_r\left(K^r\left(\sqrt[d]{n}\left(\mathbf{Y}_n \cup \{Y'\}\right)\right)\right)$. Let

$$U := \left\{\partial\left(\sigma \cup \left\{\sqrt[d]{n}Y'\right\}\right) \text{ s.t. } \sigma \subseteq \left(\sqrt[d]{n}\mathbf{Y}_n\right) \cap B_{\sqrt[d]{n}Y'}(B) \text{ and } \#\{\sigma\} = q+1\right\}.$$

Now consider x to be any cycle in $Z_q\left(K^r\left(\sqrt[d]{n}\left(\mathbf{Y}_n \cup \{Y'\}\right)\right)\right)$ with diameter at most B and a vertex at $\sqrt[d]{n}Y'$. For every simplex σ of x not containing $\sqrt[d]{n}Y'$ as a vertex, we add $\partial\left(\sigma \cup \left\{\sqrt[d]{n}Y'\right\}\right)$ to x, $\partial\left(\sigma \cup \left\{\sqrt[d]{n}Y'\right\}\right)$ necessarily having diameter less than B. This operation cannot add any new vertices to x, and thus cannot increase the total cycle diameter. What remains after completing these additions is either 0 or a cycle x' whose simplices all contain $\sqrt[d]{n}Y'$ as a vertex, the latter being an impossibility. Thus, any B-bounded cycle in $Z_q\left(K^r\left(\sqrt[d]{n}\left(\mathbf{Y}_n \cup \{Y'\}\right)\right)\right)$ having a vertex at $\sqrt[d]{n}Y'$ can be written as a linear combination of B-bounded elements from U. For $I_n := \sum_{i=1}^n \mathbb{1}\left\{\|Y_i - Y'\| \leq B/\sqrt[d]{n}\right\}$, we arrive at a worst case bound of

$$\dim\left(\frac{Z_{q,B}\left(K^r\left(\sqrt[d]{n}\left(\mathbf{Y}_n \cup \{Y'\}\right)\right)\right)}{Z_{q,B}\left(K^r\left(\sqrt[d]{n}\mathbf{Y}_n\right)\right)}\right) \leq \#\{U\} = \binom{I_n}{q+1}. \tag{A.27}$$

A similar argument for the boundaries yields

$$\dim\left(\frac{B_{q,B}\left(K^s\left(\sqrt[d]{n}\left(\mathbf{Y}_n \cup \{Y'\}\right)\right)\right)}{B_{q,B}\left(K^s\left(\sqrt[d]{n}\mathbf{Y}_n\right)\right)}\right) \leq \#\{U\} = \binom{I_n}{q+1}. \tag{A.28}$$

For any $a > 2$, via Lemma A.1 we have

$$\begin{aligned}
&\left|\beta_{q,B}^{r,s}\left(\mathcal{K}\left(\sqrt[d]{n}\left(\mathbf{Y}_n \cup \{Y'\}\right)\right)\right) - \beta_{q,B}^{r,s}\left(\mathcal{K}\left(\sqrt[d]{n}\mathbf{Y}_n\right)\right)\right|^a \\
&\leq \max\left\{\dim\left(\frac{Z_{q,B}\left(K^r\left(\sqrt[d]{n}\left(\mathbf{Y}_n \cup \{Y'\}\right)\right)\right)}{Z_{q,B}\left(K^r\left(\sqrt[d]{n}\mathbf{Y}_n\right)\right)}\right), \dim\left(\frac{B_{q,B}\left(K^s\left(\sqrt[d]{n}\left(\mathbf{Y}_n \cup \{Y'\}\right)\right)\right)}{B_{q,B}\left(K^s\left(\sqrt[d]{n}\mathbf{Y}_n\right)\right)}\right)\right\}^a \\
&\leq \binom{I_n}{q+1}^a \\
&\leq \frac{1}{\left((q+1)!\right)^a} I_n^{a(q+1)} \\
&\leq \frac{1}{\left((q+1)!\right)^a}\left(I_n^{a(q+1)} + 1\right).
\end{aligned}$$

Here $R = B$, $U_a = 1/\left((q+1)!\right)^a$, and $u_a = a\,(q+1)$. (E1) is then satisfied via Lemma 2.3. (S2) is satisfied via Lemma A.2, in this case with a constant radius of stabilization of $2B$. An application of Theorem 2.7 gives the desired result.

In this case, given that the radius of stabilization is a known constant, an explicit rate for γ_ϵ in Proposition 2.6 can be calculated. Details omitted, from the proof of Proposition 2.6 we have $\delta_\epsilon = B^d \epsilon^{\frac{p-2}{p-1}}$ up to constant factors. For $p < \infty$, using $a = (p-1)/(q+1)$ we achieve an optimal rate for γ_ϵ of

$$O\left(B^{d\left(1-\frac{2q+2}{p-1}\right)}\left(1 + B^{d(2q+2)}\right)\epsilon^{\frac{p-2}{p-1}\left(1-\frac{2q+2}{p-1}\right)}\right). \tag{A.29}$$

For $p = \infty$, using $a_\epsilon = 2 - \log(\delta_\epsilon)$ we achieve an optimal rate of

$$O\left(\epsilon B^{d(2q+3)}\left(\frac{-\log\left(B^d\epsilon\right)}{\log\left(-\log\left(B^d\epsilon\right)\right)}\right)^{2q+2}\right). \tag{A.30}$$

□

Corollary A.7. *Let $q \geq 0$ and $p > 2q+1$. Let $\mathcal{K}$ be a filtration of simplicial complexes satisfying (K1) and (D1). Then for any given $\vec{r}$, Statement A.1 holds for $\chi_q^{\vec{r}}$.*

Proof. We will show that assumption (E2) is satisfied for $\psi = \chi_q^r(\mathcal{K}) := \chi_q(K^r)$ given $r \in \mathbb{R}$. Let $\mathbf{Y}_n = \{Y_i\}_{i=1}^n$ be an iid sample in $\mathbb{R}^d$, with Y' an independent copy.

By (K1), and (D1) it suffices to count those simplices within $B_{\sqrt[d]{n}Y'}(\phi(r))$ having a

vertex at $\sqrt[d]{n}Y'$. Let $I_n = \sum_{i=1}^{n} \mathbb{1}\left\{\|Y_i - Y'\| \leq \phi(r)/\sqrt[d]{n}\right\}$. For any $a > 2$, we have

$$\begin{aligned}
&\left|\chi_q^r\left(\mathcal{K}\left(\sqrt[d]{n}\left(\mathbf{Y}_n \cup \{Y'\}\right)\right)\right) - \chi_q^r\left(\mathcal{K}\left(\sqrt[d]{n}\mathbf{Y}_n\right)\right)\right|^a \\
&= \left|\sum_{k=0}^{q}(-1)^k \#\left\{K_k^r\left(\sqrt[d]{n}\left(\mathbf{Y}_n \cup \{Y'\}\right)\right)\right\} - \sum_{k=0}^{q}(-1)^k \#\left\{K_k^r\left(\sqrt[d]{n}\mathbf{Y}_n\right)\right\}\right|^a \\
&= \left|\sum_{k=0}^{q}(-1)^k \left(\#\left\{K_k^r\left(\sqrt[d]{n}\left(\mathbf{Y}_n \cup \{Y'\}\right)\right)\right\} - \#\left\{K_k^r\left(\sqrt[d]{n}\mathbf{Y}_n\right)\right\}\right)\right|^a \\
&= \left|\sum_{k=0}^{q}(-1)^k \#\left\{K_k^r\left(\sqrt[d]{n}\left(\mathbf{Y}_n \cup \{Y'\}\right)\right) \setminus K_k^r\left(\sqrt[d]{n}\mathbf{Y}_n\right)\right\}\right|^a \\
&\leq \left(\sum_{k=0}^{q}\binom{I_n}{k}\right)^a \\
&\leq \left(\sum_{k=0}^{q}\frac{I_n^k}{k!}\right)^a \\
&\leq \left(\sum_{k=0}^{q}\frac{I_n^q}{k!}\right)^a \\
&\leq \left(eI_n^q\right)^a \\
&\leq e^a\left(1 + I_n^{aq}\right).
\end{aligned}$$

Here $R = \phi(r)$, $U_a = e^a$, and $u_a = aq$, satisfying (E2). (E1) then follows from Lemma 2.3 for $p \geq qa+1 > 2q+1$. (S2) is satisfied via Lemma A.4 with a constant radius of stabilization $\phi(r)$. (S1) is satisfied via Lemma 2.4. An application of Theorem 2.7 gives the final result.

For the rate in Proposition 2.6, for $p < \infty$ we have that $\delta_\epsilon = \epsilon^{\frac{p-2}{p-1}}$ up to constant factors. Using $a = (p-1)/q$ we achieve a final rate for γ_ϵ of

$$O\left(\epsilon^{\frac{p-2}{p-1}\left(1-\frac{2q}{p-1}\right)}\right). \tag{A.31}$$

For $p = \infty$, using $a = a_\epsilon = 2 - \log(\epsilon)$ we achieve a final rate of

$$O\left(\epsilon\left(\frac{-\log(\epsilon)}{\log(-\log(\epsilon))}\right)^{2q}\right). \tag{A.32}$$

□

Corollary A.8. *Let $q \geq 0$ and $p > 2q+3$. Let $\mathcal{K}$ be a filtration of simplicial complexes satisfying (D2). Then for any given $\vec{r}$, Statement A.1 holds for $\chi_q^{\vec{r}}$.*

Proof. The proof follows exactly that of Corollary A.7. Let $\mathbf{Y}_n = \{Y_i\}_{i=1}^n$ be an iid sample in $\mathbb{R}^d$, with Y' an independent copy.

By (D2) it suffices to consider simplices within $B_{\sqrt[d]{n}Y'}(\phi(r))$. Let

$$I_n = \sum_{i=1}^{n}\mathbb{1}\left\{\|Y_i - Y'\| \leq \phi(r)/\sqrt[d]{n}\right\}.$$

For any $a > 2$, we have

$$\begin{aligned}
&\chi_q^r\left(\mathcal{K}\left(\sqrt[d]{n}\left(\mathbf{Y}_n \cup \{Y'\}\right)\right)\right) - \chi_q^r\left(\mathcal{K}\left(\sqrt[d]{n}\mathbf{Y}_n\right)\right) \\
= &\sum_{k=0}^{q}(-1)^k \#\left\{K_k^r\left(\sqrt[d]{n}\left(\mathbf{Y}_n \cup \{Y'\}\right)\right)\right\} - \sum_{k=0}^{q}(-1)^k \#\left\{K_k^r\left(\sqrt[d]{n}\mathbf{Y}_n\right)\right\} \\
= &\sum_{k=0}^{q}(-1)^k \left(\#\left\{K_k^r\left(\sqrt[d]{n}\left(\mathbf{Y}_n \cup \{Y'\}\right)\right)\right\} - \#\left\{K_k^r\left(\sqrt[d]{n}\mathbf{Y}_n\right)\right\}\right) \\
= &\sum_{k=0}^{q}(-1)^k \left(\#\left\{K_k^r\left(\sqrt[d]{n}\left(\mathbf{Y}_n \cup \{Y'\}\right)\right) \setminus K_k^r\left(\sqrt[d]{n}\mathbf{Y}_n\right)\right\}\right) \\
&- \sum_{k=0}^{q}(-1)^k \left(\#\left\{K_k^r\left(\sqrt[d]{n}\mathbf{Y}_n\right) \setminus K_k^r\left(\sqrt[d]{n}\left(\mathbf{Y}_n \cup \{Y'\}\right)\right)\right\}\right).
\end{aligned}$$

Any simplices added by the inclusion of $\sqrt[d]{n}Y'$ may contain $\sqrt[d]{n}Y'$ as a vertex, and any removed simplices must only have vertices within $\sqrt[d]{n}\mathbf{Y}_n$. We bound the possible simplices in each dimension. Thus for any $a > 2$

$$\begin{aligned}
&\left|\chi_q^r\left(\mathcal{K}\left(\sqrt[d]{n}\left(\mathbf{Y}_n \cup \{Y'\}\right)\right)\right) - \chi_q^r\left(\mathcal{K}\left(\sqrt[d]{n}\mathbf{Y}_n\right)\right)\right|^a \\
\leq &\left(\sum_{k=0}^{q}\binom{I_n}{k+1} + \sum_{k=0}^{q}\binom{I_n+1}{k+1}\right)^a \\
\leq &\left(2\sum_{k=0}^{q}\binom{I_n+1}{k+1}\right)^a \\
\leq &\left(2\sum_{k=0}^{q}\frac{(I_n+1)^{k+1}}{(k+1)!}\right)^a \\
\leq &\left(2\sum_{k=0}^{q}\frac{(I_n+1)^{q+1}}{(k+1)!}\right)^a \\
\leq &\left(2\left(e-1\right)\right)^a \left(I_n+1\right)^{a(q+1)} \\
\leq &\, 2^{a(q+2)-1}\left(e-1\right)^a \left(1 + I_n^{a(q+1)}\right).
\end{aligned}$$

Here $R = \phi(r)$, $U_a \leq 2^{a(q+2)-1}(e-1)^a$, and $u_a = a(q+1)$, satisfying (E2). (E1) is then satisfied via Lemma 2.3 for $p \geq a(q+1)+1 > 2q+3$. (S2) is satisfied via Lemma A.4 with a constant radius of stabilization $\phi(r)$. (S1) is satisfied via Lemma 2.4. An application of Theorem 2.7 gives the final result.

For the rate in Proposition 2.6, for $p < \infty$ we have that $\delta_\epsilon = \epsilon^{\frac{p-2}{p-1}}$ up to constant factors. Using $a = (p-1)/(q+1)$ we achieve a final rate for γ_ϵ of

$$O\left(\epsilon^{\frac{p-2}{p-1}\left(1-\frac{2q+2}{p-1}\right)}\right). \tag{A.33}$$

For $p = \infty$, using $a = a_\epsilon = 2 - \log(\epsilon)$ we achieve a final rate of

$$O\left(\epsilon\left(\frac{-\log(\epsilon)}{\log\left(-\log(\epsilon)\right)}\right)^{2q+2}\right). \tag{A.34}$$

□

Appendix B: Proofs of Main Results

B.1. Necessary Inequalities

Throughout these proofs, we will make ample use of the Hölder, Jensen, and Minkowsky inequalities, along with the following. For brevity, these inequalities may be used implicitly and in combination. For $m \in \mathbb{N}$, $\{x_i\}_{i=1}^m \subset \mathbb{R}$, and $k \geq 1$,

$$\left|\sum_{i=1}^m x_i\right|^k \leq m^{k-1}\left(\sum_{i=1}^m |x_i|^k\right). \tag{B.1}$$

Likewise for $0 \leq k \leq 1$

$$\left|\sum_{i=1}^m x_i\right|^k \leq \sum_{i=1}^m |x_i|^k. \tag{B.2}$$

Next, for any density f and $1 \leq j \leq k \leq \infty$

$$\|f\|_j^j \leq \|f\|_k^{(j-1)\frac{k}{k-1}}. \tag{B.3}$$

Finally, for any set $A \subseteq \mathbb{R}^d$ with $|A|$ the Lebesgue measure of A and $k \geq 1$

$$\left|\int_A f(x)\ \mathrm{d}x\right|^k \leq |A|^{k-1}\int_A |f(x)|^k\ \mathrm{d}x. \tag{B.4}$$

Furthermore, in each of the following, we use the simplified notation

$$H_n(\mathbf{S},\mathbf{T}) = \psi\left(\sqrt[d]{n}\mathbf{S}\right) - \psi\left(\sqrt[d]{n}\mathbf{T}\right) \tag{B.5}$$

for the change in the statistic ψ when the underlying scaled point cloud is altered. In the multivariate case, given $\vec{\psi} = (\psi_j)_{j=1}^k$ we use the notation $\vec{H}_n(\mathbf{S},\mathbf{T}) = (H_{n,j}(\mathbf{S},\mathbf{T}))_{j=1}^k$, where $H_{n,j} = \psi_j\left(\sqrt[d]{n}\mathbf{S}\right) - \psi_j\left(\sqrt[d]{n}\mathbf{T}\right)$.

B.2. Proofs of Section 2.2

Proposition B.1 (Proposition 2.1)**.** *For $\mathbf{S}$ a simple point process taking values in $\mathcal{X}\left(\mathbb{R}^d\right)$, let ψ stabilize on $\mathbf{S}$ almost surely. Then ψ stabilizes on $\mathbf{S}$ in probability.*

Proof. Let ρ be a radius of stabilization satisfying Definition 2.4. Likewise, let D^∞ be a corresponding terminal addition cost. For any $\rho(\mathbf{S}) \leq l < \infty$, $D(\mathbf{S} \cap B_z(l)) = D(\mathbf{S} \cap B_z(\rho(\mathbf{S})))$ $= D^\infty(\mathbf{S})$. Thus $\{D(\mathbf{S} \cap B_z(l)) \neq D^\infty(\mathbf{S})\} \subseteq \{\rho(\mathbf{S}) > l\}$, and consequently $\mathbb{P}^*[D(\mathbf{S} \cap B_z(l)) \neq D^\infty(\mathbf{S})] \leq \mathbb{P}^*[\rho(\mathbf{S}) > l] \to 0$. We see that ψ stabilizes in probability on $\mathbf{S}$ with terminal addition cost $D^\infty(\mathbf{S})$. □

Proposition B.2 (Proposition 2.2)**.** *For $\mathcal{R}$ the space of locally-determined radii of stabilization for ψ centered at $z \in \mathbb{R}^d$, let $\rho^*\colon \mathcal{X}\left(\mathbb{R}^d\right) \to [0,\infty]$ such that $\rho^*(S) = \inf_{\rho\in\mathcal{R}} \rho(S)$. Then ρ^* is a locally determined radius of stabilization for ψ centered at z.*

Proof. If all possible radii are infinite, the result follows trivially. Else for $S, T \in \mathcal{X}\left(\mathbb{R}^d\right)$ suppose $\rho^*(S) < \infty$ with $S \cap B_z(\rho^*(S)) = T \cap B_z(\rho^*(S))$. Since S and T have no accumulation points, for any $\epsilon > 0$ sufficiently small, we have $S \cap B_z(\rho^*(S) + \epsilon) = T \cap B_z(\rho^*(S) + \epsilon)$. There exists a locally determined radius of stabilization ρ such that $\rho(S) \leq \rho^*(S) + \epsilon$. As $S \cap B_z(\rho^*(S) + \epsilon) = T \cap B_z(\rho^*(S) + \epsilon)$ with $\rho(S) \leq \rho^*(S) + \epsilon$, we have that $S \cap B_z(\rho(S)) = T \cap B_z(\rho(S))$. Thus $\rho(S) = \rho(T)$ by the local-determination criterion. Then $\rho^*(T) \leq \rho(T) = \rho(S) \leq \rho^*(S) + \epsilon$. Since the choice of ϵ was arbitrary, we have $\rho^*(T) \leq \rho^*(S)$. Thus, $S \cap B_z(\rho^*(T)) = T \cap B_z(\rho^*(T))$. By similar arguments, $\rho^*(S) \leq \rho^*(T)$. Combining, $\rho^*(S) = \rho^*(T)$ must hold, and the result follows. □

B.3. Proofs of Section 2.3

Lemma B.3 (Lemma 2.3)**.** *For $p > 2$, let ψ satisfy (E2) with $u_a \leq p - 1$ for some $a > 2$. Then for any $M < \infty$, ψ satisfies (E1) for $\mathcal{C}_{p,M}\left(\mathbb{R}^d\right)$.*

Proof. Let $R > 0$ and $a > 2$ be as given such that $u_a \leq p - 1$. Define $I_n := \#\{\mathbf{Y}_n \cap B_{Y'}(R/\sqrt[d]{n})\} = \#\left\{(\sqrt[d]{n}\mathbf{Y}_n) \cap B_{\sqrt[d]{n}Y'}(R)\right\}$. Conditional on Y', I_n follows a binomial distribution with expectation $n\int_{B_{Y'}(R/\sqrt[d]{n})} g(y)\ \mathrm{d}y$, where g is a density of G. By (E2), we have that

$$\begin{aligned}
&\mathbb{E}\left[\left|\psi\left(\sqrt[d]{n}\left(\mathbf{Y}_n \cup \{Y'\}\right)\right) - \psi\left(\sqrt[d]{n}\mathbf{Y}_n\right)\right|^a\right] && \text{(B.6)}\\
&\leq \mathbb{E}\left[U_a\left(1 + I_n^{u_a}\right)\right] && \text{(B.7)}\\
&\leq U_a\left(1 + \mathbb{E}\left[I_n^{u_a}\right]\right). && \text{(B.8)}
\end{aligned}$$

Via Corollary 3 in [35], there is a universal constant K such that the conditional u_a-th moment of I_n is at most

$$\begin{aligned}
&\left(K\frac{u_a}{\log(u_a)}\right)^{u_a} \max\left\{n\int_{B_{Y'}\left(\frac{R}{\sqrt[d]{n}}\right)} g(y)\ \mathrm{d}y, \left(n\int_{B_{Y'}\left(\frac{R}{\sqrt[d]{n}}\right)} g(y)\ \mathrm{d}y\right)^{u_a}\right\}\\
\leq\ &\left(K\frac{u_a}{\log(u_a)}\right)^{u_a}\left(n\int_{B_{Y'}\left(\frac{R}{\sqrt[d]{n}}\right)} g(y)\ \mathrm{d}y + \left(n\int_{B_{Y'}\left(\frac{R}{\sqrt[d]{n}}\right)} g(y)\ \mathrm{d}y\right)^{u_a}\right).
\end{aligned}$$

Removing the conditioning on Y', for V_d the volume of a unit ball in $\mathbb{R}^d$, we have

$$\begin{aligned}
&\int_{\mathbb{R}^d} n\left(\int_{B_x\left(\frac{R}{\sqrt[d]{n}}\right)} g(y)\ \mathrm{d}y\right) g(x)\ \mathrm{d}x\\
&= \int_{\mathbb{R}^d}\int_{B_0(R)} g\left(x + \frac{t}{\sqrt[d]{n}}\right) g(x)\ \mathrm{d}t\ \mathrm{d}x\\
&= \int_{B_0(R)}\int_{\mathbb{R}^d} g\left(x + \frac{t}{\sqrt[d]{n}}\right) g(x)\ \mathrm{d}x\ \mathrm{d}t\\
&\leq V_d R^d \|g\|_2^2\\
&\leq V_d R^d \|g\|_p^{\frac{p}{p-1}}\\
&\leq V_d R^d M^{\frac{p}{p-1}}
\end{aligned}$$

and

$$\begin{aligned}
&\int_{\mathbb{R}^d} \left(n \int_{B_x\left(\frac{R}{\sqrt[d]{n}}\right)} g(y)\ \mathrm{d}y \right)^{u_a} g(x)\ \mathrm{d}x \\
&= \int_{\mathbb{R}^d} \left(\int_{B_0(R)} g\left(x + \frac{t}{\sqrt[d]{n}}\right)\ \mathrm{d}t \right)^{u_a} g(x)\ \mathrm{d}x \\
&\leq \left(V_d R^d\right)^{u_a - 1} \int_{\mathbb{R}^d} \int_{B_0(R)} g\left(x + \frac{t}{\sqrt[d]{n}}\right)^{u_a} g(x)\ \mathrm{d}t\ \mathrm{d}x \\
&= \left(V_d R^d\right)^{u_a - 1} \int_{B_0(R)} \int_{\mathbb{R}^d} g\left(x + \frac{t}{\sqrt[d]{n}}\right)^{u_a} g(x)\ \mathrm{d}x\ \mathrm{d}t \\
&\leq \left(V_d R^d\right)^{u_a} \|g\|_{u_a+1}^{u_a+1} \\
&\leq \left(V_d R^d\right)^{u_a} \|g\|_p^{\frac{p}{p-1}u_a} \\
&\leq \left(V_d R^d M^{\frac{p}{p-1}}\right)^{u_a}. \qquad \text{(B.9)}
\end{aligned}$$

Combining, we have

$$\mathbb{E}\left[\left|\psi\left(\sqrt[d]{n}\left(\mathbf{Y}_n \cup \{Y'\}\right)\right) - \psi\left(\sqrt[d]{n}\mathbf{Y}_n\right)\right|^a\right] \qquad \text{(B.10)}$$

$$\leq U_a \left(1 + \left(K \frac{u_a}{\log(u_a)}\right)^{u_a} \left(V_d R^d M^{\frac{p}{p-1}} + \left(V_d R^d M^{\frac{p}{p-1}}\right)^{u_a}\right)\right). \qquad \text{(B.11)}$$

Since this bound does not depend on G or n, (E1) is satisfied by ψ for $\mathcal{C}_{p,M}\left(\mathbb{R}^d\right)$. □

Lemma B.4 (Lemma 2.4). *Let ψ satisfy (S2) for $F \in C_{p,M}\left(\mathbb{R}^d\right)$. Then ψ satisfies (S1) for $C_{p,M}\left(\mathbb{R}^d\right)$, F, $b = (p-2)/(d(p-1))$, and any $(l_\epsilon)_{\epsilon>0}$ such that $\lim_{\epsilon\to 0} l_\epsilon \epsilon^{(p-2)/(d(p-1))} = 0$ and $\lim_{\epsilon\to 0} l_\epsilon = \infty$.*

Proof. Let $\{X_i\}_{i\in\mathbb{N}} \overset{\text{iid}}{\sim} F$ with $X' \sim F$ an independent copy. Likewise, for $G \in \mathcal{C}_{p,M}\left(\mathbb{R}^d\right) \cap B_F(\epsilon; d_{\mathrm{TV}})$, let $\{Y_i\}_{i\in\mathbb{N}} \overset{\text{iid}}{\sim} G$ with $Y' \sim G$ an independent copy. Denote $\mathbf{X}_n := \{X_i\}_{i=1}^n$. As $d_{\mathrm{TV}}(F, G) \leq \epsilon$, it may be assumed that $\{(X_i, Y_i)\}_{i\in\mathbb{N}}$ are iid with $\mathbb{P}[X_i \neq Y_i] \leq \epsilon$ for all $i \in \mathbb{N}$.

Let $(l_\epsilon)_{\epsilon>0}$ be such that $\lim_{\epsilon\to 0} l_\epsilon \epsilon^{(p-2)/(d(p-1))} = 0$. Define the following sets:

$$A_Y := \{Y' = X'\} \qquad \text{(B.12)}$$

$$B_{Y,l_\epsilon} := \left\{ \mathbf{Y}_n \cap B_{X'}\left(\frac{l_\epsilon}{\sqrt[d]{n}}\right) = \mathbf{X}_n \cap B_{X'}\left(\frac{l_\epsilon}{\sqrt[d]{n}}\right) \right\} \qquad \text{(B.13)}$$

$$C_{l_\epsilon} := \left\{ \rho_{\sqrt[d]{n}X'}\left(\sqrt[d]{n}\mathbf{X}_n\right) \leq l_\epsilon \right\}. \qquad \text{(B.14)}$$

By the local-definition criterion, Definition 2.5, we have the following inclusion:

$$A_Y \cap B_{Y,l_\epsilon} \cap C_{X,l_\epsilon} \subseteq \left\{ \rho_{\sqrt[d]{n}Y'}\left(\sqrt[d]{n}\mathbf{Y}_n\right) \leq l_\epsilon \right\}.$$

Then

$$\begin{aligned}
&\mathbb{P}^*\left[\rho_{\sqrt[d]{n}Y'}\left(\sqrt[d]{n}\mathbf{Y}_n\right) > l_\epsilon\right] \\
&\leq \mathbb{P}^*\left[A_Y^c \cup B_{Y,l_\epsilon}^c \cup C_{l_\epsilon}^c\right] \\
&\leq \mathbb{P}\left[A_Y^c\right] + \mathbb{P}\left[B_{Y,l_\epsilon}^c\right] + \mathbb{P}^*\left[C_{l_\epsilon}^c\right]. \qquad \text{(B.15)}
\end{aligned}$$

Bounding each piece, $\mathbb{P}\left[A_Y^c\right] = \mathbb{P}\left[X' \neq Y'\right] \leq \epsilon$. Likewise, by (S2) we have $\mathbb{P}^*\left[C_{l_\epsilon}^c\right] = \mathbb{P}^*\left[\rho_{\sqrt[d]{n}X'}\left(\sqrt[d]{n}\mathbf{X}_n\right) > l_\epsilon\right] \leq p_\epsilon$, with p_ϵ not depending on G or n such that $\lim_{\epsilon\to 0} p_\epsilon = 0$. It thus remains to be shown that B_{Y,l_ϵ}^c occurs with small probability, uniformly in n and G.

As in (B.48) in the proof of Proposition 2.6, the probability that $\mathbf{X}_n$ and $\mathbf{Y}_n$ coincide within $B_{X'}\left(l_\epsilon/\sqrt[d]{n}\right)$ is at most $2M\frac{p}{p-1}V_d {l_\epsilon}^d \epsilon^{\frac{p-2}{p-1}}$. Thus we have that $\mathbb{P}\left[B_{Y,l_\epsilon}^c\right] \leq 2M\frac{p}{p-1}V_d {l_\epsilon}^d \epsilon^{\frac{p-2}{p-1}}$. The bound does not depend on G or n, with

$$\lim_{\epsilon\to 0} 2M\tfrac{p}{p-1}V_d l_\epsilon^d \epsilon^{\frac{p-2}{p-1}} = 2M\tfrac{p}{p-1}V_d\left(\lim_{\epsilon\to 0} l_\epsilon \epsilon^{\frac{p-2}{d(p-1)}}\right)^d = 0. \tag{B.16}$$

Finally, by the definition of a radius of stabilization we have that

$$\mathbb{P}\left[D_{\sqrt[d]{n}Y'}\left(\left(\sqrt[d]{n}\mathbf{Y}_n\right)\cap B_{\sqrt[d]{n}}\left(l_\epsilon\right)\right) \neq D_{\sqrt[d]{Y'}}\left(\mathbf{Y}_n\right)\right] \tag{B.17}$$

$$\leq \mathbb{P}^*\left[\rho_{\sqrt[d]{n}Y'}\left(\sqrt[d]{n}\mathbf{Y}_n\right) > l_\epsilon\right] \tag{B.18}$$

$$\leq \epsilon + p_\epsilon + 2M\tfrac{p}{p-1}V_d l_\epsilon^d \epsilon^{\frac{p-2}{p-1}}. \tag{B.19}$$

Here the final quantity does not depend on G or n, and goes to 0 as $\epsilon \to 0$. Thus (S1) is satisfied.

□

Lemma B.5 (Lemma 2.5)**.** *Let $F \in C_{p,M}$ with $p > 2$ and $M < \infty$. Let ρ_0 be a locally-determined radius of stabilization for ψ centered at 0. Suppose that for any given $a, b \in (0,\infty)$, and $\delta > 0$, there exists an $L_{a,b,\delta} < \infty$ and a measurable set $A_{a,b,\delta}$ with $\rho_0^{-1}\left((L_{a,b,\delta},\infty]\right) \subseteq A_{a,b,\delta}$ such that*

$$\sup_{\lambda\in[a,b]} \mathbb{P}^*\left[\rho_0\left(P_\lambda\right) > L_{a,b,\delta}\right] \leq \sup_{\lambda\in[a,b]} \mathbb{P}\left[\mathbf{P}_\lambda \in A_{a,b,\delta}\right] \leq \delta. \tag{B.20}$$

Then for any $\delta > 0$ there exists an $n_\delta < \infty$ and $L_\delta < \infty$ such that

$$\sup_{n\geq n_\delta} \mathbb{P}^*\left[\rho_0\left(\mathbf{X}_n - X'\right) > L_\delta\right] \leq \delta. \tag{B.21}$$

Proof. We consider $n \geq n_0$. Define two independent sets of random variables $(U_i)_{i=1}^\infty \overset{\text{iid}}{\sim} F$ and $(U_i^*)_{i=1}^\infty \overset{\text{iid}}{\sim} F$. For $N \sim \text{Pois}\left(n\right)$, denote by $\mathbf{P}_n$ the Poisson process given by $\{U_i\}_{i=1}^N$, having intensity nf over $\mathbb{R}^d$. We will couple this Poisson process to $\mathbf{X}_n$. $\{U_i\}_{i=1}^{N\vee n} \cup \{U_i^*\}_{i=1}^{(n-N)^+}$ has the same distribution as $\mathbf{X}_n$, thus we assume that the two random variables are equal. For a given random variable U_i or U_i^* and $L > 0$, the probability of falling within $B_{X'}\left(L/\sqrt[d]{n}\right)$

is bounded, as shown below. Applying the Cauchy-Schwartz inequality, we have

$$\begin{aligned}
&\int_{\mathbb{R}^d} \int_{B_x\left(\frac{L}{\sqrt[d]{n}}\right)} f(y) f(x) \ \mathrm{d}y \ \mathrm{d}x \\
&= \int_{\mathbb{R}^d} \int_{B_0\left(\frac{L}{\sqrt[d]{n}}\right)} f(x+t) f(x) \ \mathrm{d}t \ \mathrm{d}x \\
&= \int_{B_0\left(\frac{L}{\sqrt[d]{n}}\right)} \int_{\mathbb{R}^d} f(x+t) f(x) \ \mathrm{d}x \ \mathrm{d}t \\
&\leq \frac{V_d L^d}{n} \|f\|_2^2 \\
&\leq \frac{V_d L^d}{n} \|f\|_p^{\frac{p}{p-1}} \\
&\leq \frac{V_d L^d M^{\frac{p}{p-1}}}{n}.
\end{aligned}$$

The expected number of points within $B_{X'}\left(L/\sqrt[d]{n}\right)$ that contribute to $\mathbf{P}_n \triangle \mathbf{X}_n$ is then at most

$$\mathbb{E}\left[|N-n| \frac{V_d L^d M^{\frac{p}{p-1}}}{n}\right] \leq \frac{V_d L^d M^{\frac{p}{p-1}}}{n} \sqrt{\mathrm{Var}\left[N\right]} \leq \frac{M^{\frac{p}{p-1}} V_d L^d}{\sqrt{n}}. \tag{B.22}$$

As the number of differing points is an integer-valued random variable, this expectation bounds the probability that $\mathbf{X}_n$ and $\mathbf{P}_n$ differ within $B_{X'}\left(L/\sqrt[d]{n}\right)$. For a fixed value of L and sufficiently large n, the bound can be made arbitrarily small.

Next, we will couple the Poisson process $\mathbf{P}_n$ with a conditionally homogeneous approximation. We construct the following coupling: Let $\mathbf{T}$ be a homogeneous Poisson process on $\mathbb{R}^d \times [0,\infty)$ with unit intensity. The point process given by $\{U_i \text{ s.t. } (U_i, T_i) \in \mathbf{T}, T_i \leq nf(U_i)\}$ is then a nonhomogeneous Poisson process with intensity nf. We can safely assume that this process equals $\mathbf{P}_n$. Define the point process $\mathbf{H}_n := \{U_i \text{ s.t. } (U_i, T_i) \in \mathbf{T} \text{ and } T_i \leq nf(X')\}$.

Conditional on X', $\mathbf{H}_n$ is a homogeneous Poisson process with intensity $nf(X')$. The number of observations within $B_{X'}\left(L/\sqrt[d]{n}\right)$ that contribute to $\mathbf{P}_n \triangle \mathbf{H}_n$ follows a Poisson distribution with rate parameter

$$\int_{B_{X'}\left(\frac{L}{\sqrt[d]{n}}\right)} |nf(y) - nf(X')| \ \mathrm{d}y \tag{B.23}$$

Removing the conditioning on X', the expected number is

$$\int_{\mathbb{R}^d} \left(n \int_{B_x\left(\frac{L}{\sqrt[d]{n}}\right)} |f(y) - f(x)| \ \mathrm{d}y \right) f(x) \ \mathrm{d}x \tag{B.24}$$

As the expectation above is an upper bound for the probability that $\mathbf{P}_n$ and $\mathbf{H}_n$ fail to coincide within $B_{X'}\left(L/\sqrt[d]{n}\right)$, we show that this quantity can be made arbitrarily small. Consider C, the set of Lebesgue points of f. We have that C^c has Lebesgue measure 0 by the Lebesgue differentiation theorem. By the definition of a Lebesgue point, we may write

$$C = \bigcap_{\gamma>0} \bigcup_{\Delta>0} \bigcap_{\delta \leq \Delta} \left\{ x \in \mathbb{R} \text{ s.t. } \frac{\int_{B_x(\delta)} |f(y) - f(x)| \ \mathrm{d}y}{V_d \delta^d} \leq \gamma \right\} \tag{B.25}$$

Here V_d denotes the volume of a unit ball in $\mathbb{R}^d$. Now as f is a density, it may be shown that $\int_{B_x(\delta)} |f(y) - f(x)| \, \mathrm{d}y / V_d \delta^d$ is a jointly continuous function of x and δ, and therefore it is measurable. Via the continuity with respect to δ, we need only consider rational $\delta \leq \Delta$, because the rationals are dense in the reals. Thus,

$$C_{\Delta,\gamma} := \bigcap_{\delta \leq \Delta} \left\{ x \in \mathbb{R} \text{ s.t. } \frac{\int_{B_x(\delta)} |f(y) - f(x)| \, \mathrm{d}y}{V_d \delta^d} \leq \gamma \right\}$$

is a countable intersection of measurable sets. Finally, by the Archimedean principle and other standard calculus arguments, we may assume γ and Δ also come from a countable class, $\{1/n : n \in \mathbb{N}\}$, for example. Let $C_\gamma := \cup_{\Delta > 0} C_{\Delta,\gamma}$. We have that $C_{\delta,\gamma}$ and C_γ are measurable with $\lim_{\Delta \to 0} C^c_{\delta,\gamma} = C^c_\gamma$ and $\lim_{\gamma \to 0} C^c_\gamma = C^c$. By continuity of measure, the Lebesgue measure of C^c_γ must go to 0, as well for $\int_{C^c_\gamma} f(x) \, \mathrm{d}x$. We decompose the integral in (B.24) as follows. For any integer $1 < a \leq p - 1$, an application of Hölder's inequality gives

$$\begin{aligned}
&\int_{\mathbb{R}^d} \left(n \int_{B_x\left(\frac{L}{\sqrt[d]{n}}\right)} |f(y) - f(x)| \, \mathrm{d}y \right) f(x) \, \mathrm{d}x \\
&= \int_{\mathbb{R}^d} \left(n \int_{B_x\left(\frac{L}{\sqrt[d]{n}}\right)} |f(y) - f(x)| \, \mathrm{d}y \right) f(x) \, \mathbb{1}\left\{ x \in C_{\frac{L}{\sqrt[d]{n}},\gamma} \right\} f(x) \, \mathrm{d}x \\
&\quad + \int_{\mathbb{R}^d} \left(n \int_{B_x\left(\frac{L}{\sqrt[d]{n}}\right)} |f(y) - f(x)| \, \mathrm{d}y \right) f(x) \, \mathbb{1}\left\{ x \in C^c_{\frac{L}{\sqrt[d]{n}},\gamma} \right\} f(x) \, \mathrm{d}x \\
&\leq \gamma V_d L^d \\
&\quad + \left(\int_{\mathbb{R}^d} \left(\int_{B_0(L)} \left| f\left(x + \frac{t}{\sqrt[d]{n}} \right) - f(x) \right| \, \mathrm{d}t \right)^a f(x) \, \mathrm{d}x \right)^{\frac{1}{a}} \mathbb{P}\left[X' \in C^c_{\frac{L}{\sqrt[d]{n}},\gamma} \right]^{1 - \frac{1}{a}}.
\end{aligned} \tag{B.26}$$

For the integral above

$$\begin{aligned}
&\int_{\mathbb{R}^d} \left(\int_{B_0(L)} \left| f\left(x + \frac{t}{\sqrt[d]{n}} \right) - f(x) \right| \, \mathrm{d}t \right)^a f(x) \, \mathrm{d}x \\
&\leq \left(V_d L^d \right)^{a-1} \int_{\mathbb{R}^d} \int_{B_0(L)} \left| f\left(x + \frac{t}{\sqrt[d]{n}} \right) - f(x) \right|^a f(x) \, \mathrm{d}t \, \mathrm{d}x \\
&= \left(V_d L^d \right)^{a-1} \int_{B_0(L)} \int_{\mathbb{R}^d} \left| f\left(x + \frac{t}{\sqrt[d]{n}} \right) - f(x) \right|^a f(x) \, \mathrm{d}t \, \mathrm{d}x \\
&\leq 2^{a-1} \left(V_d L^d \right)^{a-1} \int_{B_0(L)} \int_{\mathbb{R}^d} \left(f\left(x + \frac{t}{\sqrt[d]{n}} \right)^a + f(x)^a \right) f(x) \, \mathrm{d}t \, \mathrm{d}x \\
&\leq \left(2 V_d L^d \right)^a \|f\|_{a+1}^{a+1} \\
&\leq \left(2 V_d L^d \right)^a \|f\|_p^{\frac{p}{p-1} a} \\
&\leq \left(2 V_d L^d M^{\frac{p}{p-1}} \right)^a.
\end{aligned}$$

Thus (B.26) is at most

$$\gamma V_d L^d + 2V_d L^d M^{\frac{p}{p-1}} \mathbb{P}\left[X' \in C^c_{\frac{L}{\sqrt[d]{n}},\gamma}\right]^{1-\frac{1}{a}} \tag{B.27}$$

$$\leq \gamma V_d L^d + 2V_d L^d M^{\frac{p}{p-1}} \mathbb{P}\left[X' \in C^c_{\frac{L}{\sqrt[d]{n}},\gamma}\right]^{\frac{p-2}{p-1}}. \tag{B.28}$$

This provides a bound for the probability that $\mathbf{P}_n$ and $\mathbf{H}_n$ fail to coincide within $B_{X'}\left(L/\sqrt[d]{n}\right)$. The bound holds in the limiting $p = \infty$ case and can be made arbitrarily small for γ sufficiently small and n sufficiently large. γ can be chosen as a function of F, L, and n to provide the tightest bound, but this requires specific knowledge of f. Combining with the previous steps, we have coupled $\mathbf{X}_n$ and $\mathbf{H}_n$ to be equal with arbitrarily high probability.

Now for $\eta, \zeta > 0$ define $D_{*,\eta} = f^{-1}\left([\eta, \infty)\right)$ and $D^*_\zeta = f^{-1}\left([0, \zeta]\right)$. For η sufficiently small and ζ sufficiently large, $\mathbb{P}\left[X' \in D^c_{*,\eta}\right]$ and $\mathbb{P}\left[X' \in D^{*c}_\zeta\right]$ can be made arbitrarily small.

By assumption, for any given η, ζ, and $\nu > 0$ there is an $L_{\eta,\zeta,\nu}$ and a measurable set $A_{\eta,\zeta,\nu}$ such that for any homogenous Poisson process $\mathbf{Q}_\lambda$ on $\mathbb{R}^d$ with intensity λ bounded between η and ζ, be have $(\rho_0)^{-1}\left((L_{\eta,\zeta,\nu}, \infty]\right) \subseteq A_{\eta,\zeta,\nu}$ and $\mathbb{P}^*\left[\rho_0\left(\mathbf{Q}_\lambda\right) > L_{\eta,\zeta,\nu}\right] \leq \mathbb{P}\left[\mathbf{Q}_\lambda \in A_{\eta,\zeta,\nu}\right] \leq \nu$. $L_{\eta,\zeta,\nu}$ is possibly increasing as $\eta \to 0$, $\zeta \to \infty$, and $\nu \to 0$.

As $\sqrt[d]{n}\left(\mathbf{H}_n - X'\right)$ is a homogeneous Poisson process, conditional on $X', \in D_{*,\eta} \cup D^*_\zeta$, we have

$$\mathbb{P}^*\left[\rho_0\left(\sqrt[d]{n}\left(\mathbf{H}_n - X'\right)\right) > L_{\eta,\zeta,\nu} \middle|\ X' \in D_{*,\eta} \cup D^*_\zeta\right] \tag{B.29}$$

$$\leq \mathbb{P}\left[\sqrt[d]{n}\left(\mathbf{H}_n - X'\right) \in A_{\eta,\zeta,\nu} \middle|\ X' \in D_{*,\eta} \cup D^*_\zeta\right] \tag{B.30}$$

$$= \mathbb{E}\left[\mathbb{P}\left[\sqrt[d]{n}\left(\mathbf{H}_n - X'\right) \in A_{\eta,\zeta,\nu} | X'\right] \middle|\ X' \in D_{*,\eta} \cup D^*_\zeta\right] \tag{B.31}$$

$$\leq \nu. \tag{B.32}$$

Combining the pieces and letting $L = L_{\eta,\zeta,\nu}$, we have that

$$\mathbb{P}^*\left[\rho_0\left(\mathbf{X}_n - X'\right) > L_{\eta,\zeta,\nu}\right] \tag{B.33}$$

$$= \mathbb{P}^*\left[\rho_0\left(\mathbf{H}_n - X'\right) > L_{\eta,\zeta,\nu} | X' \in A_{*,\eta} \cup A^*_\zeta\right] \mathbb{P}\left[X' \in D_{*,\eta} \cup D^*_\zeta\right] \tag{B.34}$$

$$+ \mathbb{P}\left[\mathbf{X}_n \cap B_{X'}\left(L_{\eta,\zeta,\nu}\right) \neq \mathbf{P}_n \cap B_{X'}\left(L_{\eta,\zeta,\nu}\right)\right] \tag{B.35}$$

$$+ \mathbb{P}\left[\mathbf{P}_n \cap B_{X'}\left(L_{\eta,\zeta,\nu}\right) \neq \mathbf{H}_n \cap B_{X'}\left(L_{\eta,\zeta,\nu}\right)\right] + \mathbb{P}\left[X' \in D^c_{*,\eta}\right] + \mathbb{P}\left[X' \in D^{*c}_\zeta\right] \tag{B.36}$$

$$\leq \nu + \frac{M^{\frac{p}{p-1}} V_d L^d_{\eta,\zeta,\nu}}{\sqrt{n}} + \gamma V_d L^d_{\eta,\zeta,\nu} + 2V_d L^d_{\eta,\zeta,\nu} M^{\frac{p}{p-1}} \mathbb{P}\left[X' \in C^c_{\frac{L_{\eta,\zeta,\nu}}{\sqrt[d]{n}},\gamma}\right]^{\frac{p-2}{p-1}} \tag{B.37}$$

$$+ \mathbb{P}\left[X' \in D^c_{*,\eta}\right] + \mathbb{P}\left[X' \in D^{*c}_\zeta\right]. \tag{B.38}$$

As $\eta, \nu \to 0$ and $\zeta \to \infty$, $L_{\eta,\zeta,\nu}$ can become unbounded. Let $\gamma \to 0$ and choose n_0 suitably large to ensure that the entire expression goes to 0. The result follows.

□

Proposition B.6 (Proposition 2.6)**.** *For $p > 2$ and $M < \infty$, let ψ satisfy (E1) and (S1) for $\mathcal{C}_{p,M}\left(\mathbb{R}^d\right)$, $F \in \mathcal{C}_{p,M}\left(\mathbb{R}^d\right)$, and some $a > 2$. Then for any $G \in \mathcal{C}_{p,M}\left(\mathbb{R}^d\right) \cap B_F\left(\epsilon, d_{TV}\right)$, there exist iid coupled random variables $\left((X_i, Y_i)\right)_{i\in\mathbb{N}}$ such that $\mathbf{X}_n = \{X_i\}_{i=1}^n \overset{iid}{\sim} F$, $\mathbf{Y}_n = \{Y_i\}_{i=1}^n \overset{iid}{\sim} G$, and*

$$\sup_{n\in\mathbb{N}} \operatorname{Var}\left[\frac{1}{\sqrt{n}}\left(\psi\left(\sqrt[d]{n}\mathbf{X}_n\right) - \psi\left(\sqrt[d]{n}\mathbf{Y}_n\right)\right)\right] \leq \gamma_\epsilon. \tag{B.39}$$

The value γ_ϵ does not depend on G and satisfies $\lim_{\epsilon\to 0}\gamma_\epsilon = 0$.

Proof. We refer to Appendix B.1 for a reference list of the general inequalities used here. Our proof technique is inspired by that of Proposition 5.4 in [31]. We expand using a martingale difference sequence (MDS). Let $\{(X_i, Y_i)\}_{i=1}^\infty$ be iid such that $\{X_i\}_{i=1}^\infty \overset{\text{iid}}{\sim} F$ and $\{Y_i\}_{i=1}^\infty \overset{\text{iid}}{\sim} G$. Each pair (X_i, Y_i) can be identically coupled such that $\mathbb{P}\left[X_i \neq Y_i\right] = d_{TV}\left(F, G\right) \leq \epsilon$. For $\mathbf{X}_j := \{X_i\}_{i=1}^j$, $\mathbf{Y}_j := \{Y_i\}_{i=1}^j$, and σ denoting a generated sigma algebra, let $\mathcal{F}_j := \sigma\{\mathbf{X}_j, \mathbf{Y}_j\}$ with $\mathcal{F}_0 := \{\Omega, \emptyset\}$. For (X', Y') an independent copy of the (X_i, Y_i), let

$$\begin{aligned}\mathbf{X}'_{n,j} &:= \{X_1, ..., X_{j-1}, X', X_{j+1}, ..., X_n\}\\ \mathbf{Y}'_{n,j} &:= \{Y_1, ..., Y_{j-1}, Y', Y_{j+1}, ..., Y_n\}.\end{aligned}$$

We apply the condensed notation $H_n\left(\mathbf{S}, \mathbf{T}\right) = \psi\left(\sqrt[d]{n}\mathbf{S}\right) - \psi\left(\sqrt[d]{n}\mathbf{T}\right)$. Using the orthogonality of a MDS and the conditional version of Jensen's inequality,

$$\begin{aligned}&\operatorname{Var}\left[\frac{1}{\sqrt{n}}H_n\left(\mathbf{X}_n, \mathbf{Y}_n\right)\right] &\text{(B.40)}\\ &= \frac{1}{n}\mathbb{E}\left[\left|\sum_{j=1}^n \mathbb{E}\left[H_n\left(\mathbf{X}_n, \mathbf{Y}_n\right)|\mathcal{F}_j\right] - \mathbb{E}\left[H_n\left(\mathbf{X}_n, \mathbf{Y}_n\right)\middle|\mathcal{F}_{j-1}\right]\right|^2\right]\\ &= \frac{1}{n}\mathbb{E}\left[\left|\sum_{j=1}^n \mathbb{E}\left[H_n\left(\mathbf{X}_n, \mathbf{Y}_n\right) - H_n\left(\mathbf{X}'_{n,j}, \mathbf{Y}'_{n,j}\right)|\mathcal{F}_j\right]\right|^2\right]\\ &= \frac{1}{n}\sum_{j=1}^n \mathbb{E}\left[\mathbb{E}\left[H_n\left(\mathbf{X}_n, \mathbf{Y}_n\right) - H_n\left(\mathbf{X}'_{n,j}, \mathbf{Y}'_{n,j}\right)|\mathcal{F}_j\right]^2\right]\\ &\leq \mathbb{E}\left[\left|H_n\left(\mathbf{X}_n, \mathbf{Y}_n\right) - H_n\left(\mathbf{X}'_{n,j}, \mathbf{Y}'_{n,j}\right)\right|^2\right]. &\text{(B.41)}\end{aligned}$$

The above holds for any $1 \leq j \leq n$. We have an upper bound for (B.41) of

$$\begin{aligned}&2\mathbb{E}\left[\left|H_n\left(\mathbf{X}_n \cup X', \mathbf{X}_n\right) - H_n\left(\mathbf{Y}_n \cup Y', \mathbf{Y}_n\right)\right|^2\right]\\ &\quad + 2\mathbb{E}\left[\left|H_n\left(\mathbf{X}_n \cup X', \mathbf{X}'_{n,j}\right) - H_n\left(\mathbf{Y}_n \cup Y', \mathbf{Y}'_{n,j}\right)\right|^2\right]\\ &= 4\mathbb{E}\left[\left|H_n\left(\mathbf{X}_n \cup X', \mathbf{X}_n\right) - H_n\left(\mathbf{Y}_n \cup Y', \mathbf{Y}_n\right)\right|^2\right]. &\text{(B.42)}\end{aligned}$$

We will decompose the expectation in (B.42) using the stabilization of ψ. Let $L > 0$, and define the following sets. Note that when all four are satisfied, $H_n(\mathbf{X}_n \cup X', \mathbf{X}_n) = H_n(\mathbf{Y}_n \cup Y', \mathbf{Y}_n)$.

$$A_Y := \{Y' = X'\} \tag{B.43}$$

$$B_{Y,L} := \left\{\mathbf{Y}_n \cap B_{X'}\left(\frac{L}{\sqrt[d]{n}}\right) = \mathbf{X}_n \cap B_{X'}\left(\frac{L}{\sqrt[d]{n}}\right)\right\} \tag{B.44}$$

$$C_{X,L} := \left\{D^{\infty}_{\sqrt[d]{n}X'}\left(\left(\sqrt[d]{n}\mathbf{X}_n\right) \cap B_{\sqrt[d]{n}X'}(L)\right) = D_{\sqrt[d]{n}X'}\left(\sqrt[d]{n}\mathbf{X}_n\right)\right\} \tag{B.45}$$

$$C_{Y,L} := \left\{D^{\infty}_{\sqrt[d]{n}Y'}\left(\left(\sqrt[d]{n}\mathbf{Y}_n\right) \cap B_{\sqrt[d]{n}Y'}(L)\right) = D_{\sqrt[d]{n}Y'}\left(\sqrt[d]{n}\mathbf{Y}_n\right)\right\}. \tag{B.46}$$

Let $C_{X,L*} \subseteq C_{X,L}$ and $C_{Y,L*} \subseteq C_{Y,L}$ be measurable. We decompose the expectation in (B.42) along these events into

$$\begin{aligned}
&\mathbb{E}\left[\left|H_n(\mathbf{X}_n \cup X', \mathbf{X}_n) - H_n(\mathbf{Y}_n \cup Y', \mathbf{Y}_n)\right|^2 \mathbb{1}\left\{A_Y \cap B_{Y,L} \cap C_{X,L*} \cap C_{Y,L*}\right\}\right] \\
&\quad + \mathbb{E}\left[\left|H_n(\mathbf{X}_n \cup X', \mathbf{X}_n) - H_n(\mathbf{Y}_n \cup Y', \mathbf{Y}_n)\right|^2 \mathbb{1}\left\{A_Y^c \cup B_{Y,L}^c \cup C_{X,L*}^c \cup C_{Y,L*}^c\right\}\right] \\
&= \mathbb{E}\left[\left|H_n(\mathbf{X}_n \cup X', \mathbf{X}_n) - H_n(\mathbf{Y}_n \cup Y', \mathbf{Y}_n)\right|^2 \mathbb{1}\left\{A_Y^c \cup B_{Y,L}^c \cup C_{X,L*}^c \cup C_{Y,L*}^c\right\}\right].
\end{aligned}$$

Let $a > 2$ satisfy (E1). Hölder's inequality gives an upper bound of

$$\left|\left|\left|H_n(\mathbf{X}_n \cup X', \mathbf{X}_n) - H_n(\mathbf{Y}_n \cup Y', \mathbf{Y}_n)\right|^2\right|\right|_{\frac{a}{2}} \mathbb{P}\left[A_Y^c \cup B_{Y,L}^c \cup C_{X,L*}^c \cup C_{Y,L*}^c\right]^{1-\frac{2}{a}}.$$

As the choice of $C_{X,L*}$ and $C_{Y,L*}$ was arbitrary, the expectation in (B.42) is at most

$$\begin{aligned}
&\left|\left|\left|H_n(\mathbf{X}_n \cup X', \mathbf{X}_n) - H_n(\mathbf{Y}_n \cup Y', \mathbf{Y}_n)\right|^2\right|\right|_{\frac{a}{2}} \mathbb{P}^*\left[A_Y^c \cup B_{Y,L}^c \cup C_{X,L}^c \cup C_{Y,L}^c\right]^{1-\frac{2}{a}} \\
&\leq \left|\left|\left|H_n(\mathbf{X}_n \cup X', \mathbf{X}_n) - H_n(\mathbf{Y}_n \cup Y', \mathbf{Y}_n)\right|^2\right|\right|_{\frac{a}{2}} \\
&\quad \times \max\left\{\mathbb{P}\left[A_Y^c\right] + \mathbb{P}\left[B_{Y,L}^c\right] + \mathbb{P}^*\left[C_{X,L}^c\right] + \mathbb{P}^*\left[C_{Y,L}^c\right], 1\right\}^{1-\frac{2}{a}}.
\end{aligned}$$

Consider the norm in the final expression above. We have an upper bound of

$$2\left(\left|\left|H_n(\mathbf{X}_n \cup X', \mathbf{X}_n)^2\right|\right|_{\frac{a}{2}} + \left|\left|H_n(\mathbf{Y}_n \cup Y', \mathbf{Y}_n)^2\right|\right|_{\frac{a}{2}}\right) \leq 4E_a^{\frac{2}{a}}. \tag{B.47}$$

This final quantity does not depend on ϵ, G, or n. It thus remains to show that, for a certain choice of L and as $\epsilon \to 0$, that each of the events A_Y^c, $B_{Y,L}^c$, $C_{X,L}^c$, and $C_{Y,L}^c$ can be made to occur with small outer probability, uniformly in G and n. For A^c, this is satisfied because $\mathbb{P}[X' \neq Y'] \leq \epsilon$.

We then consider $B_{Y,L}^c$. The sample pairs which contribute to $\mathbf{X}_n \cap B_{X'}(L/\sqrt[d]{n})$ but not $\mathbf{Y}_n \cap B_{X'}(L/\sqrt[d]{n})$ are those (X_i, Y_i) for which $X_i \neq Y_i$ and either $\|X_i - X'\| \leq L/\sqrt[d]{n}$ or $\|Y_i - X'\| \leq L/\sqrt[d]{n}$. Conditional on X', their count follows a binomial distribution with expectation at most $n\mathbb{P}[X_i \neq Y_i] \int_{B_{X'}(L/\sqrt[d]{n})} \tilde{f}(y) + \tilde{g}(y)\ \mathrm{d}y$. Here $\tilde{f}$ and $\tilde{g}$ are the densities of X_i and Y_i conditional on the event $\{X_i \neq Y_i\}$. These densities can be shown to exist via the absolute continuity of F and G with respect to the Lebesgue measure on $\mathbb{R}^d$. Subsequently,

we have that $\|\tilde{f}\|_p \leq \|f\|_p / \mathbb{P}[X_i \neq Y_i] \leq M/\mathbb{P}[X_i \neq Y_i]$ and $\|\tilde{g}\|_p \leq M/\mathbb{P}[X_i \neq Y_i]$. Removing the conditioning on X', via Hölder's inequality the expected number of pairs which contribute to $\mathbf{X}_n \triangle \mathbf{Y}_n$ within $B_{X'}(L/\sqrt[d]{n})$ is at most

$$\begin{aligned}
&\int_{\mathbb{R}^d} \left(n\mathbb{P}[X_i \neq Y_i] \int_{B_x\left(\frac{L}{\sqrt[d]{n}}\right)} \tilde{f}(y) + \tilde{g}(y) \, \mathrm{d}y \right) f(x) \, \mathrm{d}x \\
&= \mathbb{P}[X_i \neq Y_i] \int_{\mathbb{R}^d} \int_{B_0(L)} \left(\tilde{f}\left(x + \frac{t}{\sqrt[d]{n}}\right) + \tilde{g}\left(x + \frac{t}{\sqrt[d]{n}}\right) \right) f(x) \, \mathrm{d}t \, \mathrm{d}x \\
&= \mathbb{P}[X_i \neq Y_i] \int_{B_0(L)} \int_{\mathbb{R}^d} \left(\tilde{f}\left(x + \frac{t}{\sqrt[d]{n}}\right) + \tilde{g}\left(x + \frac{t}{\sqrt[d]{n}}\right) \right) f(x) \, \mathrm{d}x \, \mathrm{d}t \\
&\leq \mathbb{P}[X_i \neq Y_i] V_d L^d \left(\|\tilde{f}\|_{\frac{p}{p-1}} + \|\tilde{g}\|_{\frac{p}{p-1}} \right) \|f\|_p \\
&\leq 2\mathbb{P}[X_i \neq Y_i] M V_d L^d \left(\frac{M}{\mathbb{P}[X_i \neq Y_i]} \right)^{\frac{1}{p-1}} \\
&\leq 2M^{\frac{p}{p-1}} V_d L^d \epsilon^{\frac{p-2}{p-1}}. \qquad \text{(B.48)}
\end{aligned}$$

This final expression provides an upper bound on $\mathbb{P}\left[B^c_{Y,L}\right]$. Let $(l_\epsilon)_{\epsilon>0}$ satisfy (S1) and $L = l_\epsilon$. We have that $\mathbb{P}\left[B^c_{Y,l_\epsilon}\right] \leq 2M^{\frac{p}{p-1}} V_d {l_\epsilon}^d \epsilon^{\frac{p-2}{p-1}} \to 0$. By (S1), both $\mathbb{P}^*\left[C^c_{X,l_\epsilon}\right]$ and $\mathbb{P}^*\left[C^c_{Y,l_\epsilon}\right]$ are bounded above by a quantity p_ϵ such that $\lim_{\epsilon\to 0} p_\epsilon = 0$.

Let $\delta_\epsilon = \min\left\{\epsilon + 2M^{\frac{p}{p-1}} V_d {l_\epsilon}^d \epsilon^{\frac{p-2}{p-1}} + 2p_\epsilon, 1\right\}$ be the derived upper bound for $\mathbb{P}^*\left[A^c_Y \cup B^c_{Y,l_\epsilon} \cup C^c_{X,l_\epsilon} \cup C^c_{Y,l_\epsilon}\right]$. We achieve a final upper bound for (B.40) of

$$16 E_a^{\frac{2}{a}} \delta_\epsilon^{1-\frac{2}{a}}. \qquad \text{(B.49)}$$

If (S1) is satisfied for many $(l_\epsilon)_{\epsilon>0}$ such that $\lim_{\epsilon\to 0} l_\epsilon \epsilon^{(p-2)/(d(p-1))} = 0$, l_ϵ can be further chosen to optimize the rate of δ_ϵ, provided a rate for p_ϵ. Furthermore, if (E1) is satisfied for more than one $a > 2$, $a = a_\epsilon$ may be chosen to optimize the final rate as $\epsilon \to 0$. Such considerations depend on the specifics of the statistic ψ and the density assumptions used. □

B.4. Proofs of Section 2.4

Theorem B.7 (Theorem 2.7)**.** *Let $F \in \mathcal{P}\left(\mathbb{R}^d\right)$ with density f such that $\|f\|_p < \infty$ for some $p > 2$. Furthermore, let F and $\hat{f}_n$ be such that $\|\hat{f}_n - f\|_1 \to 0$ and $\|\hat{f}_n - f\|_p \to 0$ in probability (resp. a.s.). Suppose $\vec{\psi}\colon \tilde{\mathcal{X}}\left(\mathbb{R}^d\right) \to \mathbb{R}^k$ has component functions $\psi_j\colon \tilde{\mathcal{X}}\left(\mathbb{R}^d\right) \to \mathbb{R}$, $1 \leq j \leq k$ satisfying (E1) and (S1) for $\mathcal{C}_{p,M}\left(\mathbb{R}^d\right)$, $M > \|f\|_p$, F, and $b = (p-2)/(d(p-1))$. Then for a sample $\mathbf{X}_n = \{X_i\}_{i=1}^n \overset{iid}{\sim} F$, $(m_n)_{n\in\mathbb{N}}$ such that $\lim_{n\to\infty} m_n = \infty$, a bootstrap sample $\mathbf{X}^*_{m_n} = \{X_i^*\}_{i=1}^{m_n} \overset{iid}{\sim} \hat{F}_n|\mathbf{X}_n$, and a multivariate distribution G,*

$$\frac{1}{\sqrt{n}}\left(\vec{\psi}\left(\sqrt[d]{n}\mathbf{X}_n\right) - \mathbb{E}\left[\vec{\psi}\left(\sqrt[d]{n}\mathbf{X}_n\right)\right]\right) \xrightarrow{d} G$$

if and only if

$$\frac{1}{\sqrt{m_n}}\left(\vec{\psi}\left(\sqrt[d]{m_n}\mathbf{X}^*_{m_n}\right) - \mathbb{E}\left[\vec{\psi}\left(\sqrt[d]{m_n}\mathbf{X}^*_{m_n}\right)\middle|\mathbf{X}_n\right]\right) \xrightarrow{d} G \text{ in probability (resp. a.s.).}$$

Proof. For any bounded, Lipschitz function $v\colon \mathbb{R}^k \to \mathbb{R}$, consider the functional given by $V_{m_n} := \mathbb{E}\left[v\left(\vec{H}_{m_n}\left(\mathbf{Y}_{m_n}\right)\right)\right]$, where $\mathbf{Y}_{m_n} = \{Y_i\}_{i=1}^{m_n}$ is an iid sample, and the functional takes as input the shared distribution of the Y_i. Let v be bounded within $[-L, L]$ with a Lipschitz constant of L. First assuming that $\vec{H}_n(\mathbf{X}_n) \xrightarrow{d} G$, we have $V_n(F) \to \int_{\mathbb{R}} v \, \mathrm{d}G$.

Now, let $\mathbf{X}'_{m_n} = \left\{X'_{m_n,i}\right\}_{i=1}^{m_n} \overset{\text{iid}}{\sim} F$ be an independent copy of $\mathbf{X}_{m_n} = \{X_i\}_{i=1}^{m_n}$. Furthermore, as in the proof of Proposition 2.6, $\mathbf{X}'_{m_n}$ and $\mathbf{X}^*_{m_n}$ can be coupled so that $\mathbb{P}\left[X'_{m_n,i} \neq X_i^*\right] = d_{TV}\left(F, \hat{F}_n\right) = \|\hat{f}_n - f\|_1/2$, conditional on $\mathbf{X}_n$. Via Proposition 2.6 and

Chebyshev's inequality, for $\delta > 0$ we have almost surely that

$$
\begin{aligned}
&V_{m_n}\left(\hat{F}_n\right) \\
&= \mathbb{E}\left[v\left(\vec{H}_{m_n}\left(\mathbf{X}^*_{m_n}\right)\right)|\mathbf{X}_n\right] \\
&= \mathbb{E}\left[v\left(\vec{H}_{m_n}\left(\mathbf{X}^*_{m_n}\right)\right)\mathbb{1}\left\{\|\vec{H}_{m_n}\left(\mathbf{X}^*_{m_n}\right)-\vec{H}_{m_n}\left(\mathbf{X}'_{m_n}\right)\|\leq\delta\right\}|\mathbf{X}_n\right] \\
&\quad + \mathbb{E}\left[v\left(\vec{H}_{m_n}\left(\mathbf{X}^*_{m_n}\right)\right)\mathbb{1}\left\{\|\vec{H}_{m_n}\left(\mathbf{X}^*_{m_n}\right)-\vec{H}_{m_n}\left(\mathbf{X}'_{m_n}\right)\|>\delta\right\}|\mathbf{X}_n\right] \\
&\leq \mathbb{E}\left[v\left(\vec{H}_{m_n}\left(\mathbf{X}'_{m_n}\right)\right)+L\delta|\mathbf{X}_n\right] \\
&\quad + L\left(2-\delta\right)\sum_{j=1}^{m_n}\mathbb{P}\left[\left|H_{m_n,j}\left(\mathbf{X}^*_{m_n}\right)-H_{m_n,j}\left(\mathbf{X}'_{m_n}\right)\right|>\frac{\delta}{\sqrt{k}}|\mathbf{X}_n\right] \\
&\leq \mathbb{E}\left[v\left(\vec{H}_{m_n}\left(\mathbf{X}'_{m_n}\right)\right)+L\delta|\mathbf{X}_n\right] \\
&\quad + L\left(2-\delta\right)\left(\sum_{j=1}^{k}\mathbb{P}\left[\left|H_{m_n,j}\left(\mathbf{X}^*_{m_n}\right)-H_{m_n,j}\left(\mathbf{X}'_{m_n}\right)\right|>\frac{\delta}{\sqrt{k}}|\mathbf{X}_n\right]\right)\mathbb{1}\left\{\|\hat{f}_n\|_p\leq M\right\} \\
&\quad + L\left(2-\delta\right)\mathbb{1}\left\{\|\hat{f}_n\|_p> M\right\} \\
&\leq \mathbb{E}\left[v\left(\vec{H}_{m_n}\left(\mathbf{X}'_{m_n}\right)\right)\right]+L\delta \\
&\quad + L\left(2-\delta\right)\left(\left(\sum_{j=1}^{k}\frac{k\gamma_{\|\hat{f}_n-f\|_1/2,j}}{\delta^2}\right)\mathbb{1}\left\{\|\hat{f}_n\|_p\leq M\right\}+\mathbb{1}\left\{\|\hat{f}_n\|_p> M\right\}\right).
\end{aligned}
$$

Here $\gamma_{\|\hat{f}_n-f\|_1/2,j}$ is as given in Proposition 2.6 applied to ψ_j for $\epsilon = \|\hat{f}_n - f\|_1/2$. Similarly, almost surely

$$
\begin{aligned}
&V_{m_n}\left(\hat{F}_n\right) && \text{(B.50)}\\
&\geq \mathbb{E}\left[v\left(\vec{H}_{m_n}\left(\mathbf{X}'_{m_n}\right)\right)\right]-L\delta && \text{(B.51)}\\
&\quad - L\left(2-\delta\right)\left(\left(\sum_{i=1}^{k}\frac{k\gamma_{\|\hat{f}_n-f\|_1/2,j}}{\delta^2}\right)\mathbb{1}\left\{\|\hat{f}_n\|_p\leq M\right\}+\mathbb{1}\left\{\|\hat{f}_n\|_p> M\right\}\right). && \text{(B.52)}
\end{aligned}
$$

As $\|\hat{f}_n - f\|_p \to 0$ and $M > \|f\|_p$, we have that the lower bound for $V_{m_n}\left(\hat{F}_n\right)$ converges to $\int_{\mathbb{R}} v\,\mathrm{d}G - L\delta$ and the upper bound converges to $\int_{\mathbb{R}} v\,\mathrm{d}G + L\delta$, either in probability or a.s., depending on assumptions. Since this holds for any $\delta > 0$, we have that $V_{m_n}\left(\hat{F}_n\right) \to \int_{\mathbb{R}} v\,\mathrm{d}G$ in probability (or a.s.).

Now we will show the converse direction. Let $\mathbf{X}^*_{m_n}$ and $\mathbf{X}'_{m_n}$ be as previously defined.

We have

$$\begin{aligned}
&V_{m_n}(F)\\
&= \mathbb{E}\left[\mathbb{E}\left[v\left(\vec{H}_{m_n}\left(\mathbf{X}'_{m_n}\right)\right)|\mathbf{X}_n\right]\right]\\
&\leq \mathbb{E}\left[\mathbb{E}\left[v\left(\vec{H}_{m_n}\left(\mathbf{X}^*_{m_n}\right)\right)|\mathbf{X}_n\right]\right] + L\delta\\
&\quad + L(2-\delta)\,\mathbb{E}\left[\min\left\{\sum_{i=1}^{k}\frac{k\gamma_{\|\hat{f}_n - f\|_1/2,j}}{\delta^2}, 1\right\}\mathbb{1}\left\{\|\hat{f}_n\|_p \leq M\right\} + \mathbb{1}\left\{\|\hat{f}_n|_p > M\right\}\right]
\end{aligned}$$

and

$$\begin{aligned}
&V_{m_n}(F)\\
&\geq \mathbb{E}\left[\mathbb{E}\left[v\left(\vec{H}_{m_n}\left(\mathbf{X}^*_{m_n}\right)\right)|\mathbf{X}_n\right]\right] - L\delta\\
&\quad - L(2-\delta)\,\mathbb{E}\left[\min\left\{\sum_{i=1}^{k}\frac{k\gamma_{\|\hat{f}_n - f\|_1/2,j}}{\delta^2}, 1\right\}\mathbb{1}\left\{\|\hat{f}_n\|_p \leq M\right\} + \mathbb{1}\left\{\|\hat{f}_n\|_p > M\right\}\right].
\end{aligned}$$

Each expectation involves only bounded random variables, thus the lower bound converges to $\int_{\mathbb{R}} v\,\mathrm{d}G - L\delta$ and the upper bound to $\int_{\mathbb{R}} v\,\mathrm{d}G + L\delta$, assuming $\mathbb{E}\left[v\left(\vec{H}_{m_n}\left(\mathbf{X}^*_{m_n}\right)\right)|\mathbf{X}_n\right] \to \int_{\mathbb{R}} v\,\mathrm{d}G$. This holds if the assumed convergence is either in probability or almost sure. Since this holds for any $\delta > 0$, we have $V_{m_n}(F) \to \int_{\mathbb{R}} v\,\mathrm{d}G$. Since our initial choice of v was arbitrary, the desired result follows. □

B.5. Proofs of Section 4.3

Lemma B.8 (Lemma 4.1)**.** *Let $F \in \mathcal{C}_{p,M}(\mathbb{R}^d)$ for some $p > 2$ and $M < \infty$, and let $\mathcal{K} = \{K^r\}_{r\in\mathbb{R}}$ be a filtration of simplicial complexes satisfying (K2), (D2), and (D3). Then for any $r \in \mathbb{R}$, $s \in \mathbb{R}$, and $q \geq 0$, $\beta_q^{r,s}(\mathcal{K})$ satisfies (S2) for F.*

Proof. We start by defining a crude locally-determined radius of stabilization. Let K be either K^r or K^s. Denote $\phi = \max\{\phi(r), \phi(s)\}$ as given by (D2). For $z \in \mathbb{R}^d$, $S \in \mathcal{X}(\mathbb{R}^d)$, and $a > \phi$, consider the connected components in $K(S \cap B_z(a))$ and $K((S \cap B_z(a)) \cup \{z\})$ with at least one simplex entirely contained within $B_z(\phi)$. By (D2), if these components are entirely contained within $B_z(a-\phi)$, no simplices will be added or removed from them within $K(S \cap B_z(b))$ or $K((S \cap B_z(b)) \cup \{z\})$ for any $b > a$. This property holds for both K^s or K^r. The persistent Betti numbers are additive with respect to connected components, thus the add-z cost is entirely defined by those components altered by the inclusion of z, which necessarily must include one simplex within $B_z(\phi)$. As such, a is a locally determined radius of stabilization for S in this case. Any changes to the simplices outside of a must contribute to different connected components, and thus do not influence the add-z cost.

Now, $\mathbf{X}_n$ contains n total points. Including one point within $B_z(\phi)$, the longest possible chain of n connected points reaches at most a radius of $n\phi$. Therefore, $\rho_{\sqrt[d]{n}X'}(\sqrt[d]{n}\mathbf{X}_n) = (n+1)\phi$ is a locally-determined radius of stabilization on $\sqrt[d]{n}\mathbf{X}_n$ centered at $\sqrt[d]{n}X'$, as shown in the previous paragraph. However, since this radius grows with n, it alone cannot provide for the desired result.

Given (D2) and (D3), by Theorem 4.3 in [31] and the proof thereof, there exists a locally-determined radius of stabilization ρ_0^* for $\beta_q^{r,s}(\mathcal{K})$ centered at 0 such that the conditions of

Lemma 2.5 are satisfied. It must be noted that the original statement of the referenced lemma does not give this result directly. However, a careful analysis of the provided proof yields this more general result with minimal additions, and is not restated here. By (K2), we may define a radius of stabilization ρ_z^* for $\beta_q^{r,s}$ centered at $z \in \mathbb{R}^d$ with $\rho_z^*(S) = \rho_0^*(S - z)$. Thus, for any $\delta > 0$, there exists an $L_\delta < \infty$ and $n_\delta < \infty$ such that $\mathbb{P}\left[\rho^*_{\sqrt[d]{n}X'}\left(\sqrt[d]{n}\mathbf{X}_n\right)\right] \leq \delta$ for all $n \geq N_\delta$.

Denote by $P_z(S)$ the union of all connected components in either $K(S)$ or $K(S \cup \{0\})$ with at least one simplex entirely contained within $B_z(\phi)$. For any center point $z \in \mathbb{R}^d$, define $\rho_z : \mathcal{X} \to [0, \infty]$ with $\rho_z(S) = \min\{\operatorname{diam}(P_z(S)) + \phi, \rho^*(S - z)\}$. We have that ρ_z is a locally-determined radius of stabilization.

For $n < n_\delta$, we have that $\rho_{\sqrt[d]{n}X'}\left(\sqrt[d]{n}\mathbf{X}_n\right) \leq (n_\delta + 1)\phi$ almost surely. For $n \geq n_\delta$, $\rho_{\sqrt[d]{n}X'}\left(\sqrt[d]{n}\mathbf{X}_n\right) \leq \rho^*_{\sqrt[d]{n}X'}\left(\sqrt[d]{n}\mathbf{X}_n\right) \leq L_\delta$ with probability at least $1 - \delta$. Therefore $\sup_{n\in\mathbb{N}} \mathbb{P}\left[\rho_{\sqrt[d]{n}X'}\left(\sqrt[d]{n}\mathbf{X}_n\right) > \max\{L_\delta, (n_\delta + 1)\phi\}\right] \leq \delta$, and the result follows. □

B.6. Proofs of Section 4.4

Corollary B.9 (Corollary 4.2)**.** *Let $q \geq 0$ and $p > 2q + 3$. Let $\mathcal{K}$ be a filtration of simplicial complexes satisfying* (K1)*,* (K2)*,* (D1)*, and* (D3)*. Then for any given $\vec{r}$, $\vec{s}$, Statement 4.1 holds for $\beta_q^{\vec{r},\vec{s}}$.*

Proof. For given $r, s \in \mathbb{R}$, we will verify that assumption (E2) is satisfied for $\psi = \beta_q^{r,s}(\mathcal{K})$. Let $\mathbf{Y}_n = \{Y_i\}_{i=1}^n$ be iid and Y' an independent copy. By the Geometric Lemma 3.1, a bound for the change in persistent Betti numbers when $\{\sqrt[d]{n}Y'\}$ is added to $\sqrt[d]{n}\mathbf{Y}_n$ is given by the number of new simplices introduced to the corresponding complexes. By (K1), (D1), it suffices to count the number of points within $\phi := \max\{\phi(r), \phi(s)\}$ of $\sqrt[d]{n}Y'$, the combinations of which include any possible new simplices. Let $I_n = \sum_{i=1}^n \mathbb{1}\{\|Y_i - Y'\| \leq \phi/\sqrt[d]{n}\}$. For any $a > 2$ we have

$$
\begin{aligned}
&\left|\beta_q^{r,s}\left(\mathcal{K}\left(\sqrt[d]{n}\left(\mathbf{Y}_n \cup \{Y'\}\right)\right)\right) - \beta_q^{r,s}\left(\mathcal{K}\left(\sqrt[d]{n}\mathbf{Y}_n\right)\right)\right|^a \\
&\leq \left|\#\left\{K_q^r\left(\sqrt[d]{n}\left(\mathbf{Y}_n \cup \{Y'\}\right)\right) \setminus K_q^r\left(\sqrt[d]{n}\mathbf{Y}_n\right)\right\}\right. \\
&\quad \left. + \#\left\{K_{q+1}^s\left(\sqrt[d]{n}\left(\mathbf{Y}_n \cup \{Y'\}\right)\right) \setminus K_{q+1}^s\left(\sqrt[d]{n}\mathbf{Y}_n\right)\right\}\right|^a \\
&\leq \left|\binom{I_n}{q} + \binom{I_n}{q+1}\right|^a \\
&= \binom{I_n + 1}{q+1}^a \\
&\leq \frac{1}{((q+1)!)^a}(I_n + 1)^{a(q+1)} \\
&\leq \frac{2^{a(q+1)-1}}{((q+1)!)^a}\left(I_n^{a(q+1)} + 1\right).
\end{aligned}
$$

In this case $R = \phi$, $U_a \leq 2^{a(q+1)-1}/((q+1)!)^a$ and $u_a = a(q+1)$. (E1) then follows from Lemma 2.3 for $p \geq a(q+1) + 1 > 2q + 3$. As (K1) and (D1) together imply (D2), (S2) is satisfied as shown in Lemma 4.1. Then (S1) follows from Lemma 2.4. Finally an application of Theorem 2.7 gives the desired result.

Referring to Proposition 2.6 and the proof thereof, for $p < \infty$, using $a = (p-1)/(q+1)$ we achieve an optimal rate for γ_ϵ of

$$O\left(\delta_\epsilon^{1-\frac{2q+2}{p-1}}\right). \tag{B.53}$$

Details of the calculation are omitted here. For $p = \infty$, using $a = a_\epsilon = 2 - \log(\delta_\epsilon)$ we achieve an optimal rate of

$$O\left(\delta_\epsilon \left(\frac{-\log(\delta_\epsilon)}{\log(-\log(\delta_\epsilon))}\right)^{2q+2}\right). \tag{B.54}$$

Both of these rates depend on δ_ϵ, the upper bound for the total probability found in the proof of Proposition 2.6. The techniques found in the proofs of Lemma 2.3 and Proposition 2.6 allow for a bound on δ_ϵ, provided a tail bound for $\sup_{n\in\mathbb{N}} \rho_0\left(\sqrt[d]{n}\left(\mathbf{Y}_n - Y'\right)\right)$. At this time, such a bound is unavailable, thus no explicit rate calculation is possible. □

Corollary B.10 (Corollary 4.3)**.** *Let $q \geq 0$ and $p > 2q+5$. Let $\mathcal{K}$ be a filtration of simplicial complexes satisfying* (K2)*,* (D2)*, and* (D3)*. Then for any given $\vec{r}$, $\vec{s}$, Statement 4.1 holds for $\beta_q^{\vec{r},\vec{s}}$.*

Proof. The proof follows exactly that of Corollary 4.2, thus we will omit many replicated details. Let $\mathbf{Y}_n = \{Y_i\}_{i=1}^n$ be iid and Y' an independent copy. Define $\phi := \max\{\phi(r), \phi(s)\}$.

Since we do not assume (K1) in this case, the addition of $\sqrt[d]{n}Y'$ to the complex may both add and remove simplices, but only within $B_{\sqrt[d]{n}Y'}(\phi)$ by (D2). Any additional simplices may have $\sqrt[d]{n}Y'$ as a vertex, whereas any removed simplices may only have vertices within $\sqrt[d]{n}\mathbf{Y}_n$. For $I_n = \sum_{i=1}^n \mathbb{1}\{\|Y_i - Y'\| \leq \phi/\sqrt[d]{n}\}$, via the Geometric Lemma 3.1 we have

$$\begin{aligned}
&\left|\beta_q^{r,s}\left(\mathcal{K}\left(\sqrt[d]{n}\left(\mathbf{Y}_n \cup \{Y'\}\right)\right)\right) - \beta_q^{r,s}\left(\mathcal{K}\left(\sqrt[d]{n}\mathbf{Y}_n\right)\right)\right| \\
&\leq \left|\beta_q^{r,s}\left(\mathcal{K}\left(\sqrt[d]{n}\left(\mathbf{Y}_n \cup \{Y'\}\right)\right) \cup \mathcal{K}\left(\sqrt[d]{n}\mathbf{Y}_n\right)\right) - \beta_q^{r,s}\left(\mathcal{K}\left(\sqrt[d]{n}\left(\mathbf{Y}_n \cup \{Y'\}\right)\right)\right)\right| \\
&\quad + \left|\beta_q^{r,s}\left(\mathcal{K}\left(\sqrt[d]{n}\left(\mathbf{Y}_n \cup \{Y'\}\right)\right) \cup \mathcal{K}\left(\sqrt[d]{n}\mathbf{Y}_n\right)\right) - \beta_q^{r,s}\left(\mathcal{K}\left(\sqrt[d]{n}\mathbf{Y}_n\right)\right)\right| \\
&\leq \#\left\{K_q^r\left(\sqrt[d]{n}\mathbf{Y}_n\right) \setminus K_q^r\left(\sqrt[d]{n}\left(\mathbf{Y}_n \cup \{Y'\}\right)\right)\right\} \\
&\quad + \#\left\{K_{q+1}^s\left(\sqrt[d]{n}\mathbf{Y}_n\right) \setminus K_{q+1}^s\left(\sqrt[d]{n}\left(\mathbf{Y}_n \cup \{Y'\}\right)\right)\right\} \\
&\quad + \#\left\{K_q^r\left(\sqrt[d]{n}\left(\mathbf{Y}_n \cup \{Y'\}\right)\right) \setminus K_q^r\left(\sqrt[d]{n}\mathbf{Y}_n\right)\right\} \\
&\quad + \#\left\{K_{q+1}^s\left(\sqrt[d]{n}\left(\mathbf{Y}_n \cup \{Y'\}\right)\right) \setminus K_{q+1}^s\left(\sqrt[d]{n}\mathbf{Y}_n\right)\right\} \\
&\leq \binom{I_n}{q+1} + \binom{I_n}{q+2} + \binom{I_n+1}{q+1} + \binom{I_n+1}{q+2} \\
&\leq 2\binom{I_n+2}{q+2} \\
&\leq \frac{2^{q+3}}{(q+2)!}(I_n+1)^{q+2}.
\end{aligned}$$

Thus for any $a > 2$,

$$\left|\beta_q^{r,s}\left(\mathcal{K}\left(\sqrt[d]{n}\left(\mathbf{Y}_n \cup \{Y'\}\right)\right)\right) - \beta_q^{r,s}\left(\mathcal{K}\left(\sqrt[d]{n}\mathbf{Y}_n\right)\right)\right|^a \leq \frac{2^{(a+1)(q+2)}}{((q+2)!)^a}\left(I_n^{a(q+2)} + 1\right).$$

(E2) is satisfied for $R = \phi$, $U_a = \left(2^{(a+1)(q+2)}\right)/((q+2)!)^a$, and $u_a = a\,(q+2)$. Thus for $p \geq a\,(q+2)+1 > 2q+5$, (E1) follows by Lemma 2.3. (S2) and thus (S1) follow from Lemmas 4.1 and 2.4, respectively. An application of Theorem 2.7 gives the final result.

For the rate in Proposition 2.6, for $p < \infty$, using $a = (p-1)/(q+2)$ we achieve an optimal rate for γ_ϵ of

$$O\left(\delta_\epsilon^{1-\frac{2q+4}{p-1}}\right). \tag{B.55}$$

For $p = \infty$, using $a_\epsilon = 2 - \log(\delta_\epsilon)$ we achieve an optimal rate of

$$O\left(\delta_\epsilon\left(\frac{-\log(\delta_\epsilon)}{\log(-\log(\delta_\epsilon))}\right)^{2q+4}\right). \tag{B.56}$$

□

Corollary B.11 (Corollary 4.4)**.** *Let $m < \infty$ and $p > 2m+3$. Let $\mathcal{K}$ be a filtration of simplicial complexes satisfying* (K1)*,* (K2)*,* (D1)*,* (D3)*, and* (D4)*. Then for any given $\vec{r}$, Statement 4.1 holds for $\chi^{\vec{r}}$.*

Corollary B.12 (Corollary 4.5)**.** *Let $m < \infty$ and $p > 2m+5$. Let $\mathcal{K}$ be a filtration of simplicial complexes satisfying* (K2)*,* (D2)*,* (D3)*, and* (D4)*. Then for any given $\vec{r}$, Statement 4.1 holds for $\chi^{\vec{r}}$.*

Proof. We prove together Corollaries 4.4 and 4.5. Recall that the Euler characteristic χ can be written as an alternating (finite) sum of the Betti numbers when (D4) holds. As mentioned after the proposition statement, since Proposition 2.6 holds for the Betti numbers in dimensions $0 \leq q \leq m$ under the assumed conditions (see the proofs of Corollaries 4.2 and 4.3), then the same holds for their (alternating) sum, namely the Euler characteristic. The proof of Theorem 2.7 applies without alteration. □

Corollary B.13 (Corollary 4.6)**.** *Let $p > 2$. Furthermore, let $F \in \mathcal{D}_{\gamma,r_0}(C)$ and $\mathbb{1}\left\{\hat{F}_n \in \mathcal{D}_{\gamma,r_0}(C)\right\} \to 1$ in probability (resp. a.s.). Then Statement 4.1 holds for $l_{NN,k}$.*

Proof. First, we will show that $\mathbb{E}\left[\left|l_{\mathrm{NN},k}\left(\sqrt[d]{n}\left(\mathbf{Y}_n \cup \{Y'\}\right)\right) - l_{\mathrm{NN},k}\left(\sqrt[d]{n}\mathbf{Y}_n\right)\right|^a\right]$ is uniformly bounded for $G \in \mathcal{D}_{\gamma,r_0}(C)$ and $Y', Y_1, ..., Y_n \overset{\text{iid}}{\sim} G$. Denote by A_{k+1} the $k+1$ nearest neighbors of $\sqrt[d]{n}Y'$ in $\sqrt[d]{n}\mathbf{Y}_n$. Denote by B_k the set of points in $\sqrt[d]{n}\mathbf{Y}_n$ for which $\sqrt[d]{n}Y'$ is among the k nearest neighbors.

It may be shown that $\#\{B_k\} \leq C_{d,k}$, where $C_{d,k}$ is a constant depending only on the dimension d and k. To show this, consider a cone of angle $\pi/6$ whose point lies on $\sqrt[d]{n}Y'$. For $y_1, ..., y_k$ the k closest points of B_k to $\sqrt[d]{n}Y'$ within the cone, it follows from basic geometric arguments that any point lying within the cone, but outside a radius of $\max\{\|y_i - \sqrt[d]{n}Y'\|\}_{i=1}^k$ from $\sqrt[d]{n}Y'$ must be closer to each of $y_1, ..., y_k$ than to $\sqrt[d]{n}Y'$. Thus, any cone of this type may contain at most k points of B_n. Since $\mathbb{R}^d$ may be covered by finitely many of these cones, there must exist the required bound $C_{d,k}$.

Now, consider the points of A_{k+1} and B_k. Let $R_{k+1,n} := \max\{\|y - \sqrt[d]{n}Y'\| : y \in A_n\}$. For any point y in B_n, the distance to each point of A_n is at most $\|y - \sqrt[d]{n}Y'\| + R_{k+1,n}$ by the triangle inequality. In this case, the introduction of $\sqrt[d]{n}Y'$ to the sample may reduce the contribution to $l_{\mathrm{NN},k}$ from the points in B_n by at most

$$l_{\mathrm{NN},k}\left(\sqrt[d]{n}\mathbf{Y}_n\right) - l_{\mathrm{NN},k}\left(\sqrt[d]{n}\left(\mathbf{Y}_n \cup \{Y'\}\right)\right) \leq C_{d,k}R_{k+1,n}.$$

Likewise, the contribution of $\sqrt[d]{n}Y'$ is bounded by

$$l_{\mathrm{NN},k}\left(\sqrt[d]{n}\left(\mathbf{Y}_n \cup \{Y'\}\right)\right) - l_{\mathrm{NN},k}\left(\sqrt[d]{n}\mathbf{Y}_n\right) \leq kR_{k,n} \leq kR_{k+1,n}.$$

Thus, we proceed by bounding $\mathbb{E}\left[R_{k+1,n}^a\right]$. For any $G \in \mathcal{D}_{\gamma,r_0}$

$$\mathbb{E}\left[\int_{r_0}^{\infty} \mathbb{P}\left[\|Y_j - Y'\| > \sqrt[a]{r} \mid Y'\right]^n \, \mathrm{d}r\right] \leq \mathrm{diam}\,(C)\left(1 - \gamma r_0^{\frac{d}{a}}\right)^n.$$

We apply a bound similar to Theorem 7 in [47]. In the statement of the referenced theorem, it is assumed that the above quantity is bounded by C_T/n for an appropriate constant C_T. Here, we may improve that to an exponential bound. Consequently, we have

$$\mathbb{E}\left[R_{k+1,n}^a\right] \leq \left(\frac{k+1}{\gamma}\right)^{\frac{a}{d}} + \mathrm{diam}\,(C)\, n^{\frac{a}{d}}\left(1 - \gamma r_0^{\frac{d}{a}}\right)^n + \frac{a\left(e/\left(k+1\right)\right)^{k+1}}{d\left(\gamma\right)^{\frac{a}{d}}} \int_{k+1}^{\infty} e^{-y} y^{k+\frac{a}{d}} \, \mathrm{d}y. \tag{B.57}$$

For any $a < \infty$, this quantity limits to a constant with $n \to \infty$, thus admitting a constant upper bound which holds for all $n \in \mathbb{N}$, satisfying (E1).

The required stabilization properties (2.4) are first established for a unit-intensity homogeneous Poisson process via Lemma 6.1 in [38]. Let ρ denote the minimal locally-determined radius of stabilization for $l_{\mathrm{NN},k}$. Let $\mathbb{P}_\lambda$ denote a homogeneous Poisson process with intensity λ. By the scaling properties of $l_{\mathrm{NN},k}$, we have $\rho_0\left(\mathbb{P}_\lambda\right) = \rho_0\left(\mathbb{P}_1/\sqrt[d]{\lambda}\right) = \rho_0\left(\mathbb{P}_1\right)/\sqrt[d]{\lambda}$. Thus, $\mathbb{P}^*\left[\rho_0\left(\mathbb{P}_\lambda\right) > L\right] = \mathbb{P}^*\left[\rho_0\left(\mathbb{P}_1\right) > \sqrt[d]{\lambda}L\right]$. For any $\lambda > 1$, $\mathbb{P}^*\left[\rho_0\left(\mathbb{P}_\lambda\right) > L\right] \leq \mathbb{P}^*\left[\rho_0\left(\mathbb{P}_1\right) > L\right]$. Likewise, for any $\lambda_* < 1$, we may choose L_δ such that $\mathbb{P}^*\left[\rho_0\left(\mathbb{P}_1\right) > \sqrt[d]{\lambda_*}L_\delta\right] \leq \delta$. Then $\mathbb{P}\left[\rho_0\left(\mathbb{P}_\lambda\right) > L_\delta\right] \leq \delta$ for all $\lambda \in [\lambda_*, \infty)$. Stabilization then extends to the binomial sampling setting via Lemma 2.5 and the translation invariance of $l_{\mathrm{NN},k}$. We have for any $\delta > 0$ that there exists an $n_\delta < \infty$ and $L_\delta^* < \infty$ such that $\mathbb{P}^*\left[\rho_{\sqrt[d]{n}Y'}\left(\sqrt[d]{n}\mathbf{Y}_n\right) > L_\delta^*\right] \leq \delta$. Both quantities do not depend specifically on G.

When restricted to C, we have an absolute upper bound of $\mathrm{diam}\,(C)\sqrt[d]{n}$ for the radius of stabilization, as all points will fall inside of C almost surely. We set $L_\delta = \max\{\mathrm{diam}\,(C)\sqrt[d]{n_\delta}, L_\delta^*\}$. Then $\mathbb{P}^*\left[\rho_{\sqrt[d]{n}Y'}\left(\sqrt[d]{n}\mathbf{Y}_n\right) > L_\delta\right] \leq \delta$ for all $n \in \mathbb{N}$, satisfying (S2).

We now have the required pieces to prove bootstrap convergence. Although $\mathcal{C}_{p,M} \cap \mathcal{D}_{\gamma,r_0}(C)$ is only a subset of $\mathcal{C}_{p,M}$, the proof and conclusion of Proposition 2.6 still apply. Likewise, the proof of Theorem 2.7 is easily altered to include the additional condition $\mathbb{1}\left\{\hat{F}_n \in D_{\gamma,r_0}(C)\right\} \to 1$. We omit details here. □

Appendix C: L_p Consistency of Kernel Density Estimators

In this section we discuss the L_p-norm consistency of the kernel density estimator under very mild conditions. To the best of our knowledge, the exact proof of this result could not be found in the kernel density literature, though it employs well-known results from probability theory. In the context of our smoothed bootstrap procedure, the L_p-norm convergence assumption of the KDE follows as a direct consequence of the following theorem. Notably,

the necessary assumptions for L_p-norm convergence for the KDE are strictly weaker than those of Theorem 2.7.

For Q a kernel with $\int_{\mathbb{R}^d} Q(x)\ \mathrm{d}x = 1$, define $Q_h(x) := Q(x/h)/h^d$. Let F be a probability distribution on $\mathbb{R}^d$ with corresponding density f and $\{X_i\}_{i\in\mathbb{N}} \overset{\text{iid}}{\sim} F$. The kernel density estimator for f with bandwidth h is

$$\hat{f}_{n,h}(x) := \frac{1}{n}\sum_{i=1}^{n} Q_h(x - X_i) \tag{C.1}$$

Proposition C.1. *Given $p \geq 2$, let $||Q||_p < \infty$ and $||f||_p < \infty$. Then for any h_n such that $\lim_{n\to\infty} h_n = \infty$ and $\lim_{n\to\infty} n^{p/(2d(p-1))} h_n = \infty$*

$$\left|\left|\hat{f}_{n,h_n} - f\right|\right|_p \overset{p}{\to} 0 \tag{C.2}$$

If further $\sum_{n\in\mathbb{N}} 1/\left(n^{p/2} h_n^{d(p-1)}\right) < \infty$

$$\left|\left|\hat{f}_{n,h_n} - f\right|\right|_p \overset{a.s.}{\to} 0 \tag{C.3}$$

Proof. The expectation of $\hat{f}_{n,h_n}$ is $Q_{h_n} * f$, where $*$ denotes the convolution operator. We expand the L_p-norm using the triangle inequality.

$$\left|\left|\hat{f}_{n,h_n} - f\right|\right|_p \leq \left|\left|\hat{f}_{n,h_n} - Q_{h_n} * f\right|\right|_p + ||Q_{h_n} * f - f||_p \tag{C.4}$$

Because $\int_{\mathbb{R}^d} Q_{h_n}(x)\ \mathrm{d}x = 1$ and $||f||_p < \infty$, the second term goes to 0 with $h_n \to 0$ via Theorem 8.14 in [25]. We focus on the first term of (C.4).

$$\mathbb{E}\left[\int \left|\hat{f}_{n,h_n}(x) - (Q_{h_n} * f)(x)\right|^p \mathrm{d}x\right] = \int \mathbb{E}\left[\left|\hat{f}_{n,h_n}(x) - (Q_{h_n} * f)(x)\right|^p\right] \mathrm{d}x \tag{C.5}$$

$$= \frac{1}{n^p}\int \mathbb{E}\left[\left|\sum_{i=1}^{n} Y_i(x)\right|^p\right] \mathrm{d}x \tag{C.6}$$

where $Y_i(x) := Q_{h_n}(x - X_i) - (Q_{h_n} * f)(x)$ are iid mean-zero random variables.

We symmetrize using independent Radamacher random variables $\{e_i\}_{i\in\mathbb{N}}$, letting $Z_i(x) := e_i Y_i(x)$. We have that $\mathbb{E}\left[|\sum_{i=1}^n Y_i(x)|^p\right] \leq 2^p \mathbb{E}\left[|\sum_{i=1}^n Z_i(x)|^p\right]$. By Corollary 3 in [35], there exists a universal constant $C < \infty$ such that, for any $j \in \mathbb{N}$

$$\mathbb{E}\left[\left|\sum_{i=1}^{n} Z_i(x)\right|^p\right] \leq C^p \left(\frac{p}{\log p}\right)^p \max\left\{\left(n\mathbb{E}\left[|Z_j(x)|^2\right]\right)^{\frac{p}{2}}, n\mathbb{E}\left[|Z_j(x)|^p\right]\right\} \tag{C.7}$$

$$= C^p \left(\frac{p}{\log p}\right)^p \max\left\{\left(n\mathbb{E}\left[|Y_j(x)|^2\right]\right)^{\frac{p}{2}}, n\mathbb{E}\left[|Y_j(x)|^p\right]\right\}$$

$$\leq C^p \left(\frac{p}{\log p}\right)^p \max\left\{n^{\frac{p}{2}}\mathbb{E}\left[|Y_j(x)|^p\right], n\mathbb{E}\left[|Y_j(x)|^p\right]\right\}$$

$$= C^p \left(\frac{p}{\log p}\right)^p n^{\frac{p}{2}}\mathbb{E}\left[|Y_j(x)|^p\right]. \tag{C.8}$$

Then

$$\mathbb{E}\left[\int \left|\hat{f}_{n,h_n}(x) - (Q_{h_n} * f)(x)\right|^p \, \mathrm{d}x\right] \leq \frac{2^p C^p}{n^{\frac{p}{2}}} \left(\frac{p}{\log p}\right)^p \int \mathbb{E}\left[|Y_j(x)|^p\right] \, \mathrm{d}x. \tag{C.9}$$

$$\begin{aligned}
\int \mathbb{E}\left[|Y_j(x)|^p\right] \, \mathrm{d}x &= \mathbb{E}\left[\int |Y_j(x)|^p \, \mathrm{d}x\right] && \text{(C.10)}\\
&= \int\int |Q_{h_n}(x-y) - (Q_{h_n} * f)(x)|^p f(y) \, \mathrm{d}x \, \mathrm{d}y \\
&\leq 2^{p-1} \int\int \left(|Q_{h_n}(x-y)|^p + |(Q_{h_n} * f)(x)|^p\right) f(y) \, \mathrm{d}x \, \mathrm{d}y \\
&= 2^{p-1}\left(||Q_{h_n}||_p^p + ||Q_{h_n} * f||_p^p\right) \\
&\leq 2^p ||Q_{h_n}||_p^p \\
&= \frac{2^p}{(h_n^d)^{p-1}} ||Q||_p^p. && \text{(C.11)}
\end{aligned}$$

The last inequality follows from Young's inequality for convolutions, given that $||f||_1 = 1$, f being a probability density.

$$\mathbb{E}\left[\int \left|\hat{f}_{n,h_n}(x) - (Q_{h_n} * f)(x)\right|^p \, \mathrm{d}x\right] \leq \; 4^p C^p \left(\frac{p}{\log p}\right)^p \frac{||Q||_p^p}{\left(n^{\frac{p}{2d(p-1)}} h_n\right)^{d(p-1)}} \tag{C.12}$$

As $\lim_{n\to\infty} n^{p/(2d(p-1))} h_n = \infty$ by assumption, this final bound goes to 0 with $n \to \infty$. For any $\epsilon > 0$, Markov's inequality gives

$$\begin{aligned}
\mathbb{P}\left[\left|\left|\hat{f}_{n,h_n} - Q_{h_n} * f\right|\right|_p \geq \epsilon\right] &= \mathbb{P}\left[\left|\left|\hat{f}_{n,h_n} - Q_{h_n} * f\right|\right|_p^p \geq \epsilon^p\right] && \text{(C.13)}\\
&\leq \frac{\mathbb{E}\left[\left|\left|\hat{f}_{n,h_n} - Q_{h_n} * f\right|\right|_p^p\right]}{\epsilon^p} && \text{(C.14)}
\end{aligned}$$

As was shown earlier, the right hand side goes to 0, thus $\left|\left|\hat{f}_{n,h_n} - Q_{h_n} * f\right|\right|_p \xrightarrow{p} 0$. As $||Q_{h_n} * f - f||_p \to 0$, an application of Slutsky's theorem gives the final result. If $\sum_{n\in\mathbb{N}} 1/\left(n^{p/2} h_n^{d(p-1)}\right) < \infty$, the almost sure result follows from Borel-Cantelli. $h_n = n^{-(p-1)/2} (\log(n))^2$ satisfies this criterion.

□

Appendix D: Details of Simulation Study

Provided here are the data generating functions, written in pseudocode, for the simulation study of Section 5. Each generator below corresponds to a distribution F_1-F_7 in Table 1. A description is included, explaining each case in more detail. In all of the following, $\mathbb{S}^{d-1}$ denotes the unit sphere in $\mathbb{R}^d$, $B_z(r)$ the ball with radius r around z, and $\mathrm{Unif}(S)$ the

uniform distribution on the set S. $\mathrm{N}\left(\mu, \sigma^2\right)$ denotes the normal distribution with mean μ and variance σ^2, and $\mathrm{Exp}\left(\lambda\right)$ is the exponential distribution with rate parameter λ. $\mathrm{Cauchy}\left(\lambda\right)$ denotes the Cauchy distribution with scale parameter λ, and $(\cdot)$ is used to show vector concatenation.

Generator 1:

$\theta \sim \mathrm{Unif}\left(\mathbb{S}^1\right)$
$S \sim \mathrm{Unif}\left(\{-1, 1\}\right)$
$R \sim \mathrm{Unif}\left([0, 1]\right)$
return $X = \theta R^{.9S}$

F_1 is radially symmetric around the origin, and the radius is such that the random variable is unbounded, and the L_8 norm of the overall density is finite. Furthermore, the density approaches infinity near the origin. This case is chosen so as to test the assumptions of Corollary 4.2 with regards to the required norm bound.

Generator 2:

$\theta \sim \mathrm{Unif}\left(\mathbb{S}^1\right)$
$S \sim \mathrm{Unif}\left(\{-1, 1\}\right)$
$R \sim \mathrm{Unif}\left([0, 1]\right)$
return $X = \theta R^{.55S}$

F_2 is radially symmetric around the origin, and the radius is such that the random variable is unbounded. The L_2 norm of the overall density is finite, but the L_8 norm is infinite. As with distribution F_1, the density approaches infinity near the origin. This case violates the assumptions of Corollary 4.2.

Generator 3:

$\theta \sim \mathrm{Unif}\left(\mathbb{S}^1\right)$
$X_1, X_2 \sim \mathrm{N}\left(0, .04\right)$
return $\theta + (Y_1, Y_2)$.

F_3 represents a ring in $\mathbb{R}^2$, combined with additive Gaussian noise. The variance parameter is chosen small enough so that the ring structure is not lost within the additive noise.

Generator 4:

$\theta \sim \mathrm{Unif}\left(B_0\left(1\right)\right)$
$X_1, X_2, X_3 \sim \mathrm{N}\left(0, .01\right)$
return $\theta + (Y_1, Y_2, Y_3)$.

F_4 is the uniform distribution on the unit ball in $\mathbb{R}^3$, with a small amount of additive noise included to slightly smooth the boundary at radius 1.

Generator 5:

1: $X \sim \text{Unif}\left(\left\{\begin{array}{l}(0.38741799, 0.24263535, 0.09535272)\\(0.25147839, 0.63824409, 0.62425101)\\(0.73988542, 0.80749034, 0.84972394)\\(0.26811913, 0.35911205, 0.08316547)\\(0.65954757, 0.04704809, 0.02113341)\end{array}\right\}\right)$
2: $Y_1, Y_2, Y_3 \sim \text{Exp}(25)$
return $X + (Y_1, Y_2, Y_3)$.

F_5 consists of 5 clusters, one around each of the provided opints in $\mathbb{R}^3$. Exponential noise is included to test the effects of heavier tails on the final coverage probability. The rate parameter was chosen large enough so that the 5 clusters remain distinct after noise addition.

Generator 6:

1: $\theta \sim \text{Unif}\left(\mathbb{S}^2\right)$
2: $Y_1, ..., Y_5 \sim \text{Cauchy}(.1)$
return $(\theta, 0, 0) + (Y_1, ..., Y_5)$

F_6 represents a 2-dimensional unit sphere embedded in a higher dimension $\mathbb{R}^6$. We have included additive Cauchy noise to investigate the effects of very heavy tails.

Generator 7:

1: $(\theta_1, \theta_2) \sim \text{Unif}\left(\mathbb{S}^1\right)$
2: $S \sim \text{Unif}(\{-1, 1\})$
3: $Y_1, ..., Y_{10} \sim \text{N}(0, .04)$
return $(\theta_1 + S, \theta_2, 0, ..., 0) + (Y_1, ..., Y_{10})$

F_7 represents a dual ring, or figure-8 embedded in $\mathbb{R}^{10}$. Full-dimensional Gaussian noise is added, with variance chosen small enough so that the dual rings are not closed upon noise addition. F_7 is included to illustrate the effects of the "curse of dimensionality" expected in higher dimensions.