SUPPLEMENT TO "BOOTSTRAPPING PERSISTENT BETTI NUMBERS AND OTHER STABILIZING STATISTICS"

BY BENJAMIN ROYCRAFT[1,a], JOHANNES KREBS[2,c] AND WOLFGANG POLONIK[1,b]

[1]Department of Statistics, University of California Davis, [a]btroycraft@ucdavis.edu; [b]wpolonik@ucdavis.edu

[2]Department of Mathematics, KU Eichstätt-Ingolstadt, [c]johannes.krebs@ku.de

APPENDIX A: DATA ANALYSIS

In this section we show how smoothed bootstrap estimation performs on a real dataset. Source code is available at github.com/btroycraft/stabilizing_statistics_bootstrap [8]. We consider a selection of galaxies from the Sloan Digital Sky Survey [1], 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 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 [3], we estimate the density of the matter distribution using the adaptive bandwidth selector described in [2]. 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 3. 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 3. 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 $r = 3, \ldots, 30$ 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.

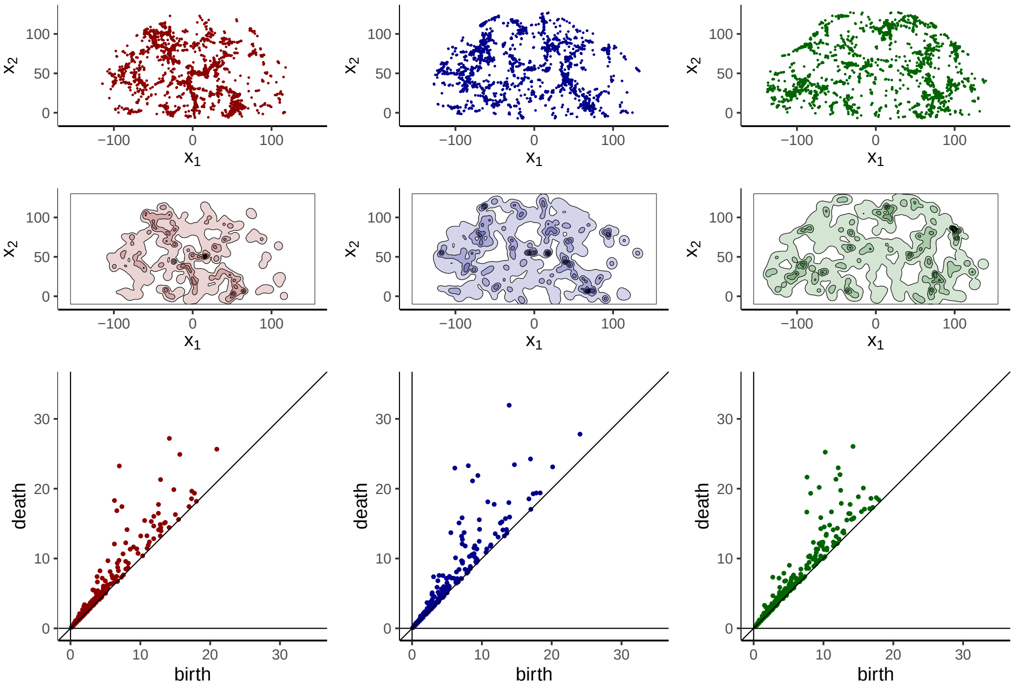

FIG 3. *Top row: Transformed point clouds. Middle row: Density estimates using adaptive bandwidth. Bottom row: Persistence diagrams in dimension $q = 1$ for the Vietoris-Rips complex. Columns from left to right: Galaxies with redshifts within* $(0.025, 0.026)$*,* $(0.027, 0.028)$*, and* $(0.029, 0.030)$*, respectively. Axis units are given in Megaparsecs (Mpc).*

We use bootstrap estimation to construct nominal 98% confidence intervals for the population mean values, both pointwise and simultaneous within each regime across $r = 3, \dots, 30$ Mpc. The number of bootstrap replicates used was $B = 20,000$, with results seen in Figure 4.

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

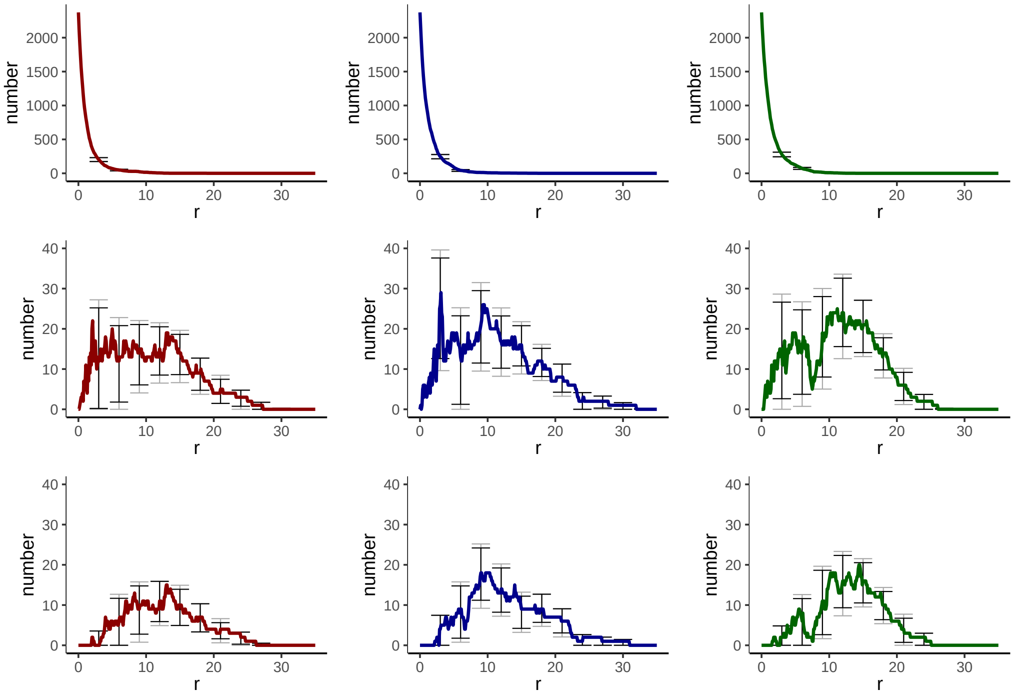

FIG 4. *Betti curves for the Vietoris-Rips complex. Top row: Betti numbers β_0^r. Middle row: Betti numbers β_1^r. Bottom Row: persistent Betti numbers $\beta_1^{r,r+1}$. Columns correspond with those of Figure 3. Axis units are given in Megaparsecs (Mpc). For each of $r = 3, \ldots, 30$ Mpc, simultaneous bootstrap confidence bands are given in gray, drawn from bootstrap samples of size $B = 20,000$. Likewise, pointwise intervals are given in black.*

APPENDIX B: PROOF OF MAIN RESULTS

B.1. Necessary Inequalities and Probability Results. 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 $k \geq 1$,

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

Likewise for $0 \leq k \leq 1$

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

Next, for any function $f\colon \mathbb{R}^d \to \mathbb{R}$ and $1 \leq a < b < c \leq \infty$,

$$\int |f|^b \, \mathrm{d}\lambda \leq \left(\int |f|^a \, \mathrm{d}\lambda\right)^{\frac{c-b}{c-a}} \left(\int |f|^c \, \mathrm{d}\lambda\right)^{\frac{b-a}{c-a}}.$$

Finally, for a locally-integrable function $f\colon \mathbb{R}^d \to \mathbb{R}$, the Hardy-Littlewood operator M is defined as $\mathrm{M}f(x) := \sup_{r>0} (1/V_d r^d) \int_{B_x(r)} |f| \, \mathrm{d}\lambda$. We have two maximal inequalities for M relating the behavior of $\mathrm{M}f$ to f. The weak-type inequality states that there exists a constant C_1 depending only on d such that $\int \mathbb{1}\{\mathrm{M}f > k\} \leq (C_1/k)\|f\|_1$. The strong-type inequality states that, for any $1 < p \leq \infty$ there exists a constant C_p depending only on p such that $\|\mathrm{M}f\|_p \leq C_p \|f\|_p$. These constants appear in several of the proofs throughout this work.

PROPOSITION B.1. *Let f and g be two probability density functions on $\mathbb{R}^d$ with $\|f\|_p < \infty$ for some $p > 1$. Then for any $q \in [1, p)$ there exists a continuous, increasing function $\xi \colon \mathbb{R}_+ \to \mathbb{R}_+$ depending only on f, p, and q such that $\lim_{\alpha \to 0} \xi(\alpha) = 0$ and*

$$\|f - g\|_q \leq \xi\Big(\|f - g\|_p\Big).$$

PROOF. Let λ denote the Lebesgue measure on $\mathbb{R}^d$. Via Markov's inequality, for any $a > 0$ we have

$$\begin{aligned}
\frac{1}{2}\int |f-g| \,\mathrm{d}\lambda &= \int |f-g| \,\mathbb{1}\{g < f\} \,\mathrm{d}\lambda \\
&= \int |f-g| (\mathbb{1}\{f - a \leq g < f\} + \mathbb{1}\{g \leq f - a\}) \,\mathrm{d}\lambda \\
&\leq \int \min\{f, a\} + |f-g| \,\mathbb{1}\{|f-g| \geq a\} \,\mathrm{d}\lambda \\
&\leq \int \min\{f, a\} \,\mathrm{d}\lambda + \frac{1}{a^{p-1}} \int |f-g|^p \,\mathrm{d}\lambda. \qquad (1)
\end{aligned}$$

By the Dominated Convergence Theorem, $\lim_{a \to 0} \int \min\{f, a\} \,\mathrm{d}\lambda = 0$. Thus $a \to \infty$ may be chosen such that both terms of (1) go to 0 as $\alpha \to 0$. Then for $\xi^* \colon \mathbb{R}_+ \to \mathbb{R}_+$ where $\xi^*(\alpha) := \inf_{a > 0} 2(\int \min\{f, a\} \,\mathrm{d}\lambda + \alpha^p / a^{p-1})$ we have $\|g - f\|_1 \leq \xi^*(\|g - f\|_p)$.

Now let $q \in [1, p)$ be given. The desired result holds for $\xi \colon \mathbb{R}_+ \to \mathbb{R}_+$ such that $\xi(\alpha) := \xi^*(\alpha)^{(p-q)/(q(p-1))} \alpha^{(p(q-1))/(q(p-1))}$. We have

$$\begin{aligned}
\|g - f\|_q &\leq \|g - f\|_1^{\frac{p-q}{q(p-1)}} \|g - f\|_p^{\frac{q-1}{q(p-1)}} \\
&\leq \xi^*\Big(\|g - f\|_p\Big)^{\frac{p-q}{q(p-1)}} \|g - f\|_p^{\frac{q-1}{q(p-1)}} \\
&= \xi\Big(\|g - f\|_p\Big).
\end{aligned}$$

□

PROPOSITION B.2. *Let $(\Omega, \sigma(\mu), \mu)$ be a measure space and η an extension of μ to $\sigma(\eta) \supseteq \sigma(\mu)$. Then given a $\sigma(\mu)$-measurable function $X : \Omega \to \mathbb{R}_+$ there exists a continuous, increasing function $\zeta \colon \mathbb{R}_+ \to \mathbb{R}_+$ depending only on X and μ such that $\lim_{\alpha \to 0} \zeta(\alpha) = 0$ and*

$$\int_A X \,\mathrm{d}\eta \leq \zeta(\eta(A)).$$

PROOF. Let $k_\alpha := \inf\{k \geq 0 \colon \int \mathbb{1}\{X > k\} \,\mathrm{d}\mu \leq \alpha\}$. For any $\sigma(\eta)$-measurable set A, we have

$$\begin{aligned}
\int_A X \,\mathrm{d}\eta &= \int_A X(\mathbb{1}\{X > k_\alpha\} + \mathbb{1}\{X \leq k_\alpha\}) \,\mathrm{d}\eta \\
&\leq \int_A X \,\mathbb{1}\{X > k_\alpha\} + k_\alpha \,\mathbb{1}\{X \leq k_\alpha\} \,\mathrm{d}\eta \\
&= \int_A X \,\mathbb{1}\{X > k_\alpha\} \,\mathrm{d}\eta + k_\alpha \int_{A^c} \mathbb{1}\{X > k_\alpha\} \,\mathrm{d}\eta
\end{aligned}$$

$$+ k_\alpha\left(\int_A \mathbb{1}\{X \le k_\alpha\}\,\mathrm{d}\eta - \int_{A^c} \mathbb{1}\{X > k_\alpha\}\,\mathrm{d}\eta\right)$$
$$\le \int X\,\mathbb{1}\{X > k_\alpha\}\,\mathrm{d}\eta + k_\alpha\left(\eta(A) - \int \mathbb{1}\{X > k_\alpha\}\,\mathrm{d}\eta\right).$$

Because X is $\sigma(\mu)$-measurable and $\eta(A) \le \alpha$, we have a final bound of $\xi(\alpha) := \int X\,\mathbb{1}\{X > k_\alpha\}\,\mathrm{d}\mu + k_\alpha(\alpha - \int \mathbb{1}\{X > k_\alpha\}\,\mathrm{d}\mu)$. Considering first only those α such that $\int \mathbb{1}\{X > k_\alpha\}\,\mathrm{d}\mu = \alpha$, the second term of ξ vanishes and $\lim_{\alpha\to 0}\xi(\alpha) = 0$ by the Dominated Convergence Theorem. Linear interpolation is applied via the second term whenever $\mathbb{1}\{X > k_\alpha\}\,\mathrm{d}\mu < \alpha$, extending to a continuous bound defined for all $\alpha \ge 0$. □

B.2. Proofs of Section 2.2.

PROPOSITION 2.5. *For* $\mathbf{S}$ *a simple point process taking values in* $\mathcal{X}(\mathbb{R}^d)$, *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 $\rho(\mathbf{S}) \le l < \infty$, $D(\mathbf{S}\cap B_z(l)) = D(\mathbf{S}\cap B_z(\rho(\mathbf{S}))) = D^\infty(\mathbf{S})$. Thus $\{D(\mathbf{S}\cap B_z(l)) \ne D^\infty(\mathbf{S})\} \subseteq \{\rho(\mathbf{S}) > l\}$, and consequently $\mathbb{P}^*[D(\mathbf{S}\cap B_z(l)) \ne D^\infty(\mathbf{S})] \le \mathbb{P}^*[\rho(\mathbf{S}) > l] \to 0$. We see that ψ stabilizes in probability on $\mathbf{S}$ with terminal addition cost $D^\infty(\mathbf{S})$. □

PROPOSITION 2.7. *For* $\mathcal{R}$ *the space of locally-determined radii of stabilization for* ψ *centered at* $z \in \mathbb{R}^d$, *let* $\rho^*\colon \mathcal{X}(\mathbb{R}^d) \to [0,\infty]$ *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 $S, T \in \mathcal{X}(\mathbb{R}^d)$ 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 $\rho(S) \le \rho^*(S) + \epsilon$. As $S \cap B_z(\rho^*(S)+\epsilon) = T \cap B_z(\rho^*(S)+\epsilon)$ with $\rho(S) \le \rho^*(S)+\epsilon$, 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 $\rho^*(T) \le \rho(T) = \rho(S) \le \rho^*(S) + \epsilon$. Since the choice of ϵ was arbitrary, we have $\rho^*(T) \le \rho^*(S)$. Thus, $S \cap B_z(\rho^*(T)) = T \cap B_z(\rho^*(T))$. By similar arguments, $\rho^*(S) \le \rho^*(T)$. Combining, $\rho^*(S) = \rho^*(T)$ must hold, and the result follows. □

B.3. Proofs of Section 2.3.

LEMMA 2.9. *Let* ψ *satisfy* (E2). *Then the following hold:*

1. *If* $\|f\|_{\max\{2u+1,2\}} < \infty$ *then* ψ *satisfies* (E1).
2. *If* $\|f\|_{\max\{p,2\}} < \infty$ *for some* $p > 2u+1$ *then* ψ *satisfies Statement 2.8.*

PROOF. We begin with some facts from elementary probability. Let $B_n \sim \mathrm{Binom}\,(n, \mu/n)$. Denote $p_j(\mu) := \mu^j e^{-\mu}/j!$. For any $j > \mu$, $\mathbb{P}[B_n = j] \le p_j(\mu)$, achieving equality in the limit as $n \to \infty$. Thus the upper tail probability and moments of B_n are bounded by those of a Poisson distribution with the same expectation. We have $\mathbb{E}[B_n^q\,\mathbb{1}\{B_n \ge k\}] \le \mu^q\,\mathbb{1}\{\mu \le k\} + \sum_{j=0}^\infty j^q p_q(\mu)\,\mathbb{1}\{j \ge k\}\,\mathbb{1}\{j > \mu\} \le \mu^q\,\mathbb{1}\{k \le \mu\} + \sum_{j=0}^\infty j^q p_q(\mu)\,\mathbb{1}\{j \ge k\}$ for any $q \in \mathbb{R}$, $k \in \mathbb{N}$.

Furthermore, for any $q \in [0,\infty)$, $\sum_{j=0}^{\infty} j^q p_j(\mu) = \lim_{n\to\infty} \mathbb{E}[B_n^q] \le \sup_{n\in\mathbb{N}} \mathbb{E}[B_n^q]$. We bound this quantity via Corollary 3 [6]. There exists a universal constant $K>0$ such that

$$\sup_{n\in\mathbb{N}} \mathbb{E}[B_n^q] \le \left(K\frac{q}{\log(q)}\right)^q \max\{\mu,\mu^q\} \le \left(K\frac{q}{\log(q)}\right)^q (\mu+\mu^q).$$

We continue with these facts established. Define $I_n := \#\{\mathbf{Y}_n \cap B_{Y'}(R/\sqrt[d]{n})\}$. Conditional on Y', I_n follows a binomial distribution with expectation $n\int_{B_{Y'}(R/\sqrt[d]{n})} g\,\mathrm{d}\lambda \le V_d R^d\,\mathrm{M}g(Y')$, where V_d is the volume of a unit ball in $\mathbb{R}^d$, $g := \mathrm{d}G/\mathrm{d}\lambda$, and M is the Hardy–Littlewood maximal operator such that $\mathrm{M}g(x) := \sup_{r\in\mathbb{R}_+} (1/V_d r^d)\int_{B_x(r)} g\,\mathrm{d}\lambda$.

We may prove both conclusions with a single argument. Let $p \ge 2u+1$ and $\|f\|_{\max\{p,2\}} < \infty$. Note because $\|f\|_1 = 1$, both $\|f\|_p<\infty$ and $\|f\|_2<\infty$. Because (E2) holds, we have that $|D_{\sqrt[d]{n}Y'}(\sqrt[d]{n}\mathbf{Y}_n)|^{(p-1)/u} \le 2^{(p-1)/u-1}U^{(p-1)/u}(1+I_n^{p-1})$. Then for $k=(k'/U-1)^{1/u}$,

$$\begin{aligned}
&\mathbb{E}\Big[\big|D_{\sqrt[d]{n}Y'}(\sqrt[d]{n}\mathbf{Y}_n)\big|^{\frac{p-1}{u}}\,\mathbb{1}\big\{\big|D_{\sqrt[d]{n}Y'}(\sqrt[d]{n}\mathbf{Y}_n)\big| > k'\big\}\Big] \\
&\le 2^{\frac{p-1}{u}-1}U^{\frac{p-1}{u}}\big(\mathbb{P}[I_n > k] + \mathbb{E}\big[I_n^{p-1}\,\mathbb{1}\{I_n > k\}\big]\big) \\
&\le 2^{\frac{p-1}{u}}U\,\mathbb{E}\big[I_n^{p-1}\,\mathbb{1}\{I_n > k\}\big]. \qquad (2)
\end{aligned}$$

Denote $\delta_2 := \|g-f\|_2$ and $\delta_p := \|g-f\|_p$. It then suffices to show that as $\delta_{\max\{p,2\}} \to 0$ there exists a choice $k\to\infty$ such that $\mathbb{E}[I_n^{p-1}\,\mathbb{1}\{I_n > k\}] \to 0$ uniformly in n and G. For a given $k \ge 0$, the expectation in (2) is bounded above by

$$\int \left(\Big(V_d R^d\,\mathrm{M}g\Big)^{p-1}\mathbb{1}\Big\{V_d R^d\,\mathrm{M}g \ge k\Big\} + \sum_{j=0}^{\infty} j^{p-1}p_j\Big(V_d R^d\,\mathrm{M}g\Big)\,\mathbb{1}\{j \ge k\}\right) g\,\mathrm{d}\lambda. \qquad (3)$$

Consider the first term in (3). we have

$$\begin{aligned}
\int (\mathrm{M}g)^{p-1}\,\mathbb{1}\Big\{V_d R^d\,\mathrm{M}g > k\Big\} g\,\mathrm{d}\lambda &\le \int (\mathrm{M}g)^{p-1}\Big(|g-f| + f\,\mathbb{1}\Big\{V_d R^d\,\mathrm{M}g \ge k\Big\}\Big)\,\mathrm{d}\lambda \\
&\le \|\mathrm{M}g\|_p^{p-1}\left(\delta_p + \left(\int f^p\,\mathbb{1}\Big\{V_d R^d\,\mathrm{M}g > k\Big\}\,\mathrm{d}\lambda\right)^{\frac{1}{p}}\right).
\end{aligned}$$

From the strong type Hardy-Littlewood maximal inequality, there exists a constant $C_p < \infty$ depending on p and d such that $\|\mathrm{M}g\|_p \le C_p\|g\|_p \le C_p(\|f\|_p + \delta_p)$. Likewise, because $\|g\|_1 = 1$, from the weak type Hardy-Littlewood maximal inequality, there is a constant $C_1 < \infty$ depending only on d such that $\int \mathbb{1}\{V_d R^d\,\mathrm{M}g > k\}\,\mathrm{d}\lambda \le C_1 V_d R^d/k$. Thus because $\|f\|_p < \infty$, Proposition B.2 gives that $\int f^p\,\mathbb{1}\{V_d R^d\,\mathrm{M}g > k\}\,\mathrm{d}\lambda \le \zeta_p(C_1 V_d R^d/k)$ for some function $\zeta_p\colon \mathbb{R}_+ \to \mathbb{R}_+$ depending only on f and p such that $\lim_{\alpha\to 0}\zeta_p(\alpha) = 0$.

Now consider the second term in (3):

$$\begin{aligned}
&\int \sum_{j=0}^{\infty} j^{p-1}p_j\Big(V_d R^d\,\mathrm{M}g\Big)\,\mathbb{1}\{j \ge k\} g\,\mathrm{d}\lambda \\
&\le \int \sum_{j=0}^{\infty} j^{p-1}p_j\Big(V_d R^d\,\mathrm{M}g\Big)\big(|g-f| + f\,\mathbb{1}\{j > k\}\big)\,\mathrm{d}\lambda.
\end{aligned}$$

Separating, we have

$$\int \sum_{j=0}^{\infty} j^{p-1}p_j\Big(V_d R^d\,\mathrm{M}g\Big)|g-f|\,\mathrm{d}\lambda$$

$$\le \left(K\frac{p-1}{\log(p-1)}\right)^{p-1}\int\left(V_dR^d\,\mathrm{M}g+\left(V_dR^d\,\mathrm{M}g\right)^{p-1}\right)|g-f|\,\mathrm{d}\lambda$$

$$\le \left(K\frac{p-1}{\log(p-1)}\right)^{p-1}\left(V_dR^dC_2\left\|g\right\|_2\delta_2+\left(V_dR^dC_p\left\|g\right\|_p\right)^{p-1}\delta_p\right)$$

$$\le \left(K\frac{p-1}{\log(p-1)}\right)^{p-1}\left(V_dR^dC_2(\|f\|_2+\delta_2)\delta_2+\left(V_dR^dC_p\left(\|f\|_p+\delta_p\right)\right)^{p-1}\delta_p\right).$$

Because the Poisson distribution is divisible,

$$\int\sum_{j=0}^{\infty}j^{p-1}p_j\left(V_dR^d\,\mathrm{M}g\right)\mathbb{1}\{j>k\}f\,\mathrm{d}\lambda$$

$$\le 2^{p-2}\left(\int\sum_{j=0}^{\infty}j^{p-1}p_j\left(V_dR^d\,\mathrm{M}|g-f|\right)f\,\mathrm{d}\lambda\right.$$

$$\left.+\int\sum_{i=0}^{\infty}\sum_{j=0}^{\infty}j^{p-1}p_i\left(V_dR^d\,\mathrm{M}|g-f|\right)p_j\left(V_dR^d\,\mathrm{M}f\right)\mathbb{1}\{i+j>k\}f\,\mathrm{d}\lambda\right).$$

We have

$$\int\sum_{j=0}^{\infty}j^{p-1}p_j\left(V_dR^d\,\mathrm{M}|g-f|\right)f\,\mathrm{d}\lambda$$

$$\le \left(K\frac{p-1}{\log(p-1)}\right)^{p-1}\left(V_dR^dC_2\delta_2\left\|f\right\|_2+\left(V_dR^dC_p\delta_p\right)^{p-1}\left\|f\right\|_p\right).$$

From Markov's inequality

$$\int\sum_{i=0}^{\infty}\sum_{j=0}^{\infty}p_i\left(V_dR^d\,\mathrm{M}|g-f|\right)p_j\left(V_dR^d\,\mathrm{M}f\right)\mathbb{1}\{i+j>k\}f\,\mathrm{d}\lambda$$

$$\le \frac{1}{k}\left(\int\sum_{j=0}^{\infty}j\left(p_j\left(V_dR^d\,\mathrm{M}|g-f|\right)+p_j\left(V_dR^d\,\mathrm{M}f\right)\right)f\,\mathrm{d}\lambda\right)$$

$$\le \frac{V_dR^dC_2}{k}(\delta_2+\|f\|_2)\left\|f\right\|_2.$$

Likewise

$$\int\sum_{j=0}^{\infty}j^{p-1}p_j\left(V_dR^d\,\mathrm{M}f\right)f\,\mathrm{d}\lambda$$

$$\le \left(K\frac{p-1}{\log(p-1)}\right)^{p-1}\left(V_dR^dC_2\left\|f\right\|_2^2+\left(V_dR^dC_p\right)^{p-1}\left\|f\right\|_p^p\right).$$

Then, via Proposition B.2 there exists an increasing function $\zeta\colon\mathbb{R}_+\to\mathbb{R}_+$ depending only on f with $\lim_{\alpha\to0}\zeta(\alpha)=0$ and

$$\int\sum_{i=0}^{\infty}\sum_{j=0}^{\infty}j^{p-1}p_i\left(V_dR^d\,\mathrm{M}|g-f|\right)p_j\left(V_dR^d\,\mathrm{M}f\right)\mathbb{1}\{i+j>k\}f\,\mathrm{d}\lambda$$

$$\le \zeta\left(\frac{V_dR^dC_2}{k}(\delta_2+\|f\|_2)\left\|f\right\|_2\right)$$

Via Proposition B.1 there exist increasing functions $\xi_2\colon \mathbb{R}_+ \to \mathbb{R}_+$ and $\xi_p\colon \mathbb{R}_+ \to \mathbb{R}_+$ depending only on f and $\max\{p,2\}$ with $\lim_{\alpha\to 0}\xi_2(\alpha) = \lim_{\alpha\to 0}\xi_p(\alpha) = 0$ such that $\delta_2 \le \xi_2(\delta_{\max\{p,2\}})$ and $\delta_p \le \xi_p(\delta_{\max\{p,2\}})$, with one of ξ_2 or ξ_p being the identity function.

Combining all of the preceding pieces yields the required uniform bound for (2). As $\delta_{\max\{2,p\}} \to 0$, $k \to 0$ may be chosen such that the overall bound is made arbitrarily small. When $p = 2u+1$, we see that (E1) holds.

Likewise, for $p > 2u+1$ and any $K > 0$,

$$\begin{aligned}&\mathbb{E}\Big[\big|D_{\sqrt[d]{n}Y'}\big(\sqrt[d]{n}\mathbf{Y}_n\big)\big|^{\frac{p-1}{u}}\Big]\\ &\le K^{\frac{p-1}{u}} + \mathbb{E}\Big[\big|D_{\sqrt[d]{n}Y'}\big(\sqrt[d]{n}\mathbf{Y}_n\big)\big|^{\frac{p-1}{u}}\,\mathbb{1}\big\{\big|D_{\sqrt[d]{n}Y'}\big(\sqrt[d]{n}\mathbf{Y}_n\big)\big| > K\big\}\Big].\end{aligned}$$

As demonstrated, this quantity may be bounded uniformly in G and n. Thus, Statement 2.8 holds for $a = (p-1)/u$. □

LEMMA 2.10. *Let ψ satisfy* (S2). *Then if $\|f\|_2 < \infty$, ψ satisfies* (S1).

PROOF. Let $\{X_i\}_{i\in\mathbb{N}} \overset{\text{iid}}{\sim} F$ and $\{Y_i\}_{i\in\mathbb{N}} \overset{\text{iid}}{\sim} G$ with $X' \sim F$ and $Y' \sim G$ independent copies. Denote $\delta_1 := \|g-f\|_1$ and $\delta_2 := \|g-f\|_2$, where $f := \mathrm{d}F/\mathrm{d}\lambda$ and $g := \mathrm{d}G/\mathrm{d}\lambda$. Define $\mathbf{X}_n := \{X_i\}_{i=1}^n$ and $\mathbf{Y}_n := \{Y_i\}_{i=1}^n$. By Proposition B.1, there exists a continuous, increasing function $\xi\colon \mathbb{R}_+ \to \mathbb{R}_+$ depending only on f such that $\lim_{\alpha\to 0}\xi(\alpha) = 0$ and $\delta_1 \le \xi(\delta_2)$. It may be assumed that $\{(X_i, Y_i)\}_{i\in\mathbb{N}}$ are iid coupled random variables with $\mathbb{P}[X_i \ne Y_i] = \delta_1/2$ for all $i \in \mathbb{N}$.

For $L > 0$, define the following events:

$$\begin{aligned}A &:= \{Y' = X'\}\\ B &:= \left\{\mathbf{Y}_n \cap B_{X'}\left(\frac{L}{\sqrt[d]{n}}\right) = \mathbf{X}_n \cap B_{X'}\left(\frac{L}{\sqrt[d]{n}}\right)\right\}\\ C_* &:= \big\{\rho_{\sqrt[d]{n}X'}\big(\sqrt[d]{n}\mathbf{X}_n\big) \le L\big\}.\end{aligned}$$

Let $C \subseteq C_*$ be measurable such that $\mathbb{P}[C^c] = \mathbb{P}^*[C_*^c]$. By the local-definition criterion, Definition 2.6, $A \cap B \cap C \subseteq \{\rho_{\sqrt[d]{n}Y'}(\sqrt[d]{n}\mathbf{Y}_n) \le L\}$. Then $\mathbb{P}^*[\rho_{\sqrt[d]{n}Y'}(\sqrt[d]{n}\mathbf{Y}_n) > L] \le \mathbb{P}[A^c] + \mathbb{P}[B^c] + \mathbb{P}[C^c]$. Bounding each piece in turn, $\mathbb{P}[A^c] = \mathbb{P}[X' \ne Y'] \le \xi(\delta_2)/2$. Also, by (S2), we have a bound for $\mathbb{P}[C^c]$ which holds uniformly in n and can be made arbitrarily small as $L \to \infty$.

It remains to be shown that B^c occurs with small probability uniformly in n and G. As shown in the proof of Proposition 2.12, the probability that $\mathbf{X}_n$ and $\mathbf{Y}_n$ fail to coincide within $B_{X'}(L/\sqrt[d]{n})$ is at most $C_2 V_d L^d \delta_2 \|f\|_2$ for some constant $C_2 < \infty$. As $L^d\delta_2 \to 0$ we may choose $L \to \infty$ such that both our bounds for $\mathbb{P}[B^c]$ and $\mathbb{P}[C^c]$ become arbitrarily small. Combining, we see that (S1) is satisfied. □

B.4. Proofs of Section 4.3.

LEMMA 4.1. *Let $\|f\|_2 < \infty$ and $\mathcal{K}$ satisfy* (K2), (D2), *and* (D3). *Then $\beta_q^{r,s}(\mathcal{K})$ satisfies* (S2) *for any $r \in \mathbb{R}$, $s \in \mathbb{R}$, and $q \ge 0$.*

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 m ust 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 [5] 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 2.11 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 $\mathbb{P}[\rho^*_{\sqrt[d]{n}X'}(\sqrt[d]{n}\mathbf{X}_n)] \leq \delta$ 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 $\rho_{\sqrt[d]{n}X'}(\sqrt[d]{n}\mathbf{X}_n) \leq (n_\delta+1)\phi$ almost surely. For $n \geq n_\delta$, $\rho_{\sqrt[d]{n}X'}(\sqrt[d]{n}\mathbf{X}_n) \leq \rho^*_{\sqrt[d]{n}X'}(\sqrt[d]{n}\mathbf{X}_n) \leq L_\delta$ with probability at least $1-\delta$. Therefore
$\sup_{n\in\mathbb{N}} \mathbb{P}[\rho_{\sqrt[d]{n}X'}(\sqrt[d]{n}\mathbf{X}_n) > \max\{L_\delta, (n_\delta+1)\phi\}] \leq \delta$, and the result follows. □

B.5. Proofs of Section 4.4.

COROLLARY 4.3. *Let $q \geq 0$ and $p \geq 2q+3$. Let $\mathcal{K}$ satisfy* (K1), (K2), (D1), *and* (D3). *Then for any given $\vec{r}$, $\vec{s}$, Statement 4.2 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

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

$$
\begin{aligned}
&\leq \frac{1}{((q+1)!)^a}(I_n+1)^{a(q+1)} \\
&\leq \frac{2^{a(q+1)-1}}{((q+1)!)^a}\Big(I_n^{a(q+1)}+1\Big).
\end{aligned}
$$

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 2.9 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 2.10. Finally, an application of Theorem 2.13 gives the desired result.

Referring to Proposition 2.12 and the proof thereof, for $p<\infty$, using $a=(p-1)/(q+1)$ we achieve an optimal rate for γ_ϵ of

$$
O\bigg(\delta_\epsilon^{1-\frac{2q+2}{p-1}}\bigg).
$$

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

$$
O\Bigg(\delta_\epsilon\bigg(\frac{-\log(\delta_\epsilon)}{\log(-\log(\delta_\epsilon))}\bigg)^{2q+2}\Bigg).
$$

Both of these rates depend on δ_ϵ, the upper bound for the total probability found in the proof of Proposition 2.12. The techniques found in the proofs of Lemma 2.9 and Proposition 2.12 allow for a bound on δ_ϵ, provided a tail bound for $\sup_{n\in\mathbb{N}}\rho_0(\sqrt[d]{n}(\mathbf{Y}_n-Y'))$. At this time, such a bound is unavailable, thus no explicit rate calculation is possible. □

COROLLARY 4.4. *Let $q\geq 0$ and $p\geq 2q+5$. Let $\mathcal{K}$ satisfy* (K2), (D2), *and* (D3). *Then for any given $\vec{r}$, $\vec{s}$, Statement 4.2 holds for $\beta_q^{\vec{r},\vec{s}}$.*

PROOF. The proof follows exactly that of Corollary 4.3, 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

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

$$\le \frac{2^{q+3}}{(q+2)!}(I_n+1)^{q+2}.$$

Thus for any $a > 2$,

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

(E2) is satisfied for $R = \phi$, $U_a = (2^{(a+1)(q+2)})/((q+2)!)^a$, and $u_a = a(q+2)$. Thus for $p \ge a(q+2)+1 > 2q+5$, (E1) follows by Lemma 2.9. (S2) and thus (S1) follow from Lemmas 4.1 and 2.10, respectively. An application of Theorem 2.13 gives the final result.

For the rate in Proposition 2.12, for $p < \infty$, using $a = (p-1)/(q+2)$ we achieve an optimal rate for γ_ϵ of

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

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

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

□

COROLLARY 4.5. *Let $m < \infty$ and $p \ge 2m+3$. Let $\mathcal{K}$ be a filtration of simplicial complexes satisfying* (K1), (K2), (D1), (D3), *and* (D4). *Then for any given $\vec{r}$, Statement 4.2 holds for $\chi^{\vec{r}}$.*

COROLLARY 4.6. *Let $m < \infty$ and $p \ge 2m+5$. Let $\mathcal{K}$ be a filtration of simplicial complexes satisfying* (K2), (D2), (D3), *and* (D4). *Then for any given $\vec{r}$, Statement 4.2 holds for $\chi^{\vec{r}}$.*

PROOF. We prove together Corollaries 4.5 and 4.6. 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 statement of the result, since Proposition 2.12 holds for the Betti numbers in dimensions $0 \le q \le m$ under the assumed conditions (see the proofs of Corollaries 4.3 and 4.4), then the same holds for their (alternating) sum, namely the Euler characteristic. The proof of Theorem 2.13 applies without alteration. □

COROLLARY 4.7. *Let $p > 2$. Furthermore, let $F \in \mathcal{D}_{\gamma,r_0}(C)$ and $\mathbb{1}\{\hat{F}_n \in \mathcal{D}_{\gamma,r_0}(C)\} \to 1$ in probability (resp. a.s.). Then Statement 4.2 holds for $l_{\mathrm{NN},k}$.*

PROOF. First, we will show that $\mathbb{E}[|l_{\mathrm{NN},k}(\sqrt[d]{n}(\mathbf{Y}_n \cup \{Y'\})) - l_{\mathrm{NN},k}(\sqrt[d]{n}\mathbf{Y}_n)|^a]$ is uniformly bounded for $G \in \mathcal{D}_{\gamma,r_0}(C)$ and $Y_1, \dots, Y_n, Y' \overset{\text{iid}}{\sim} G$. 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 $\#\{B_k\} \le C_{d,k}$, 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 $y_1, \dots, y_k$ 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 $y_1, \dots, y_k$ 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

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

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

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

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

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

We apply a bound similar to Theorem 7 [9]. 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

$$\mathbb{E}\big[R^a_{k+1,n}\big] \leq \left(\frac{k+1}{\gamma}\right)^{\frac{a}{d}} + \operatorname{diam}(C)n^{\frac{a}{d}}\Big(1 - \gamma r_0^{\frac{d}{a}}\Big)^n + \frac{a(e/(k+1))^{k+1}}{d(\gamma)^{\frac{a}{d}}}\int_{k+1}^{\infty} e^{-y}y^{k+\frac{a}{d}}\,\mathrm{d}y.$$

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 Statement 2.8, and thus (E1).

The required stabilization properties are first established for a unit-intensity homogeneous Poisson process via Lemma 6.1 [7]. 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 $\rho_0(\mathbb{P}_\lambda) = \rho_0(\mathbb{P}_1/\sqrt[d]{\lambda}) = \rho_0(\mathbb{P}_1)/\sqrt[d]{\lambda}$. Thus, $\mathbb{P}^*[\rho_0(\mathbb{P}_\lambda) > L] = \mathbb{P}^*[\rho_0(\mathbb{P}_1) > \sqrt[d]{\lambda}L]$. For any $\lambda > 1$, $\mathbb{P}^*[\rho_0(\mathbb{P}_\lambda) > L] \leq \mathbb{P}^*[\rho_0(\mathbb{P}_1) > L]$. Likewise, for any $\lambda_* < 1$, we may choose L_δ such that $\mathbb{P}^*[\rho_0(\mathbb{P}_1) > \sqrt[d]{\lambda_*}L_\delta] \leq \delta$. Then $\mathbb{P}[\rho_0(\mathbb{P}_\lambda) > L_\delta] \leq \delta$ for all $\lambda \in [\lambda_*, \infty)$. Stabilization then extends to the binomial sampling setting via Lemma 2.11 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 $\mathbb{P}^*[\rho_{\sqrt[d]{n}Y'}(\sqrt[d]{n}\mathbf{Y}_n) > L^*_\delta] \leq \delta$. Both quantities do not depend specifically on G.

When restricted to C, we have an absolute upper bound of $\operatorname{diam}(C)\sqrt[d]{n}$ for the radius of stabilization, as all points will fall inside of C almost surely. We set $L_\delta = \max\{\operatorname{diam}(C)\sqrt[d]{n_\delta}, L^*_\delta\}$. Then $\mathbb{P}^*[\rho_{\sqrt[d]{n}Y'}(\sqrt[d]{n}\mathbf{Y}_n) > L_\delta] \leq \delta$ 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 2.12 still apply. Likewise, the proof of Theorem 2.13 is easily altered to include the additional condition $\mathbb{1}\{\hat{F}_n \in D_{\gamma,r_0}(C)\} \to 1$. We omit details here. □

APPENDIX C: L_p CONSISTENCY OF KERNEL DENSITY ESTIMATORS

In this appendix 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 2.13.

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

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

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 $\lim_{n\to\infty} n^{p/(2d(p-1))} h_n = \infty$*

$$\left\|\hat{f}_{n,h_n} - f\right\|_p \overset{\mathrm{P}}{\to} 0$$

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

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

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.

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

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 [4]. We focus on the first term of (4).

$$\begin{aligned}\mathbb{E}\left[\int \left|\hat{f}_{n,h_n} - Q_{h_n} * f\right|^p \mathrm{d}\lambda\right] &= \int \mathbb{E}\left[\left|\hat{f}_{n,h_n} - Q_{h_n} * f\right|^p\right] \mathrm{d}\lambda \\ &= \frac{1}{n^p}\int \mathbb{E}\left[\left|\sum_{i=1}^{n} Y_i\right|^p\right] \mathrm{d}\lambda\end{aligned}$$

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 Rademacher random variables $\{e_i\}_{i\in\mathbb{N}}$, letting $Z_i(x) := e_i Y_i(x)$. We have that $\mathbb{E}[|\sum_{i=1}^n Y_i(x)|^p] \leq 2^p\,\mathbb{E}[|\sum_{i=1}^n Z_i(x)|^p]$. By Corollary 3 [6], there exists a universal constant $C < \infty$ such that, for any $j \in \mathbb{N}$

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

Then

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

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

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

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

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

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

As was shown earlier, the right hand side goes to 0, thus $\|\hat{f}_{n,h_n} - Q_{h_n} * f\|_p \xrightarrow{p} 0$. As $\|Q_{h_n} * f - f\|_p \to 0$, an application of Slutsky's theorem gives the final result. If $\sum_{n\in\mathbb{N}} 1/(n^{p/2} h_n^{d(p-1)}) < \infty$, the almost sure result follows from Borel-Cantelli. For any $\alpha > 1$, $h_n = \left(\log(n)^\alpha / n^{p/2-1}\right)^{1/d(p-1)}$ 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. $\mathrm{N}\,(\mu, \sigma^2)$ denotes the normal distribution with mean μ and variance σ^2, and $\mathrm{Exp}\,(\lambda)$ is the exponential distribution with rate parameter λ. $\mathrm{Cauchy}(\lambda)$ denotes the Cauchy distribution with scale parameter λ, and $(\cdot, \cdot)$ 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 4.3 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.

Generator 3:

```
$\theta \sim \mathrm{Unif}\left(\mathbb{S}^1\right)$
$X_1, X_2 \overset{\mathrm{iid}}{\sim} \mathrm{N}\left(0, .04\right)$
    **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:

```
$\theta \sim \mathrm{Unif}\left(B_0(1)\right)$
$X_1, X_2, X_3 \overset{\mathrm{iid}}{\sim} \mathrm{N}\left(0, .01\right)$
    **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:

```
$X \sim \mathrm{Unif}\left(\left\{\begin{matrix}(0.38741799, 0.24263535, 0.09535272)\\(0.25147839, 0.63824409, 0.62425101)\\(0.73988542, 0.80749034, 0.84972394)\\(0.26811913, 0.35911205, 0.08316547)\\(0.65954757, 0.04704809, 0.02113341)\end{matrix}\right\}\right)$
$Y_1, Y_2, Y_3 \overset{\mathrm{iid}}{\sim} \mathrm{Exp}\,(25)$
    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:

```
$\theta \sim \mathrm{Unif}\,(\mathbb{S}^2)$
$Y_1, \ldots, Y_5 \overset{\mathrm{iid}}{\sim} \mathrm{Cauchy}\,(.1)$
    return $(\theta, 0, 0) + (Y_1, \ldots, Y_5)$
```

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:

```
$(\theta_1, \theta_2) \overset{\mathrm{iid}}{\sim} \mathrm{Unif}\,(\mathbb{S}^1)$
$S \sim \mathrm{Unif}\,(\{-1, 1\})$
$Y_1, \ldots, Y_{10} \overset{\mathrm{iid}}{\sim} \mathrm{N}\,(0, .04)$
    return $(\theta_1 + S, \theta_2, 0, \ldots, 0) + (Y_1, \ldots, Y_{10})$
```

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.

REFERENCES

[1] 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. https://doi.org/10.3847/1538-3881/aa7567

[2] Breiman, L., Meisel, W. and Purcell, E. (1977). Variable Kernel Estimates of Multivariate Densities. *Technometrics* **19** 135–144. https://doi.org/10.1080/00401706.1977.10489521

[3] 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. https://doi.org/10.1051/0004-6361/201116878

[4] Folland, G. B. (1999). *Real Analysis: Modern Techniques and Applications*. Wiley.

[5] Krebs, J. T. and Polonik, W. (2019). On the Asymptotic Normality of Persistent Betti Numbers. *arXiv preprint arXiv:1903.03280*.

[6] Latała, R. (1997). Estimation of Moments of Sums of Independent Real Random Variables. *Ann. Probab.* **25** 1502–1513. https://doi.org/10.1214/aop/1024404522 MR1457628

[7] Penrose, M. D. and Yukich, J. E. (2001). Central Limit Theorems for Some Graphs in Computational Geometry. *Ann. Appl. Probab.* **11** 1005–1041. https://doi.org/10.1214/aoap/1015345393 MR1878288

[8] Roycraft, B. (2021). github.com/btroycraft/stabilizing_statistics_bootstrap. https://doi.org/10.5281/ZENODO.4627098

[9] Singh, S. and Póczos, B. (2016). Analysis of k-Nearest Neighbor Distances with Application to Entropy Estimation. *arXiv preprint arXiv:1603.08578*.