Supplement to: Subsampling Methods for Persistent Homology

A. Technical results

In this section, we present some technical results that will be used to prove the main theorems. First, we expand the notation introduced in the body of the paper (Section 3). For any positive integer m, let $\phi_m : \mathbb{X}^m \to \mathcal{D}_T$ be the diagram and $\psi_m : \mathcal{D}_T \to \mathcal{L}_T$ the landscape, i.e. $\phi_m(X) = D_X$, for any $X = \{x_1, \cdots, x_m\} \subset \mathbb{X}$ and $\psi(D_X) = \lambda_{D_X} = \lambda_X$, for any $D_X \in \mathcal{D}_T$. Given $\mu \in \mathcal{P}(\mathbb{X})$, we denote by Φ_μ^m the push-forward measure of $\mu^{\otimes m}$ by ϕ_m, that is $\Phi_\mu^m = (\phi_m)_*\mu$. Similarly, we denote by Ψ_μ^m the push-forward (induced) measure of $\mu^{\otimes m}$ by $\psi \circ \phi_m$, that is $\Psi_\mu^m = (\psi \circ \phi_m)_*\mu$.

For a fixed integer $m > 0$, consider the metric space $(\mathbb{X}, \rho)$ and the space $\mathbb{X}^m$ endowed with a metric ρ_m. We impose two conditions on ρ_m:

(C1) Given a real number $p \geq 1$, for any $X = \{x_1, \ldots, x_m\} \subset \mathbb{X}$ and $Y = \{y_1, \ldots, y_m\} \subset \mathbb{X}$,

$$\rho_m(X, Y) \leq \left(\sum_{i=1}^m \rho(x_i, y_i)^p \right)^{\frac{1}{p}}, \tag{9}$$

(C2) For any $X = \{x_1, \ldots, x_m\} \subset \mathbb{X}$ and $Y = \{y_1, \ldots, y_m\} \subset \mathbb{X}$,

$$H(X, Y) \leq \rho_m(X, Y). \tag{10}$$

Two examples of distance that satisfy conditions (C1) and (C2) are the Hausdorff distance and the L_p-distance $\rho_m(X, Y) = (\sum_{i=1}^m \rho(x_i, y_i)^p)^{\frac{1}{p}}$.

Lemma 15. *For any probability measures $\mu, \nu \in \mathcal{P}(\mathbb{X})$ and any metric $\rho_m : \mathbb{X}^m \times \mathbb{X}^m \to \mathbb{R}$ that satisfies (C1), we have*

$$W_{\rho_m,p}(\mu^{\otimes m}, \nu^{\otimes m}) \leq m^{\frac{1}{p}} W_{\rho,p}(\mu, \nu).$$

Remark: The bound of the above lemma is tight: it is an equality when μ is a Dirac measure and ν any other measure.

Proof. Let $\Pi \in \mathcal{P}(\mathbb{X} \times \mathbb{X})$ be a transport plan between μ and ν. Up to reordering the components of $\mathbb{X}^{2m}$, $\Pi^{\otimes m}$ is a transport plan between $\mu^{\otimes m}$ and $\nu^{\otimes m}$ whose p-cost is given by

$$\begin{aligned}
&\int_{\mathbb{X}^m \times \mathbb{X}^m} \rho_m(X, Y)^p d\Pi^{\otimes m}(X, Y) \\
&\leq \int_{\mathbb{X}^m \times \mathbb{X}^m} \sum_{i=1}^m \rho(x_i, y_i)^p \, d\Pi(x_1, y_1) \cdots d\Pi(x_m, y_m) \\
&= m \int_{\mathbb{X} \times \mathbb{X}} \rho(x_1, y_1)^p d\Pi(x_1, y_1).
\end{aligned}$$

The lemma follows by taking the minimum over all transport plans on both sides of this inequality. □

Lemma 16. *For any probability measures $\mu, \nu \in \mathcal{P}(\mathbb{X})$ and any metric $\rho_m : \mathbb{X}^m \times \mathbb{X}^m \to \mathbb{R}$ that satisfies (C2), we have*

$$W_{d_b,p}\left(\Phi_\mu^m, \, \Phi_\nu^m\right) \leq 2 W_{\rho_m,p}(\mu^{\otimes m}, \nu^{\otimes m}).$$

Proof. This is a consequence of the stability theorem for persistence diagrams. Given $X, Y \subset \mathbb{X}^m$, define

$$\Lambda_m(X, Y) = (D_X, D_Y).$$

If $\Pi \in \mathcal{P}(\mathbb{X}^m \times \mathbb{X}^m)$ is a transport plan between $\mu^{\otimes m}$ and $\nu^{\otimes m}$ then $\Lambda_{m,*}\Pi$ is a transport plan between Φ_μ^m and Φ_ν^m. Its p-cost is given by

$$\begin{aligned}
&\int_{\mathcal{D}_T \times \mathcal{D}_T} d_b(D_X, D_Y)^p d\Lambda_{m,*}\Pi(D_X, D_Y) \\
&= \int_{\mathbb{X}^m \times \mathbb{X}^m} d_b(\phi_m(X), \phi_m(Y))^p d\Pi(X, Y) \\
&\leq 2 \int_{\mathbb{X}^m \times \mathbb{X}^m} H(X, Y)^p d\Pi(X, Y) \;\; \text{(stability theorem)} \\
&\leq 2 \int_{\mathbb{X}^m \times \mathbb{X}^m} \rho_m(X, Y)^p d\Pi(X, Y).
\end{aligned}$$

The lemma follows by taking the minimum over all transport plans on both sides of this inequality. □

Lemma 17. *Let μ and ν be two probability measures on $\mathbb{X}$. Let $\lambda_X \sim \Psi_\mu^m$ and $\lambda_Y \sim \Psi_\nu^m$. Then*

$$\left\| \mathbb{E}_{\Psi_\mu^m}[\lambda_X] - \mathbb{E}_{\Psi_\nu^m}[\lambda_Y] \right\|_\infty \leq W_{d_b,p}\left(\Phi_\mu^m, \, \Phi_\nu^m\right).$$

Proof. Let Π be a transport plan between Φ_μ^m and Φ_ν^m. For any $t \in \mathbb{R}$ we have

$$\begin{aligned}
&\left| \mathbb{E}_{\Psi_\mu^m}[\lambda_X](t) - \mathbb{E}_{\Psi_\nu^m}[\lambda_Y](t) \right|^p \\
&= |\mathbb{E}[\lambda_X(t) - \lambda_Y(t)]|^p \\
&\leq \mathbb{E}\left[|\lambda_X(t) - \lambda_Y(t)|^p\right] \;\; \text{(Jensen inequality)} \\
&\leq \mathbb{E}\left[d_b(D_X, D_Y)^p\right] \;\; \text{(Stability of landscapes)} \\
&= \int_{\mathcal{D}_T \times \mathcal{D}_T} d_b(D_X, D_Y)^p d\Pi(D_X, D_Y) \\
&= C_p(\Pi)^p.
\end{aligned}$$

□

The following lemma is similar to Theorem 2 in Chazal et al. (2014c).

Lemma 18. *Let X be a sample of size m from a measure $\mu \in \mathcal{P}(\mathbb{X})$ that satisfies the (a, b, r_0)-standard assumption. Let $r_m = 2\left(\frac{\log m}{am}\right)^{1/b}$. Then*

$$\mathbb{E}\left[H(X, \mathbb{X}_\mu)\right] \leq r_0 + 2\left(\frac{\log m}{am}\right)^{1/b} \mathbb{1}_{(r_0,\infty)}(r_m) + \\ + 2\, C_1(a,b) \left(\frac{\log m}{am}\right)^{1/b} \frac{1}{(\log m)^2},$$

where $C_1(a, b)$ is a constant depending on a and b.

Proof. Let $r > r_0$. It can be proven that $q := \mathrm{Cv}\left(\mathbb{X}_\mu, r/2\right) \leq \frac{4^b}{ar^b} \vee 1$, where $\mathrm{Cv}(\mathbb{X}_\mu, 2r)$ denotes the number of balls of radius $r/2$ that are necessary to cover $\mathbb{X}_\mu$. Let $\mathcal{C} = \{x_1, \ldots, x_p\}$ be a set of centers such that $B(x_1, r/2), \ldots, B(x_p, r/2)$ is a covering of $\mathbb{X}_\mu$. Then,

$$\begin{aligned}
&\mathbb{P}\left(H(X, \mathbb{X}_\mu) > r\right) \\
&\leq \mathbb{P}\left(H(X, \mathcal{C}) + H(\mathcal{C}, \mathbb{X}_\mu) > r\right) \\
&\leq \mathbb{P}\left(H(X, \mathcal{C}) > r/2\right) \\
&\leq \mathbb{P}\left(\exists i \in \{1, \cdots, p\} \text{ such that } X \cap B(x_i, r/2) = \emptyset\right) \\
&\leq \sum_{i=1}^{p} \mathbb{P}\left(X \cap B(x_i, r/2) = \emptyset\right) \\
&\leq \frac{4^b}{ar^b}\left[1 - \inf_{i=1\ldots p} \mathbb{P}(B(x_i, r/2))\right]^m \\
&\leq \frac{4^b}{ar^b}\left[1 - \frac{ar^b}{2^b}\right]^m \\
&\leq \frac{4^b}{ar^b} \exp\left(-m\frac{a}{2^b} r^b\right).
\end{aligned}$$

Then

$$\begin{aligned}
\mathbb{E}\left[H(X, \mathbb{X}_\mu)\right] &= \int_{r>0} \mathbb{P}\left(H(X, \mathbb{X}_\mu) > r\right) dr \\
&\leq r_0 + \int_{r>r_0} \mathbb{P}\left(H(X, \mathbb{X}_\mu) > r\right) dr. \quad (11)
\end{aligned}$$

If $r_m \leq r_0$ then the last quantity in (11) is bounded by

$$r_0 + \int_{r>r_m} \mathbb{P}\left(H(X, \mathbb{X}_\mu) > r\right) dr,$$

otherwise (11) is bounded by

$$\begin{aligned}
&r_0 + \int_{r>0} \mathbb{P}\left(H(X, \mathbb{X}_\mu) > r\right) dr \\
&\leq r_0 + r_m + \int_{r>r_m} \mathbb{P}\left(H(X, \mathbb{X}_\mu) > r\right) dr.
\end{aligned}$$

In either case, we follow the strategy in Chazal et al. (2014c) to obtain the following bound:

$$\begin{aligned}
&\int_{r>r_m} \mathbb{P}\left(H(X, \mathbb{X}_\mu) > r\right) dr \\
&\leq 2\, C(a,b) \left(\frac{\log m}{am}\right)^{1/b} \frac{1}{(\log m)^2},
\end{aligned}$$

which implies that

$$\mathbb{E}\left[H(X, \mathbb{X}_\mu)\right] \leq r_0 + r_m \mathbb{1}_{(r_0,\infty)}(r_m) + \\ + 2\, C(a,b) \left(\frac{\log m}{am}\right)^{1/b} \frac{1}{(\log m)^2}.$$

□

B. Main Proofs

Proof of Theorem 5 It immediately follows from the three following inequalities of Lemmas 15, 16 and 17:

- upper bound on the Wasserstein distance between the tensor product of measures:
$$W_{\rho_m,p}(\mu^{\otimes m}, \nu^{\otimes m}) \leq m^{\frac{1}{p}} W_{\rho,p}(\mu, \nu)$$
- from measures on $\mathbb{X}^m$ to measures on $\mathcal{D}$:
$$W_{d_b,p}\left(\Phi_\mu^m,\ \Phi_\nu^m\right) \leq 2W_{\rho_m,p}(\mu^{\otimes m}, \nu^{\otimes m})$$
- from measures on $\mathcal{D}$ to difference of the expected landscapes:
$$\left\|\mathbb{E}_{\Psi_\mu^m}[\lambda_X] - \mathbb{E}_{\Psi_\nu^m}[\lambda_Y]\right\|_\infty \leq 2W_{d_b,p}\left(\Phi_\mu^m,\ \Phi_\nu^m\right)$$

□

Proof of Theorem 7

$$\begin{aligned}
&\|\mathbb{E}_{\Psi_\mu^m}(\lambda_X) - \mathbb{E}_{\Psi_\nu^m}(\lambda_Y)\|_\infty \\
&= \int_{\varepsilon>0} \mathbb{P}_{\Psi_\mu^m \otimes \Psi_\nu^m}\left(\|\lambda_X - \lambda_Y\|_\infty > \varepsilon\right) d\varepsilon \\
&= \varepsilon_0 + \int_{\varepsilon>\varepsilon_0} \mathbb{P}_{\Psi_\mu^m \otimes \Psi_\nu^m}\left(\|\lambda_X - \lambda_Y\|_\infty > \varepsilon\right) d\varepsilon. \quad (12)
\end{aligned}$$

The event $\{\|\lambda_X - \lambda_Y\|_\infty > \varepsilon\}$ inside the integral implies that

$$\frac{\varepsilon_0}{2} \leq \frac{\varepsilon}{2} < H(X, Y) \leq H(X, \mathbb{X}_\mu) + H(\mathbb{X}_\mu, \mathbb{X}_\nu) + H(Y, \mathbb{X}_\nu), \quad (13)$$

where X and Y are two samples of m points from μ and ν, respectively. Let $\varepsilon_0 = 2H(\mathbb{X}_\mu, \mathbb{X}_\nu)$. By (13), it follows that at least one of the following conditions holds:

$$H(X, \mathbb{X}_\mu) \geq \frac{\varepsilon - \varepsilon_0}{4},$$
$$H(Y, \mathbb{X}_\nu) \geq \frac{\varepsilon - \varepsilon_0}{4}.$$

We assume that the first condition holds (the other case follows similarly). Then the last quantity in equation (12) can be bounded by

$$\begin{aligned}
&\varepsilon_0 + \int_{\varepsilon>\varepsilon_0} \mathbb{P}\left(H(X,\mathbb{X}_\mu) \geq \frac{\varepsilon-\varepsilon_0}{4}\right) d\varepsilon \\
=&2H(\mathbb{X}_\mu,\mathbb{X}_\nu) + 4\int_{u>0} \mathbb{P}\left(H(X,\mathbb{X}_\mu)\geq u\right) du \\
=&2H(\mathbb{X}_\mu,\mathbb{X}_\nu) + 4\mathbb{E}\left[H(X,\mathbb{X}_\mu)\right] \\
\leq&2H(\mathbb{X}_\mu,\mathbb{X}_\nu) + 4r_0 \; + 8\left(\frac{\log m}{am}\right)^{1/b} \mathbb{1}_{(r_0,\infty)}(r_m) + \\
&+ 8\,C_1(a,b)\left(\frac{\log m}{am}\right)^{1/b} \frac{1}{(\log m)^2},
\end{aligned}$$

where the last inequality follows from Lemma 18. □

Proof of Theorem 9 It follows directly from (8) and Lemma 18 . □

Proof of Theorem 10

$$\begin{aligned}
&\mathbb{E}\left[\|\lambda_{\mathbb{X}_\mu} - \widehat{\lambda_n^m}\|_\infty\right] \\
&\leq 2\mathbb{E}\left[H(\mathbb{X}_\mu,\widehat{C_n^m})\right] \\
&\leq 2\int_{r>0} \mathbb{P}\left(H(\mathbb{X}_\mu,\widehat{C_n^m}) > r\right) dr \\
&\leq 2r_0 + 2\int_{r>r_0} \left[\mathbb{P}\left(H(\mathbb{X}_\mu,S_1^m)>r\right)\right]^n dr \\
&\leq 2r_0 + 2\int_{r>r_0} \left[\frac{4^b}{ar^b}\exp\left(-m\frac{a}{2^b}r^b\right)\right]^n dr,
\end{aligned}$$

where the last inequality follows from Lemma 18. If $r_m \leq r_0$ then the last term is upper bounded by

$$2r_0 + 2\int_{r>r_m} \left[\frac{4^b}{ar^b}\exp\left(-m\frac{a}{2^b}r^b\right)\right]^n dr,$$

otherwise it is bounded by

$$2r_0 + 2r_m + 2\int_{r>r_m} \left[\frac{4^b}{ar^b}\exp\left(-m\frac{a}{2^b}r^b\right)\right]^n dr.$$

In either case,

$$\begin{aligned}
&\int_{r>r_m} \left[\frac{4^b}{ar^b}\exp\left(-m\frac{a}{2^b}r^b\right)\right]^n dr \\
&= 2\frac{2^{bn}}{b}(ma)^{-1/b}m^n \int_{u>\log m} u^{1/b-n-1}\exp(-nu)du \\
&\leq 2C_2(a,b)\left(\frac{\log(2^b m)}{am}\right)^{1/b} \frac{1}{n\left[\log(2^b m)\right]^{n+1}},
\end{aligned}$$

where in the last inequality we applied the same strategy used to prove Theorem 2 in Chazal et al. (2014c). □

C. About the (a,b,r_0)-standard assumption

The aim of this section is to explain why the (a,b,r_0)-standard assumption is relevant, in particular when μ is a discrete measure. Our argument is related to weighted empirical processes, which have been studied in details by Alexander; see Alexander (1985; 1987b;a). A new look on this problem has been proposed more recently in Giné & Koltchinskii (2006); Giné et al. (2003) by using Talagrand concentration inequalities. The following result from Alexander (1985) will be sufficient here. Let $(\mathbb{X},\rho,\eta)$ be a measure metric space and let η_N be the empirical counterpart of η.

Proposition 19. *Let $\mathcal{C}$ be a VC class of measurable sets of index v of $\mathbb{X}$. Then for every δ, $\varepsilon > 0$ there exists K such that*

$$\begin{aligned}
&\eta\left[\sup\left\{\left|\tfrac{\eta_N(C)-\eta(C)}{\eta(C)}\right| \;:\; \eta(C) \geq Kv\tfrac{\log N}{N},\, C\in\mathcal{C}\right\} > \varepsilon\right] \\
&= O(N^{-(1+\delta)v}).
\end{aligned} \tag{14}$$

Assume that μ is the discrete uniform measure on a point cloud $X_N = \{x_1,\ldots,x_N\}$ which has been sampled from η, thus $\mu = \eta_N$. Assume moreover that η satisfies an (a',b)-standard assumption ($r_0 = 0$). Let r_0 be a positive function of N chosen further. For any $r > r_0(N)$ and any $y \in \mathbb{X}_\mu$:

$$\begin{aligned}
&\inf_{y\in\mathbb{X}_\mu} \mu(B(y,r)) \\
&= \inf_{y\in\mathbb{X}_\mu} \eta_N(B(y,r)) \\
&= \inf_{y\in\mathbb{X}_\mu} \left\{\eta(B(y,r))\left[1 - \frac{\eta(B(y,r)) - \eta_N(B(y,r))}{\eta(B(y,r))}\right]\right\} \\
&\geq (1\wedge a'r^b) \inf_{y\in\mathbb{X}_\mu} \left\{1 - \sup_{x\in\mathbb{X}}\left|\frac{\eta(B(x,r)) - \eta_N(B(x,r))}{\eta(B(x,r))}\right|\right\} \\
&\geq (1\wedge a'r^b) \inf_{y\in\mathbb{X}_\mu} \times \\
&\qquad \times \left\{1 - \sup_{x\in\mathbb{X}, r'\geq r_0(N)}\left|\frac{\eta(B(x,r')) - \eta_N(B(x,r'))}{\eta(B(x,r'))}\right|\right\}
\end{aligned} \tag{15}$$

Assume that the set of balls in $(\mathbb{X},\rho)$ has a finite VC-dimension v. For instance, in $\mathbb{R}^d$, the VC-dimension of balls is d+1. Under this assumption we apply Alexander's Proposition with (for instance) $\delta = 1$ and $\varepsilon = 1/2$. Let $K > 0$ such that (14) is satisfied. Then, by setting

$$r_0(N) := \left(\frac{Kv}{a'}\frac{\log N}{N}\right)^{1/b},$$

we finally obtain using (14) and (15) that

$$\eta\left[\inf_{y\in\mathbb{X}_\mu,\, r\geq r_N} \mu(B(y,r)) \geq 1\wedge\frac{a'}{2}r^b\right] = O(N^{-2v}).$$

In this quite general context, we see that by taking r_0 of the order of $\left(\frac{\log N}{N}\right)^{1/b}$, for large values of N the (a, b, r_0)-standard assumption is satisfied with high probability (in η).

D. Robustness to Outliers

The average landscape method is insensitive to outliers, as can be seen by the stability result of Theorem 5. The probability mass of an outlier gives a minimal contribution to the Wasserstein distance on the right hand side of the inequality. For example, suppose that $X_N = \{x_1, \ldots, x_N\}$ is a random sample from the unit circle $\mathbb{S}^2$, and let $Y_N = X_N \backslash \{x_1\} \cup \{(0,0)\}$. See Figure 7. The landscapes λ_{X_N} and λ_{Y_N} are very different because of the presence of the outlier $(0,0)$. On the other hand, the average landscapes constructed by multiple subsamples of $m < N$ points from X_N and Y_N are close to each other. Formally, let μ be the discrete uniform measure that put mass $1/N$ on each point of X_N and similarly let ν be the discrete uniform measure on Y_N. The 1st Wasserstein distance between the two measure is 1/N and, according to Theorem 5, the difference between the average landscapes is $\left\| \mathbb{E}_{\Psi_\mu^m}[\lambda_X] - \mathbb{E}_{\Psi_\nu^m}[\lambda_Y] \right\|_\infty \leq 2\frac{m}{N}$.

More formally, we can show that the average landscape $\overline{\lambda_n^m}$ can be more accurate than the closest subsample method when there are outliers. In fact, $\overline{\lambda_n^m}$ can even be more accurate than the landscape corresponding to a large sample of N points.

Suppose that the large, given point cloud $X_N = \{x_1, \ldots, x_N\}$ has a small fraction of outliers. Specifically, $X_N = \mathcal{G} \bigcup \mathcal{B}$ where $\mathcal{G} = \{x_1, \ldots, x_G\}$ are the good observations and $\mathcal{B} = \{y_1, \ldots, y_B\}$ are the outliers (bad observations). Let $G = |\mathcal{G}|$, $B = |\mathcal{B}|$ so that $N = G + B$ and let $\epsilon = B/N$ which we assume is small but positive. Our target is the landscape based on the non-outliers, namely, $\lambda_{\mathcal{G}}$. The presence of the outliers means that $\lambda_{X_N} \neq \lambda_G$. Let $\beta = \inf_S ||\lambda_S - \lambda_{\mathcal{G}}||_\infty > \delta$, for some $\delta > 0$, where the infimum is over all subsets that contain at least one outlier. Thus, β denotes the minimal bias due to the outliers. We consider three estimators:
λ_{X_N}: landscape from full given sample X_N;
$\overline{\lambda_n^m}$: average landscape from n subsamples of size m;
$\widehat{\lambda_n^m}$: landscape of closest subsample, from n subsamples of size m.

The last two estimators are defined in Section 3, and are constructed using n independent samples of size m from the discrete uniform measure that puts mass $1/N$ on each point of X_N.

Proposition 20. *If $\epsilon = o(1/n)$, then, for large enough n and m,*

$$\mathbb{E}\|\overline{\lambda_n^m} - \lambda_{\mathcal{G}}\|_\infty < \mathbb{E}\|\lambda_{X_N} - \lambda_{\mathcal{G}}\|_\infty. \tag{16}$$

In addition, if $nm\epsilon \to \infty$ then

$$\mathbb{P}\left[\mathbb{E}\|\overline{\lambda_n^m} - \lambda_{\mathcal{G}}\|_\infty < \|\widehat{\lambda_n^m} - \lambda_{\mathcal{G}}\|_\infty\right] \to 1. \tag{17}$$

Proof. We say that a subsample is clean if it contains no outliers and that it is infected if it contains at least one outlier. Let $S_1, \ldots, S_n$ be the subsamples of X_N of size m. Let $I = \{i : \ S_i \text{ is infected}\}$ and $C = \{i : \ S_i \text{ is clean}\}$. Then

$$\overline{\lambda_n^m} = \frac{n_0}{n}\lambda_0 + \frac{n_1}{n}\lambda_1$$

where n_0 is the number of clean subsamples, $n_1 = n - n_0$ is the number of infected subsamples, $\lambda_0 = (1/n_0)\sum_{i\in C}\lambda_{S_i}$ and $\lambda_1 = (1/n_1)\sum_{i\in I}\lambda_{S_i}$. Hence,

$$\begin{aligned}\|\overline{\lambda_n^m} - \lambda_{\mathcal{G}}\|_\infty &\leq \frac{n_0}{n}\|\lambda_0 - \lambda_{\mathcal{G}}\|_\infty + \frac{n_1}{n}\|\lambda_1 - \lambda_{\mathcal{G}}\|_\infty \\ &\leq \frac{n_0}{n}\|\lambda_0 - \lambda_{\mathcal{G}}\|_\infty + \frac{Tn_1}{2n}.\end{aligned}$$

A subsample is clean with probability $(1-\epsilon)^m$. Thus, $n_0 \sim \text{Binomial}(n, (1-\epsilon)^m)$ and $n_1 \sim \text{Binomial}(n, 1-(1-\epsilon)^m)$. Let $\pi = 1-(1-\epsilon)^m$. By Hoeffding's inequality

$$\begin{aligned}\mathbb{P}\left(\frac{Tn_1}{2n} > \frac{\beta}{2}\right) &= \mathbb{P}\left(\frac{Tn_1}{2n} - \frac{T\pi}{2} > \frac{\beta}{2} - \frac{T\pi}{2}\right) \\ &\leq \exp\left(-2n\left(\frac{\beta}{T} - \pi\right)^2\right).\end{aligned}$$

Since $\epsilon = o(1/n)$, we eventually have that

$$\pi = 1-(1-\epsilon)^m < \frac{\beta}{T} - \sqrt{\frac{\log n}{2n}},$$

which implies that $\mathbb{P}\left(\frac{Tn_1}{2n} > \frac{\beta}{2}\right) < 1/n$. So, except on a set of probability tending to 0,

$$\|\overline{\lambda_n^m} - \lambda_{\mathcal{G}}\|_\infty \leq \frac{n_0}{n}\|\lambda_0 - \lambda_{\mathcal{G}}\|_\infty + \frac{\beta}{2} \leq \|\lambda_0 - \lambda_{\mathcal{G}}\|_\infty + \frac{\beta}{2}$$

and thus, as soon as n, m and N are large enough,

$$\mathbb{E}\|\overline{\lambda_n^m} - \lambda_{\mathcal{G}}\|_\infty \leq \frac{\beta}{2} + \frac{\beta}{2} = \beta \leq \|\lambda_{X_N} - \lambda_{\mathcal{G}}\|_\infty.$$

This proves the first claim. To prove the second claim, note that the probability that at least one subsample is infected is $1-(1-\epsilon)^{nm} \sim 1-e^{-\epsilon nm} \to 1$. So with probability tending to one, there will be an infected subsample. This subsample will minimize $H(X, S_j)$ and the landscape based on this selected subsample will have a bias of order β. □

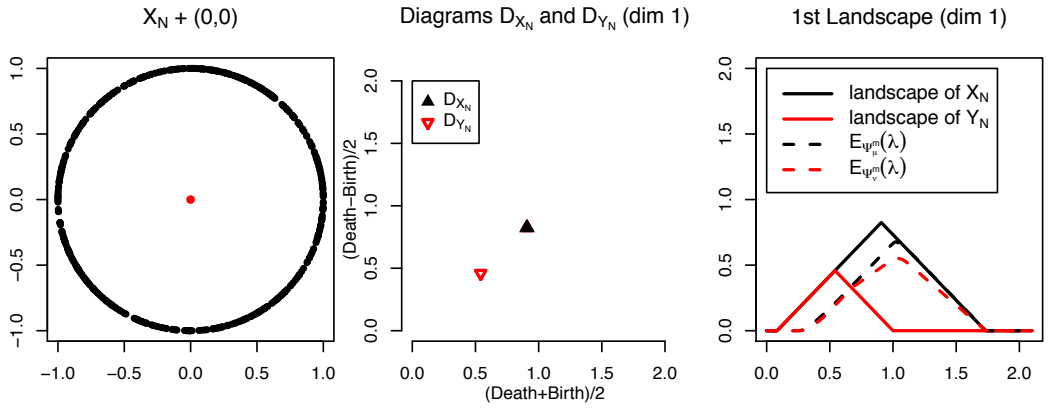

Figure 7. Left: X_N is the set of $N = 500$ points on the unit circle; $Y_N = X_N \backslash \{x_1\} \cup \{(0,0)\}$. Middle: persistence diagrams (dim 1) of the VR filtrations on X_N and Y_N, in the same plot, with different symbols. Right: landscapes of X_N, Y_N and the corresponding average landscapes constructed by subsampling $m = 100$ points from the two sets, for $n = 30$ times.

In practice, we can increase the robustness further, by using filtered subsampling. This can be done using the distance to the k-th nearest neighbor or using a kernel density estimator. For example, let

$$\widehat{p}_h(x) = \frac{1}{N}\sum_{j=1}^{N} K\left(\frac{||x - X_i||}{h}\right)$$

be a kernel density estimator with bandwidth h and kernel K. Suppose that all subsamples are chosen from the filtered set $\mathcal{F} = \{X_i : \widehat{p}_h(X_i) > t\}$. Suppose that the good observations $\mathcal{G}$ are sampled from a distribution on a set $A \subset [0,1]^d$ satisfying the (a,b)-standard condition with $b < d$, $a > 0$ and that $\mathcal{B}$ consists of B observations sampled from a uniform on $[0,1]^d$. For any $x \in A$,

$$\mathbb{E}[\widehat{p}_h(x)] \approx \frac{ah^b}{h^d}$$

and for any outlier X_i we have (for h small enough) that $\widehat{p}_h(X_i) = 1/(nh^d)$. Hence, if we choose t such that

$$\frac{1}{nh^d} < t < \frac{a}{h^{d-b}}$$

then $\mathcal{F} = \mathcal{G}$ with high probability.