A PROOFS OF STABILITY THEOREMS

Definition of diagram distances. Recall (see Section 3) that persistence diagrams are generally represented as multisets of points (i.e., points counted with multiplicity) supported on the upper half plane $\Omega = \{(b,d) \in \mathbb{R}^2, b > d\}$. Let $\mu = \{x_1, \ldots , x_n\}$ and $\nu = \{y_1, \ldots , y_m\}$ be two such diagrams and $s \geq 1$ be a parameter. Note in particular that $n \neq m$ in general. Let $\Delta = \{(t,t) \in \mathbb{R}\}$ denote the diagonal, and let $\Pi(\mu, \nu)$ denote the set of all bijections between $\mu \cup \Delta$ and $\nu \cup \Delta$. Then, the $s$-diagram distance between $\mu$ and $\nu$ is defined as:

$$d_s(\mu, \nu) = \inf_{\pi \in \Pi(\mu, \nu)} \left( \sum_{x \in \pi(\mu \cup \Delta)} \|x - s(x)\|^p \right)^{\frac{1}{p}}. \quad (6)$$

In particular, if $s = \infty$, we recover the bottleneck distance defined as:

$$d_B(\mu, \nu) = \inf_{\pi \in \Pi(\mu, \nu)} \sup_{x \in \pi(\mu \cup \Delta)} \|x - s(x)\|. \quad (7)$$

Proof of Theorem 2.2 The proof directly follows from the following two theorems. This first one, proved in [Hu et al., 2014], is a consequence of classical arguments from matrix perturbation theory.

**Theorem A.1** ([Hu et al., 2014], Theorem 1). Let $t \geq 0$ and let $L_w$ be the Laplacian matrix of a graph $G$ with $n$ vertices. Let $\lambda_1 < \cdots < \lambda_k$, $k \leq n$ be the distinct eigenvalues of $L_w$ and denote by $\delta > 0$ the smallest distance between two distinct eigenvalues: $\delta = \min_{j=1,\ldots,k-1} |\lambda_j + 1 - \lambda_j|$. Let $G'$ be another graph with $n$ vertices and Laplacian matrix $L_w = L_w + W$ with $\|W\| < \delta$, where $\|W\|$ denotes the Frobenius norm of $W$. Then, if $k = n$, there exists a constant $C_0(G,t) > 0$ such that for any vertex $v \in V$,

$$|h_{\mathcal{K}}(v) - h_{\mathcal{K}}(v)| \leq C_0(G,t)\|W\|;$$

if $k < n$, there exists two constants $C_1(G,t), C_2(G,t) > 0$ such that for any vertex $v \in V$,

$$|h_{\mathcal{K}}(v) - h_{\mathcal{K}}(v)| \leq C_1(G,t)\frac{\|W\|}{\delta - \|W\|} + C_2(G,t)\|W\|.$$

In particular, if $\|W\| < \frac{\delta}{2}$, there exists a constant $C(G,t) > 0$—notice that $\delta$ also depends on $G$—such that in the two above cases,

$$|h_{\mathcal{K}}(v) - h_{\mathcal{K}}(v)| \leq C(G,t)\|W\|.$$  

Theorem 2.2 then immediately follows from the second following theorem, which is a special case of general stability results for persistence diagrams.

**Theorem A.2** ([Chazal et al., 2016; Cohen-Steiner et al., 2009]). Let $G = (V,E)$ be a graph and $f,g : V \rightarrow \mathbb{R}$ be two functions defined on its vertices. Then:

$$d_B(Dg(G,f), Dg(G,g)) \leq \|f - g\|_\infty, \quad (8)$$

where $d_B$ stands for the so-called bottleneck distance between persistence diagrams and $\|f - g\|_\infty = \sup_{v \in G} |f(v) - g(v)|$. Moreover, this inequality is also satisfied for each of the subtypes Ord, Rel, Ext and Ext individually.

Proof of Theorem 2.3 Fix a graph $G = (V,E)$. With the same notations as in Section 2.2 recall that the eigenvalues of the normalized graph Laplacian satisfy $0 \leq \lambda_1 \leq \cdots \leq \lambda_n \leq 2$, and the corresponding eigenvectors $\{\psi_1, \ldots , \psi_n\}$ define an orthonormal family. In particular, $t \rightarrow \exp(-t\lambda_k)$ is 2-Lipschitz continuous for $t > 0$. Let $t, t'$ be two positive diffusion parameters. We have, for any $v \in V$:

$$\left| \sum_{k=1}^{n} (\exp(-t\lambda_k) - \exp(-t'\lambda_k))\psi_k(v)^2 \right| \leq 2 \cdot |t' - t| \sum_{k=1}^{n} \psi_k(v)^2.$$  

Thus in particular,

$$\sup_{v \in V} \left| \sum_{k=1}^{n} (h_{\mathcal{K}}(v) - h_{\mathcal{K}}(v)) \psi_k(v)^2 \right| \leq 2|t' - t|.$$  

As in the previous proof, we conclude using the stability of persistence diagrams w.r.t. the bottleneck distance (see Thm. A.2).

B DATASETS DESCRIPTION

Tables 3 and 4 summarize key information of each dataset for both our experiments. We also provide in Figure 5 an illustration of the weight grid we generated in Section 3.2.

C COMPLEMENTARY EXPERIMENTAL RESULTS

C.1 Weight learning

Figure 9 provides an illustration of the weight grid $w$ learned after training on the MUTAG dataset. Roughly speaking, activated cells highlight the areas of the plane where the presence of points was discriminating in the classification process. These learned grids thus emphasize the points of the persistence diagrams that matter w.r.t. learning task.
As mentioned in Section 2.2, the parameter $t$ sampled in log-scale is enough. Figure 6 illustrates the evolution of parameter $t$ over 40 epochs when trained on the MUTAG dataset (one epoch correspond to a stochastic gradient descent performed on the whole dataset). As one can see, parameter $t$ converges quickly. More importantly, it remains almost constant when initialized at $t_0 = 10.0$, suggesting that this choice is a (locally) optimal one. Fortunately, this is the parameter we use in our experiment (see Table 5). On the other hand, each time $t$ is updated (that is, at each epoch), one must recompute the diagrams for all the graphs in the training set, significantly increasing the running time of the algorithm.

### C.3 Experimental settings

Input data was fed to the network with mini-batches of size 128. For each dataset, various parameters are given (extended persistence diagrams, neural network architecture, optimizers, etc.) that were used to obtain the scores from Table 2. In Table 5, we use the following shortcuts:

- **Alpha:** persistence diagrams obtained with Gudhi’s $d$-dimensional AlphaComplex filtration.
- **hks:** extended persistence diagram obtained with HKS on the graph with parameter $t$.
- **prom($k$):** preprocessing step selecting the $k$ points that are the farthest away from the diagonal.
- **PERSLAY:** channel $\text{Im}(\phi, (a, b), q, \text{op})$ stands for a function $\phi$ obtained by using a Gaussian point
Mathieu Carrière, Frédéric Chazal, Yuichi Ike

Figure 5: Weight function $w$ when chosen to be a grid with size $20 \times 20$ before and after training (MUTAG dataset). Here, Ord0, Rel1, Ext0, and Ext1 denote the extended diagrams corresponding to downwards branches, upwards branches, connected components and loops respectively (cf Section 2.1).

Figure 6: Evolution of HKS parameter $t$ when considered as a trainable variable (i.e. differentiating $t \mapsto D_g(G, t)$ for all $G$) across 40 epochs for three different initializations of $t$, namely 0.1, 1 and 10, on the MUTAG dataset.

Figure 7: Evolution of $t \mapsto D_g(G, t)$ for one graph from the MUTAG dataset ($t \in [0.1, 100]$, $t$ in log-scale).

As our approach mix our topological features and some standard graph features, we provide two ablations studies.

C.4 Hyper-parameters influence

As our approach mix our topological features and some standard graph features, we provide two ablations studies, we provide two ablations studies, we provide two ablations studies, we provide two ablations studies, we provide two ablations studies. We have performed a grid search over the learning rate $\lambda$, the number of epochs $e$, and the size of the grid $p \times p$. We have also considered the use of a convolutional layer with filters of size $b \times b$, and a permutation equivariant function for dimension $d_2$. For the operation $\text{op}$, we have considered several options, such as the Adam optimizer (Kingma & Ba, 2014) with learning rate $\lambda$, using an Exponential Moving Average with decay rate $d$, and run during $e$ epochs.
Table 5: Settings used to generate our experimental results.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Train/Test acc (%)</th>
<th>Run time, CPU (s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>MUTAG</td>
<td>76.5/75.3</td>
<td>26.0</td>
</tr>
<tr>
<td>COLLAB</td>
<td>76.8/75.3</td>
<td>26.0</td>
</tr>
</tbody>
</table>

Table 6: Influence of hyper-parameters and ablation study. When varying a single hyper-parameter (e.g. grid size), all the others (e.g. perm op) are fixed to the values described in Supplementary Material, Table 5. Accuracies and running times are averaged over 100 runs (i5-8350U 1.70GHz CPU for the small MUTAG dataset, P100 GPU for the large COLLAB one). Bold-blue font refers to the experimental setting used in Section 4.

Similarly, we give in Table 6 the influence of the grid size that we choose as weight function $w$. In particular, we also perform an ablation study: grid size being None meaning that we enforce $w(p) = 1$ for all $p$. As expected, increasing the grid size improves train accuracy but leads to overfitting for too large values. However, this increase has only a small impact on running times whereas not using any grid significantly lowers it.

Finally, Figure 8 illustrates the variation of accuracy for both MUTAG and COLLAB datasets when varying the HKS parameter $t$ used when generating the extended persistence diagrams. One can see that the accuracy reached on MUTAG does not depend on the choice of $t$, which could intuitively be explained by the small size of the graphs in this dataset, making the $t$ parameter not very relevant. Experiments are performed on a single 10-fold, with 100 epochs. Parameters of PERSLAY are set to $\text{Im}(20,\cdot,20)$ for this experiment.

Table 7: Complementary report of experimental results.