2022-10-10
Progress Report

Pingyao Feng

9-24 Machine Learning Results for Classification

2	Tree	Accuracy (Validation): **79.2%**	
Last change: Optimizable Tree			10/10 features
6	Ensemble	Accuracy (Validation): 77.1%	
Last change: Optimizable Ensemble			10/10 features
1	Tree	Accuracy (Validation): 75.0%	
Last change: Fine Tree			10/10 features
5	KNN	Accuracy (Validation): 75.0%	
Last change: Optimizable KNN			10/10 features
8	Tree	Accuracy (Validation): 75.0%	
Last change: Medium Tree			10/10 features
3	Optimizable Discr…	Accuracy (Validation): 72.9%	
Last change: Optimizable Discriminant			10/10 features
4	SVM	Accuracy (Validation): 70.8%	
Last change: Optimizable SVM			10/10 features
7	Neural Network	Accuracy (Validation): 70.8%	
Last change: Optimizable Neural Network			10/10 features
9	KNN	Accuracy (Validation): 66.7%	
Last change: Hyperparameter option(s)			10/10 features

32 vowels, 16 consonants. 10 features: 5 are barcodes number of 5 diag, other 5 are number of barcodes that reaches inf(both consider barcode of 1 dimension for only)

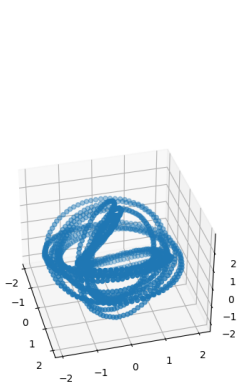
1	Tree	Accuracy (Validation): **81.5%**	
Last change: Fine Tree			4/4 features
2	Tree	Accuracy (Validation): **81.5%**	
Last change: Optimizable Tree			4/4 features
7	Tree	Accuracy (Validation): **81.5%**	
Last change: Medium Tree			4/4 features
4	Tree	Accuracy (Validation): 78.5%	
Last change: Coarse Tree			4/4 features
3	KNN	Accuracy (Validation): 69.2%	
Last change: Optimizable KNN			4/4 features
5	Neural Network	Accuracy (Validation): 46.2%	
Last change: Hyperparameter option(s)			4/4 features
6	Neural Network	Accuracy (Validation): 46.2%	
Last change: Narrow Neural Network			4/4 features

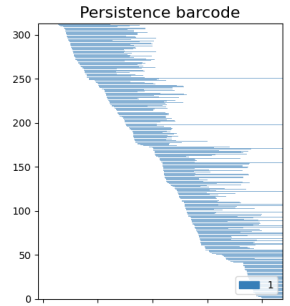
32 vowels, 33 consonants. 4 features: bottleneck distance between neighborhood barcode(currently the best result)

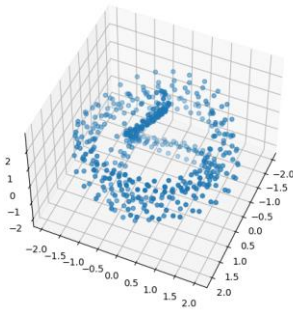
One of the reasons contributing to the bad result of classification is that the barcode is sensitive to some parameters of SW.

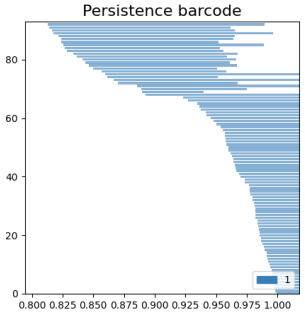
10-1 The influence of skip towards SW

I am surprised to see that ‘skip’ in SW seems to have crucial importance on persistent diagram: when skip minus one, the number of barcodes can sometimes change from hundreds to zero. ‘skip’(integer-valued) seems to complement max_edge_length(continue-valued), both perform somehow the same function. Unfortunately, here seems to be no article focusing on how to choose skip. (There are plenty focusing on dim and delay)

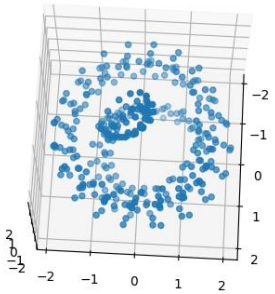


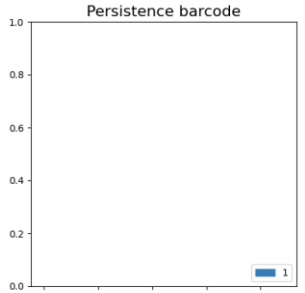


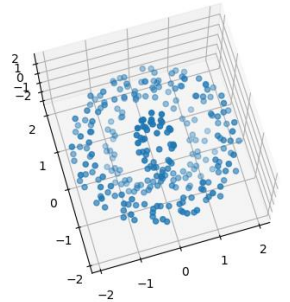




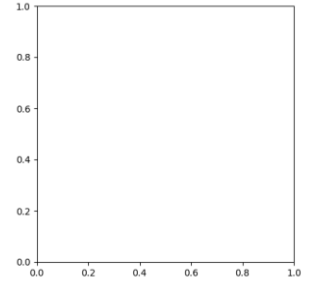

```
[(1, (0.9841401926107683, inf))]
```







```
[]
```

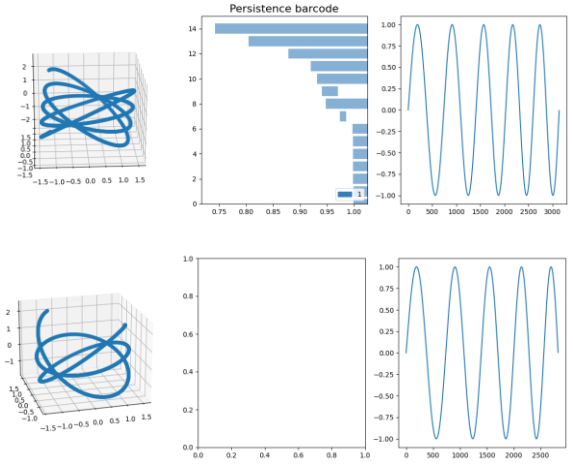


From left to right: the same data with just a different skip: 2,3,4,5 respectively. Careful with the difference between 4 and 5: Even if they seem to have the same per_diag: 4 has one barcode of dimension 1(but Gudhi can not draw it because it will result in singular transformation. Even if it can not be drawn, it’s a really good result), yet 5 has no barcode result in a runtime error. x axis may give a reason for this radical change. Here data is E:/phonetic/wav_file/vowel#32/Mid-central_vowel.wav, fraction 1, M=100, delay=4.

```
d:\python\lib\site-packages\gudhi\persistence_graphical_tools.py:210: UserWarning: Attempting to set identical
left == right == 0.9841401926107683 results in singular transformations; automatically expanding.
  axes.axis([axis_start, infinity, 0, ind])
```

10-5 Case study1: change of period

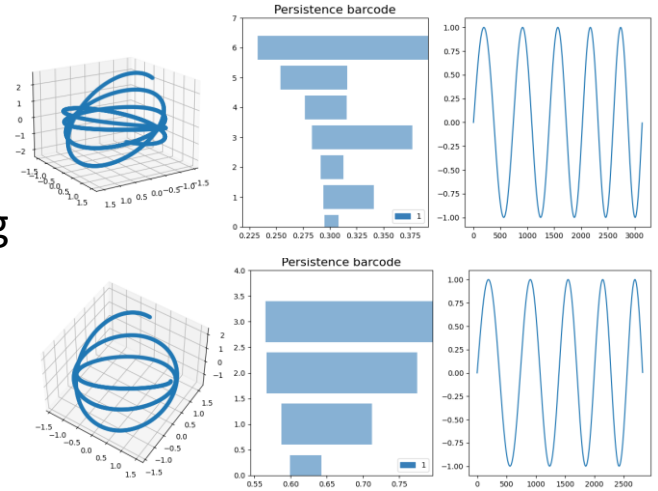
How will it affect SW and persistent diagram?



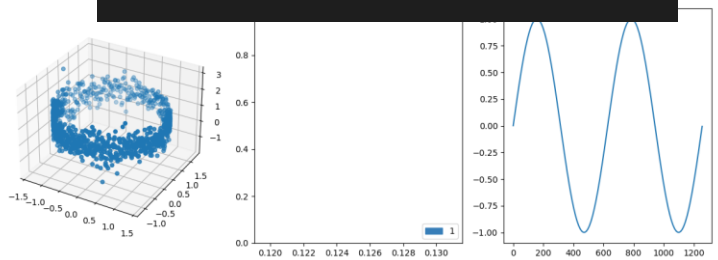
Above:[0.8, 1] equally apply to [0, 10pi], left in dim 100, right in dim 10. dim really need to be large? ;Below: [0.8,1] equally apply to [0, 9pi], left in dim 100, right in dim 10; Same thing happen when 11pi, 12pi

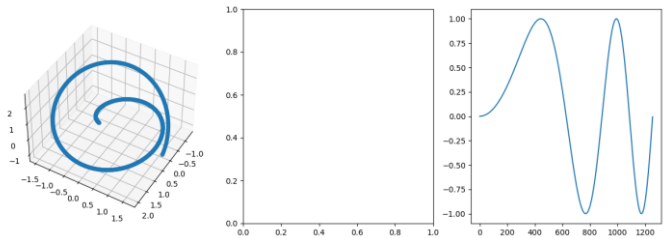
When dim is higher, it's faster to compute. Except for this, nothing is better for now to use high dim. The more experiment I do, the more confused I get about this parameter.

Q: if there is a loop after PCA, is it true that there is a loop in the original data? I think it's true because PCA just projects data to its principal component. Then why there is no barcode below on the left?



```
[(1, (0.12529346769429683, inf))]
```

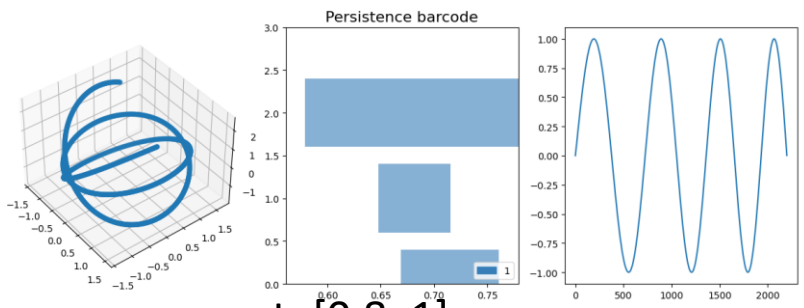




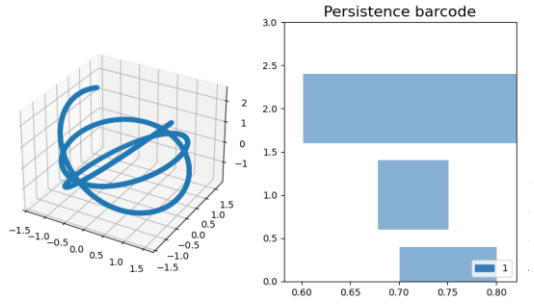
The really, really surprising thing I find about SW is that changing period makes SW seems more 'smooth'. I am confused about this, and wonder why this happen. It will affect 1-dim barcode. In the past, I supposed that the more smooth the curve is, the better the SE. Left: no period change. Right: [0,1] equally apply to data.

10-9 Continue: Change of period M=25

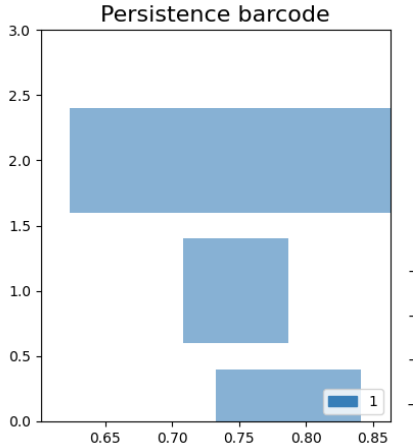
```
a=np.arange(0,7*np.pi,0.01)
t=np.linspace(0.8,1,len(a))
data_case1=np.sin(t*a)
```



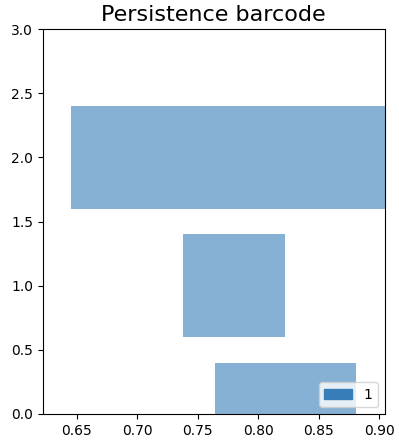
t=[0.8, 1]



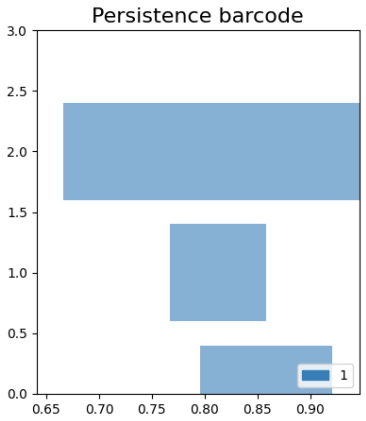
t= [0.79,1]



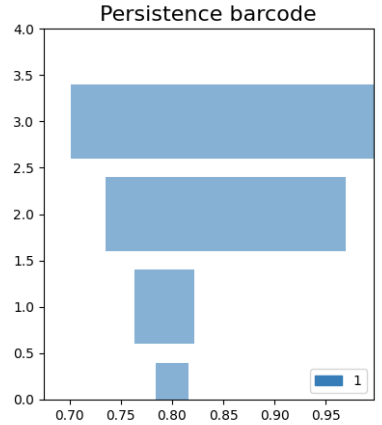
t= [0.78,1]



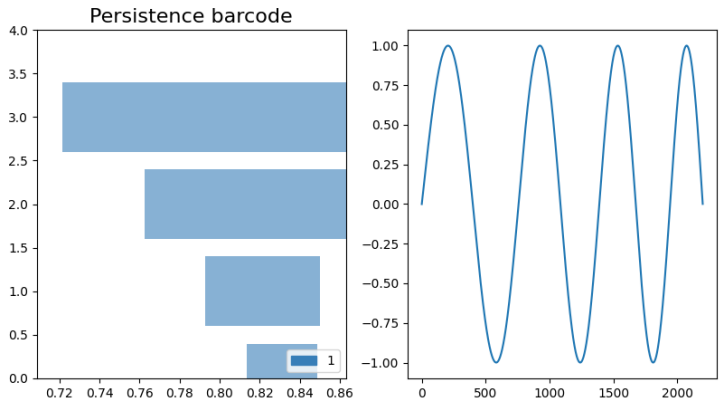
t=[0.77,1]



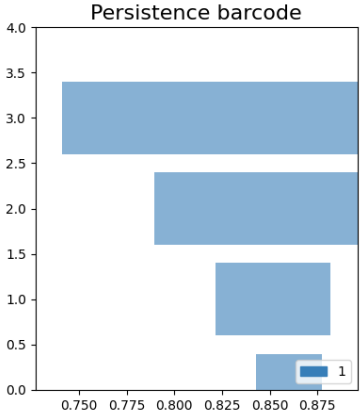
t=[0.76,1]



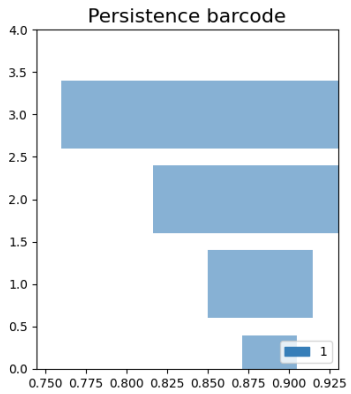
t=[0.75, 1]



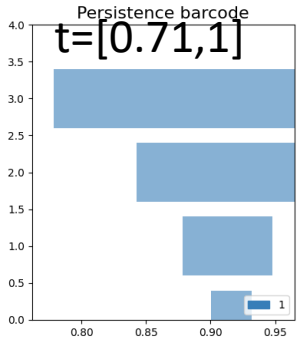
t=[0.74,1]



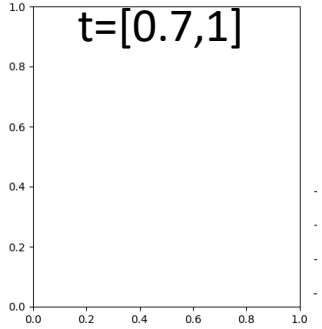
t=[0.73,1]



t=[0.72,1]



t=[0.71,1]



t=[0.7,1]

When the error becomes larger, the birth of barcode will go later, and so do the death (as expected). There will be some sudden occurrence of short barcode. The barcode, however, is quite sensitive to the error (what I do not expect). Try to describe it in a mathematical way and give an explanation?

Stubbornness & Future work

1. (10-1) Instead of trying to find a proper parametrization for classifying vowels and consonants (which is a tremendous challenge, I feel like tda can never achieve that), I will use classification as a case study to give a relatively comprehensive evaluation of tda. For example, using different complexes to do persistent diag, how will each parametrization influence the classification, and illustrating its strengths and drawbacks. I will try to do it in a mathematical way.

2. (10-2) Here're the important parameters I can think of: ①SE(dim, delay, skip). There are articles about dim and delay already. Skip is more subtle. ②different complex type. I've never tried this before. The currently used complex is rips complex. ③simplex_tree.persistence(min_persistence) and max_edge_length in building complex. The two parameters change continuously.

Relationship between SE and persistent diag will be the key point.

3. (10-10) Parameters for SE and persistent diagram are more sensitive than I expected, persistent diag seems to be no longer robust for now. (I almost begin to wonder if Gudhi performs in the right way) I will check on the previous articles about the robustness of persistent diagram, to see what kind of robustness it is and why I can not see any sign of robustness in this case study.(If I have time)