We thank all the reviewers for their valuable comments. We carefully address all the raised issues accordingly below.

1 Response to Reviewer 1: We appreciate your positive feedback.

2 Response to Reviewer 2:

Thank you for the comments. We will proofread and improve the readability. Please see details below.

2.1 Q: Definition of the forward pass for the main network $G_{x,w}$. How they go from x to z.

A: $G_{x,w}$ is the network parameterized by the kernel x and linear layers' weights w. Input is y, and output $z = G_{x,w}(y)$.

2.2 Q: In line 4 of Algorithm 1, it refers to $f_t(x, y)$ in Eq. (2), but it is not found there.

A: Sorry for the confusion. $f_t(x, y)$ is the same as $dtw^2(x, y)$ in Eq.2, with the subscript t being the iteration index.

2.3 Q: "Assume that $\tilde{x}$ is close to global minima". This assumption may never hold.

A: We shouldn't abuse the term "assumption". In fact, this "assumption" is not necessary for the analysis. We simply want to emphasize that we are interested in areas having many local minima, which happen to be always close to the global minimum from empirical observations. Note that the local-minima areas of interests can be anywhere, and the analysis still holds. The only issue arises when the left region of w's stationary point $x^{(w)*}$ is located to the right of region w. In this case, we can combine w and the center region u to form a new quasi-bowl center region $u' = [w, u]$ (similar to the analysis at the top of Page 6 in the paper), and the analysis still holds. We will remove this "assumption".

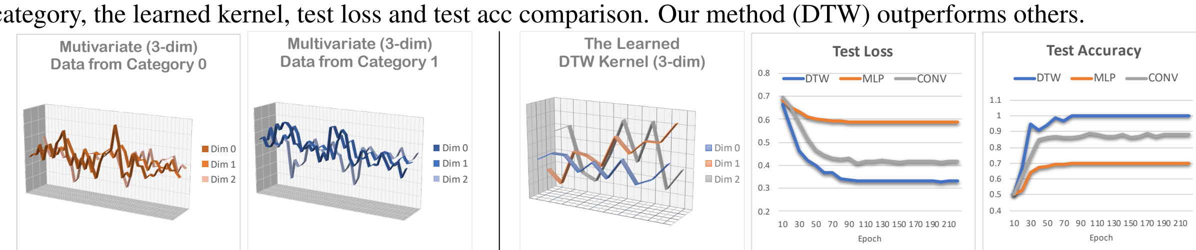
2.4 Q: Proof of convergence: after t iterations, it will be eps-close to the exact DTW.

Since it is highly non-convex and non-smooth, to the best of our knowledge, without strong assumptions it is unlikely to prove global convergence. We also analyze the escaping behavior (from local minima) but not the global convergence. The shape of DTW loss is identified through the alternating algorithm, which is one of the contributions in our paper.

2.5 Q: The idea will only work for the univariate time series.

A: The proposed approach works for both univariate and multivariate time series. Multivariate DTWs are often computed in two forms: MDTW-I and MDTW-D [1]. MDTW-I treats each dimension independently, so it is simply a stack of multiple univariate DTWs, thus directly applies to our method. MDTW-D needs to compute multivariate distance $||\mathbf{x_i} - \mathbf{y_j}||^2, \mathbf{x_i} \in \mathbf{R^m}, \mathbf{y_j} \in \mathbf{R^m}$ in the Dynamic Programming step, instead of the scaler version $||x_i - y_j||^2$. As long as the norm is well defined, e.g., Euclidean distance, the forward pass and the backpropagation are performed in the same manner. We can even define other distances, as long as their gradients w.r.t. to $\mathbf{x}$ can be computed.

We run a 3-dim multivariate time series classification task here, using MDTW-D and Euclidean distance in our approach. The experiment settings follow Section 6.1 in the paper. The following figures show: one sample data from each category, the learned kernel, test loss and test acc comparison. Our method (DTW) outperforms others.



(a) Data sample type-0 (b) Data sample type-1 (c) The learned kernel (d) Test Loss (e) Test Acc

2.6 Q: The experiments are insufficient. One dataset from UCR repo is not enough.

A: The UCR repo is a collection of a large variety of time series data, being the standard benchmark repo in related publications. We performed comprehensive experiments on 85 datasets from UCR repo (details provided in appendix of the paper), to make a fair comparison with Soft-dtw (they only use the UCR repo in their original paper as well). Additional experiments on synthetic data are also carried out. But we agree that more datasets should be considered in this area and we are exploring them in the final version and subsequent work.

2.7 Q: Some handwaving claims such as interpretability.

A: Some interpretability is revealed by the learned kernel's shape matching true patterns in the data (Fig 4, 5 in paper).

3 Response to Reviewer 3: Thank you for your comments and please see details below.

3.1 Q: For example, it will be helpful to know how to decide the number of DTW layers.

A: Empirically speaking, 1 or 2 layers of DTW are good enough. We observe no clear improvement with more than 2 layers, but the model complexity would be increased dramatically.

3.2 In alg 1's INPUT, kernels are initially set as input. But they are randomly initialised in OUTPUT part.

A: Kernels x and weights w are not inputs but model parameters being randomly initialized. We will clarify it.

3.3 Q: It will be helpful if the proposed method can be tested with more real datasets for the application part.

A: We agree with testing more datasets. Due to page limit, we randomly select the Haptics dataset from UCR repo as an expressive example in the application section, but more datasets will be considered (please also see response 2.6).

[1] Mohammad Shokoohi-Yekta, Bing Hu, Hongxia Jin, Jun Wang, and Eamonn Keogh. Generalizing dtw to the multi-dimensional case requires an adaptive approach. *Data mining and knowledge discovery*, 31(1):1–31, 2017.