Response to Reviewer #1

• *"I would like to see Theorem 4 reworded. It assumes that the underlying process has a correct clustering of states?"*

RE: You are right. Thm 4 assumes there is an underlying partition that attains the smallest value of distortion (eq(4)). The minimal distortion can be nonzero and the optimal solution may not be exactly correct. We will reword Thm 4 to make it easier to interpret.

• *"How to find state pairs in DQN analysis"* **RE:** We computed and ranked pairwise embedding distances for 2000 randomly picked states. Then we screened the top 100 closest pairs and pick those with large raw-data distances.

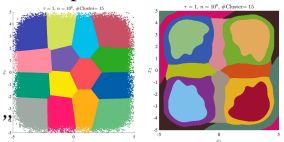
• *"... carefully written, theoretically/empirically supported ... The paper is already of a very high standard in my opinion ... clarity, readability.."* **RE:** Thanks! We are excited to have your support, and we will make an effort to improve readability of the paper.

Response to Reviewer #2

• *Relation with existing work on Koopman and dynamic model decomposition (DMD)*:

RE: Thank you for mentioning the related paper (Takeishi etal 18). We will cite it and add more discussion about DMD. DMD and our method share the same spirit - finding low-dimensional subspace of the transition function/kernel using decomposition. For comparison, DMD is originally developed for linear systems and then generalized to nonlinear dynamics. Our focus is different - we focus on the probabilistic transition of unstructured stochastic process and statistical error from finite dependent data. Our method applies to randomly jumping process (eg discrete-state games). We hope our analysis will inspire new developments on DMD.

• *"In Figure 2, experiment with clustering raw time series (without embedding) and compare."*
RE: Clustering raw time series produces partitions that look like grids and loses the temporal information (see the new Figure (*) for comparison). Figure 3 in the submission also gives a useful comparison and shows that embedding improves the clusters.



(*)15 clusters of states. Left: raw data; Right: after embeddings (Fig 2 in paper).

• *"Uniqueness of optimal solution required by Thm 4. What properties of the time-series are needed"*
RE: The uniqueness is a theoretical simplification. In practice, it means that the process admits a low-dim block structure or internal physical state, which is often unique.

• *Dependency on number of random features:* **RE:** According to (Rahimi&Recht 08), we pick the number of features to be $O(\frac{d}{\epsilon^2})$ to approximate the RKHS (ϵ is a user-picked accuracy level). Experiments suggest more than $O(\frac{d}{\epsilon^2})$ features doesn't add value.

Response to Reviewer #3

• *"Not written very well because it does not emphasize the intuition behind projecting to random features"*

RE: The results are based on projection to general RKHS space, not limited to random features. One needs random features only if the RKHS is known but explicit bases functions are unknown/infinite. In this case, randomizing features to approximate RKHS is a standard technique with rich theory developed by (Rahimi and Recht 2008).

• *"Explanation of the reversibility of normal diffusion maps and the idea here ... Not everyone knows diffusion map."*

RE: Thanks for the comment. We should have explained better, but were limited by the page limit.

- Irreversibility means we do *not* assume reversibility of the Markov process (a restricted technical condition), so the results apply to most practical time series. It also means that the frequency matrix/operator is asymmetric and doesn't admit eigendecomposition. Therefore typical analysis no longer applies, and proving the theorems requires nontrivial work.
- Diffusion map is a standard technique for dimension reduction (see Coifman etal 05, 06, Nedlar etal 09). Although it is related, it is not a prerequisite for understanding our paper. Our main idea is spectral decomposition of the transition kernel projected onto an RKHS space. We will make an effort to improve the writing and explanation.

• *"Potential for streaming the proposed algorithm"*

RE: Good point. It is possible to make the algorithm streaming (eg. using the Oja's method), but is beyond the current scope. (Our Algorithm 1 makes a single pass over time series. It runs in $O(d^2)$ space and $O(nd^2 + dr)$ total time, which is quite efficient.)

Response to Reviewer #4

• *Comparison to [1],[2],[3] (which were cited in our submission).*

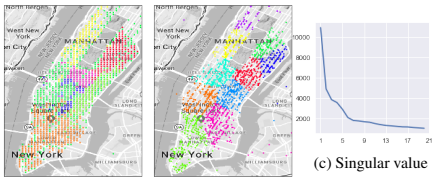
RE: [1] assumes reversibility of the underlying Markov chain and wavalet bases; its analysis relies on eigendocomposition and eigenvalue decay condition. In contrast, our method applies to practical time series (typically not reversible) and general RKHS, where eigendecomposition of P doesn't exist. [2,3] work for finite-state time series and lack scalability. Our method leverages kernel information and applies to multi-variate (or even unstructured) time series. Thm 1 generalizes [2,3]. Thms 2,3,4 provide theory for nonparametric estimation of metastable sets and preserving diffusion distance, which were not available in the literature.

• *"Assumption 2 is strong ... exactly represented by a finite sum of basis functions ... no approximation error ..."*

RE: The conditions are for simplicity and the analysis can be generalized to include approximation error. Computers operate in low dimensions. So we leverage (RR08) to approximate the kernel space and compute nonparametric estimation in finite dimensions.

• *"Why take the last layer of DQN? What is the benefit and will it improve the score? ... DQN is not stable. "*

RE: Very good questions. We should have better explained: Our DQN is pre-trained and only used as a static feature map for reducing the dimension of raw images. Our experiment can be viewed as using the composition between neural tangent kernel (Jacot etal 18) and Gaussian kernel. This experiment is to visualize and interpret state embeddings from game trajectories. Improving the score is beyond our scope and is an interesting direction for future work.



(a) 10 clusters by [2] (b) 10 embedding clusters (c) Singular value decay

• **"Additional experiments. Justify the low-rankness."** We further conduct a taxi-trip experiment (same setting as in [2]), using Gaussian kernel for pickup/dropoff locations. Figure (a) is the result from [2] without using kernel information; (b) is our clusters using kernelized state embeddings; (c) illustrates the singular value decay of transition matrix and justifies the low-rank assumption. In (a), the clusters are often mixed up and overlapping. In (b), state embedding leads to more meaningful zones and higher granularity, using the same data size. Due to the rebuttal's space limit, we can't explain every detail. We hope this additional experiment is convincing and we will add more details in the final paper.