On Rank Energy Statistics via Optimal Transport: Continuity, Convergence, and Change Point Detection

Matthew Werenski, Shoaib Bin Masud, *Member, IEEE*, James M. Murphy, *Member, IEEE*, Shuchin Aeron, *Senior Member, IEEE*

***Abstract*—This paper considers the use of recently proposed optimal transport-based multivariate goodness-of-fit (GoF) test statistics, namely rank energy and its variant the soft rank energy derived from entropy-regularized optimal transport, for unsupervised non-parametric change point detection (CPD) in multivariate time series data. We show that the soft rank energy enjoys both fast rates of statistical convergence and robust continuity properties which lead to strong performance on real datasets. Our analyses remove the need for resampling and out-of-sample extensions previously required to obtain such rates. Our theoretical results show that the rank energy suffers from the curse of dimensionality in statistical estimation and moreover can signal a change point from arbitrarily small perturbations, which leads to a high rate of false alarms in CPD. Additionally, under mild regularity conditions, we quantify the discrepancy between soft rank energy and rank energy in terms of the regularization parameter. Finally, we show our approach performs favorably in numerical experiments compared to several other optimal transport-based methods as well as maximum mean discrepancy (MMD), which is a popular multivariate GoF statistic.**

***Index Terms*—Optimal transport, goodness-of-fit, change point detection**

I. Introduction

The problem of detecting changes or transitions in multivariate time series data $(X_t) \subset \mathbb{R}^d$, referred to henceforth as *change point detection (CPD)*, is a central problem in a number of scientific domains [1]–[8]. The CPD problem amounts to partitioning the time series data into disjoint segments, with data in each consecutive segment being statistically distinct. In this context, we consider an *unsupervised* setting in which no prior examples of change points are made available and focus on non-parametric methods.

Motivated by recent developments in multivariate goodness-of-fit (GoF) tests based on notions of multivariate ranks derived from the theory of optimal transport (see [9] for a recent survey), we propose the use of the rank energy [10] and its numerically and sample efficient variant, the soft rank energy [11], for performing unsupervised non-parametric CPD. As noted in [10], [12] there are several advantages in considering rank energy and soft rank energy for GoF testing. First, rank energy is distribution-free under the null, a property which we numerically observe to be approximately also shared by the soft rank energy and that is lacking in other popular multivariate GoF measures such as the maximum mean discrepancy (MMD) [13], Wasserstein distances [14], and Sinkhorn divergences [15]. In the context of CPD, distribution-freeness potentially allows one to select a threshold for detection that is independent of the underlying distribution. Furthermore, statistical testing based on rank energy is shown to be robust to outliers and has better power for heavy tailed distributions [10]. Note while one can consider other optimal transport (OT) based rank GoF measures such as the rank MMD [10], Hotelling's-T^2 [12], and soft rank MMD [11], in this paper we focus on rank energy and soft rank energy and leave analogous development for these cases to future investigations.

We make the following fundamental contributions in this paper, keeping in view the practical utility of these tests towards robust CPD.

1) **Wasserstein Continuity Properties of GoF Statistics**: Theorems IV.1 and IV.2 provide novel analytic insights into rank energy and soft rank energy which explain their behaviors in practice. We show that the soft rank energy is Lipschitz with respect to the Wasserstein-1 metric while the rank energy fails to even be continuous. These properties translate to the smoothness (or lack thereof) of the GoF statistics in the proposed CPD algorithm with respect to small perturbations that are typical of real data.
2) **Convergence of Soft Rank Energy to Rank Energy**: In Theorem IV.4, under appropriate technical conditions, we provide an explicit convergence rate of the soft rank energy to the rank energy in terms of the regularization parameter. These results relax the conditions required to obtain convergence in existing work on the soft rank energy [11].
3) **Realistic and Fast Sample Convergence:** In Theorem IV.6 we establish the fast convergence rate of $n^{-1/2}$ of the plugin estimate of the soft rank energy to its population counterpart. Importantly, this result does not require using an out-of-sample extension as is required for the theoretical analysis in [11] thereby making most use of the limited samples and justifying the practical

The first two authors contributed equally.

Matthew Werenski is with the Department of Computer Science at Tufts University. Email: matthew.werenski@tufts.edu.

James M. Murphy is with the Department of Mathematics at Tufts University. Email: jm.murphy@tufts.edu.

Shoaib Bin Masud and Shuchin Aeron are with the Department of Electrical and Computer Engineering at Tufts University. Email: shoaib_bin.masud@tufts.edu, shuchin@ece.tufts.edu.

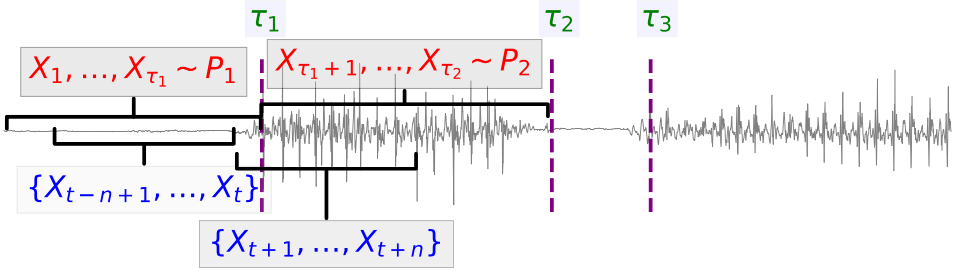
τ_1
τ_2
τ_3
$X_1, ..., X_{\tau_1} \sim P_1$
$X_{\tau_1+1}, ..., X_{\tau_2} \sim P_2$
$\{X_{t-n+1}, ..., X_t\}$
$\{X_{t+1}, ..., X_{t+n}\}$

Fig. 1. $(X_t)_{t=1}^T$ is time series data (gray) with several change points (dashed purple lines). The window of n samples between t and $t+n$ is $X_t, ..., X_{t+n}$.

implementation.

4) **Applications to CPD**: We numerically investigate and compare the performance of the soft rank energy with other OT-based and popular GoF statistics for CPD [5], [6], [16], [17]. Our results demonstrate the effectiveness of the soft rank energy in detecting change points on a variety of real-world datasets.

II. Overview of Change Point Detection

We begin by outlining the problem set-up and a brief overview of a class of methods that we focus on in this paper.

Data model: Consider a sequence of samples $(X_t) \subset \mathbb{R}^d, t = 1, 2, ...$ and assume that the sequence can be sequentially partitioned into disjoint segments $[\tau_{j-1}+1, ..., \tau_j]$ such that $X_{\tau_{j-1}+1}, ..., X_{\tau_j} \overset{i.i.d.}{\sim} P_j, j = 1, 2, ...$ Here $P_j \in \mathcal{P}(\mathbb{R}^d)$, the set of distributions on $\mathbb{R}^d$, are such that $P_j \neq P_{j+1}$ and the time indices τ_j are referred to as change points. The distributions (P_j) and the change points (τ_j) are arbitrary and not known in advance. We further consider a non-parametric setting in which the data generating distributions (P_j) are not assumed to belong to a parametric family of distributions.

Problem: Given the observed sequence (X_t), output a sequence of predicted change point indices $(\hat{\tau}_k)$ such that the sequence $(\hat{\tau}_k)$ is close to the true sequence (τ_j).

One method to estimate change points is the "sliding window" approach [5], [6], [18], [19] visualized in Figure 1 and outlined in Algorithm 1. For each time $t \in \{n, n+1, ..., T-n\}$ let z_t be a GoF statistic computed between the time-adjacent sets $\{X_{t-n+1}, ..., X_t\}$, and $\{X_{t+1}, ..., X_{t+n}\}$.

Repeating this for every time t creates a sequence (z_t) from which a sequence $(\hat{\tau}_i)$ of predicted change points can be extracted. For example, one can predict that $\hat{\tau}$ is a change point when $z_{\hat{\tau}}$ takes a large value or is a local maximizer within the sequence (z_τ). Since the predicted change points $(\hat{\tau}_k)$ are extracted from the sequence (z_τ) which is determined by a GoF statistic, the choice of GoF statistic will make a substantial difference in the quality of the predicted change points. The purpose of this paper is to argue theoretically and empirically that the soft rank energy (given in Definition III.6) is a strong choice of statistic, due to its favorable theoretical and computational properties.

Within the context of the sliding window approach we discuss several ways to quantify how close the sequence $(\hat{\tau}_k)$ is to (τ_j) which are made precise in Section V. Heuristically a sequence $(\hat{\tau}_k)$ is close to (τ_j) if it has the following two properties.

Algorithm 1: Sliding window-based CPD

```
Input  : Time series data $(X_t)_{t=1}^T$, window size $n$,
         threshold $\eta$, GoF statistic function GoFstat,
         peak search procedure PeakSearch
Output: Predicted change point sequence $(\hat{\tau}_k)$.
for $t = n, n+1, ..., T-n$ do
    $z_t =$ GoFstat$(\{X_{t-n+1}, ...X_t\}, \{X_{t+1}, ..., X_{t+n}\})$
end for
$(\hat{\tau}_k) =$ PeakSearch$((z_t), \eta)$
```

1) **(High True Change Point Detection Rate)** For most τ_j, there should be a point $\hat{\tau}_k$ close to it. For a pre-specified tolerance $\xi > 0$, we say that the change point τ_j is detected if there is a $\hat{\tau}_k$ such that $|\tau_j - \hat{\tau}_k| \leq \xi$, otherwise it is missed. This requires the algorithm to identify and localize changes in the distribution.
2) **(Low False Alarm Rate)** For almost every $\hat{\tau}_k$ there should be a point τ_j such that $|\hat{\tau}_k - \tau_j| \leq \xi$. A predicted change point $\hat{\tau}_k$ that is far from every true change point τ_j is considered a *false alarm*. This rules out algorithms which find true change points by proposing many spurious ones.

In Section V we measure the quality of the sequence $(\hat{\tau}_k)$ using two popular metrics in CPD literature [5], [18] to evaluate the performance, (a) area under the precision-recall curve, (b) F1-score. The F1-score, precision, and recall are defined as:

$$\text{F1-score} = \frac{2 \cdot \text{precision} \times \text{recall}}{\text{precision} + \text{recall}},$$
$$\text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \qquad \text{recall} = \frac{\text{TP}}{\text{TP} + \text{FN}},$$

where TP, FP, and FN represent the total number of true positive, false positive, and false negative points, respectively.

The recall is only high if most of the change points in the data are predicted and few of them are missed (the missed change points are FN). The precision is only high if most of the predicted change points truly are change points (TP) while at the same time the spurious change points (FP) are minimized. These directly correspond to the two heuristics of what makes a sequence $(\hat{\tau}_k)$ close to (τ_j). Note that these metrics do not consider true negatives (TN), and this is natural in our problem because the vast majority of time-points in a sequence are not change points so metrics such as specificity (TN / (TN + FP)) are always extremely close to 1.

The CPD setting described above is idealized in two important ways. First, in practice there is often no sharp threshold where a shift in the sampling distribution occurs. The distribution may undergo a short phase transition, and one would like to register this as only a single change point instead of a change point at every time step during the transition [20]. Second, real data distributions may exhibit subtle fluctuations around a typical distribution, and only occasionally undergo meaningful transitions. This can make statistical tests which are *too* powerful ineffective in practice because one often does not want to register small fluctuations as change points. The discrepancy between practical and theoretical change points should inform the design of a CPD algorithm. Any CPD algorithm should be *sensitive* enough to identify potentially subtle shifts and capture all the true changes, while being *robust* to insignificant fluctuations.

A. Related Works

There are many flavors of CPD: supervised [21] or unsupervised [17]; online [3] or offline [17]; number of change points (single [22] or multiple [23]); univariate [24] or multivariate [25] signals; and if the signal model is parametric or non-parametric. Supervised CPD methods require annotated training data with labeled change points to train the model, while unsupervised CPD methods operate without labeled change points, allowing them to detect change points in new signals without prior knowledge of their locations. Online CPD methods analyze data in real-time or streaming scenarios, enabling immediate detection of change points using only historical data, while offline methods process the entire dataset retrospectively. Single CPD methods are designed to detect a single change in the data, whereas multiple CPD methods are able to detect multiple potential change points. Parametric approaches make specific assumptions about the underlying data distributions and detect change points based on statistics computed from pre-change and post-change distributions [26], [27]. The most widely used parametric approaches are cumulative sum- [28], and generalized likelihood ratio (GLR)- [29] based CPD. These parametric approaches are mostly suited for *quickest* change point detection where the goal is to detect a change in the quickest time. In contrast, nonparametric methods are able to detect change points without any assumptions on the underlying distribution. In this work we are interested in unsupervised, offline, and non-parametric CPD methods.

One popular method for nonparametric CPD is the sliding window technique [18], [19] (Algorithm 1), which measures the similarity at every possible point of the signal using two-sample GoF statistics. While these methods may not be optimal for offline CPD in parametric settings, such as the one considered in [30], we focus our attention on these methods because they provide fast alternatives to optimal methods [19]. Classical and popular statistics such as Kolmogorov-Smirnov [31]–[33] and Cramér-von-Mises [34]–[37] statistics have been used for CPD. However, these statistics rely on comparing empirical CDFs, and only apply when the data dimension $d = 1$. Maximum mean discrepancy (MMD) [13] is a GoF statistic that comes from a family of integral probability metrics [38] and has been used to detect change points when $d > 2$ [17]. Recently, OT-based statistics have also been proposed for sliding-window-based CPD for multivariate signals: Wasserstein-1 (W1) distance [6], a distribution-free variant of Wasserstein distance that measures the Wasserstein distance of the Q-Q function to the uniform measure known as Wasserstein-Quantile test (WQT) [5], [14], and Sinkhorn divergence [16].

III. Background

A. Optimal Transport and Rank Energy

Let $\mathcal{P}(\Omega)$ denote the space of probability measures over an open set $\Omega \subseteq \mathbb{R}^d$ and let $\mathcal{P}_{ac}(\Omega)$ be those measures which are absolutely continuous with respect to the Lebesgue measure on Ω (i.e. those that admit a density function). For two measures $P, Q \in \mathcal{P}(\Omega)$, the optimal transport problem with squared Euclidean ground cost seeks an optimal *coupling* π between the source distribution P and the target distribution Q via solving [39]

$$W_2^2(P,Q) \triangleq \min_{\pi \in \Pi(P,Q)} \int \frac{1}{2}\|x-y\|^2 d\pi(x,y), \qquad (1)$$

where $\Pi(P,Q)$ is the set of joint probability measures on $\mathcal{P}(\Omega \otimes \Omega)$ with marginals P and Q.

The connection between optimal transport and ranking can be understood starting with $d = 1$, where if $P \in \mathcal{P}_{ac}(\Omega)$ and $Q = \text{Unif}[0,1]$, the optimal plan is supported on $\{(x, \text{CDF}_P(x))\}$ [39] ($\text{CDF}_P(x)$ is the cumulative distribution function of P), which corresponds to a cyclically monotone rearrangement that in turn aligns with the natural ordering on $\mathbb{R}$. In higher dimensions, as implicitly noted in the seminal papers extending the notion of ranks to higher dimensions via optimal transport [10], [40], [41], the *key geometric* property of cyclical monotonicty is preserved in that when $P \in \mathcal{P}_{ac}(\Omega)$, by the Brenier-McCann theorem [42], [43], the optimal transport plans are supported on cyclically monotone sets i.e., on $\{(x, T(x)) : x \in \text{supp}(P)\}$ for a map $T : \mathbb{R}^d \to \mathbb{R}^d$, which is a gradient of a convex function (hence cyclically monotone by a well-known theorem of Rockafeller [44]) and satisfies $(T\#P)[A] = P[T^{-1}(A)]$ for all measurable sets A.

This allows one to meaningfully interpret *multivariate ranks via optimal transport maps* as corresponding to a cyclically monotone rearrangement with respect to a target measure Q. Fixing the target measure Q to be $\text{Unif}([0,1]^d)$ motivates the following definition of the *multivariate rank map*.

Definition III.1 ([10]). Let $P \in \mathcal{P}_{ac}(\Omega)$ and let $Q = \text{Unif}([0,1]^d)$. The *(multivariate) rank map* of P is defined as $\mathtt{R} = \nabla\phi : \mathbb{R}^d \to \mathbb{R}^d$ where ϕ is the convex function such that $\nabla\phi$ optimally transports P to Q.

Using this notion of rank, rank energy is defined as follows.

Definition III.2 (Definition 3.2, [10]). Let $P_X, P_Y \in \mathcal{P}_{ac}(\Omega)$ and let $X, X' \overset{i.i.d.}{\sim} P_X$ and $Y, Y' \overset{i.i.d.}{\sim} P_Y$. Let $P_\lambda = \lambda P_X + (1-\lambda)P_Y$ denote the mixture distribution for any $\lambda \in (0,1)$

and let $\mathtt{R}_\lambda$ be the multivariate rank map of P_λ as in Definition III.1. The *(population) rank energy* is given by

$$\begin{aligned}\mathtt{RE}_\lambda(P_X, P_Y)^2 \triangleq\; & 2\mathbb{E}\big\|\mathtt{R}_\lambda(X) - \mathtt{R}_\lambda(Y)\big\| \\ & -\mathbb{E}\big\|\mathtt{R}_\lambda(X) - \mathtt{R}_\lambda(X')\big\| - \mathbb{E}\big\|\mathtt{R}_\lambda(Y) - \mathtt{R}_\lambda(Y')\big\|.\end{aligned} \tag{2}$$

In [10] it is shown that the rank energy is distribution free under the null, consistent against alternatives (under the alternative hypothesis the probability of accepting the null hypothesis goes to zero as the number of samples goes to infinity), and computationally feasible as long as the number of samples is not too large. These make the rank energy a good candidate for GoF testing since it is zero if $P_X = P_Y$ and is distribution-free under the null.

B. Entropic Optimal Transport and Soft Rank Energy

In order to define the soft rank energy we begin by introducing the *entropy-regularized OT* (EOT) problem. The entropy-regularized version of (1) adds an additional term to the objective [45]–[47]. For $\varepsilon > 0$, the primal formulation of EOT is given by

$$\min_{\pi\in\Pi(P,Q)} \int \frac{1}{2}\|x-y\|^2 \mathrm{d}\pi(x,y) + \varepsilon \mathrm{KL}(\pi \,\|\, P\otimes Q), \tag{3}$$

where

$$\mathrm{KL}(\pi|P\otimes Q) \triangleq \int \ln\left(\frac{d\pi(x,y)}{dP(x)dQ(y)}\right) d\pi(x,y).$$

Let π_ε denote the solution to (3) and let π_ε^x denote the conditional distribution of π_ε with first coordinate fixed at $X = x$. Extending the ideas in [48], [11] proposed the following.

Definition III.3 ([11]). Let $P \in \mathcal{P}_{ac}(\Omega)$ and $Q = \mathrm{Unif}([0,1]^d)$. Define the *entropic rank map* via

$$\mathtt{R}_\varepsilon(x) \triangleq \mathbb{E}_{Y\sim\pi_\varepsilon}[Y|X=x] = \mathbb{E}_{Y\sim\pi_\varepsilon^x}[Y],$$

the conditional expectation under the coupling π_ε.

Remark III.4. We note that $\mathtt{R}_\varepsilon$ is a gradient of a convex function [49] thereby maintaining the key geometric property of rank maps, namely cyclical monotonicity.

Based on this notion, and motivated by the nicer sample and computational complexity as well as differentiability of entropic rank maps in [11] the following variant of rank energy was proposed and utilized for learning generative models.

Definition III.5 (Soft Rank Energy, [11]). Let $P_X, P_Y \in \mathcal{P}_{ac}(\Omega)$ and let $X, X' \overset{i.i.d.}{\sim} P_X, Y, Y' \overset{i.i.d.}{\sim} P_Y$. Let $P_\lambda = \lambda P_X + (1-\lambda)P_Y$ for $\lambda \in (0,1)$ and let $\mathtt{R}_\lambda^\varepsilon$ be the entropic rank map of P_λ. The *soft rank energy* (sRE) is defined as:

$$\begin{aligned}\mathtt{sRE}_\lambda^\varepsilon(P_X, P_Y)^2 =\; & 2\mathbb{E}\big\|\mathtt{R}_\lambda^\varepsilon(X) - \mathtt{R}_\lambda^\varepsilon(Y)\big\| \\ & - \mathbb{E}\big\|\mathtt{R}_\lambda^\varepsilon(X) - \mathtt{R}_\lambda^\varepsilon(X')\big\| - \mathbb{E}\big\|\mathtt{R}_\lambda^\varepsilon(Y) - \mathtt{R}_\lambda^\varepsilon(Y')\big\|.\end{aligned} \tag{4}$$

Note that while $\mathtt{sRE}_\lambda^\varepsilon$ is not a metric, it is symmetric and $\mathtt{sRE}_\lambda^\varepsilon(P_X, P_Y) = 0$ if $P_X = P_Y$, which is useful in CPD applications.

C. Estimating $\mathtt{RE}_\lambda$ and $\mathtt{sRE}_\lambda^\varepsilon$ from Samples

Let $X_1, ..., X_n \overset{i.i.d.}{\sim} P$ and $Y_1, ..., Y_n \overset{i.i.d.}{\sim} Q$ be jointly independent samples. Using these samples one constructs the empirical measures $P^n = \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$, $Q^n = \frac{1}{n}\sum_{i=1}^n \delta_{Y_i}$. The plug-in estimates of the OT map T^n and the optimal EOT coupling π_ε^n are obtained by solving

$$T^n = \operatorname*{arg\,min}_{T:T\#P^n=Q^n} \frac{1}{n}\sum_{i=1}^n \big\|T(X_i) - X_i\big\|^2,$$

$$\pi_\varepsilon^n = \operatorname*{arg\,min}_{\pi\in\Pi(P^n,Q^n)} \sum_{i,j=1}^n \pi_{ij}\big\|X_i - Y_j\big\|^2 + \varepsilon\pi_{ij}\log\pi_{ij}.$$

The plug-in estimate of the entropic map T_ε^n is given by

$$T_\varepsilon^n(X_i) = \mathbb{E}_{Y\sim\hat{\pi}_\varepsilon}[Y|X=X_i] = n\sum_{j=1}^n (\pi_\varepsilon^n)_{ij} Y_j.$$

Note that like T^n, the map T_ε^n is only defined on the samples $\{X_1, ..., X_n\}$.

When $Q = \mathrm{Unif}([0,1]^d)$ we say that the estimate T^n is the *sample rank* and denote it by $\mathtt{R}^n$. In our setting we consider a mixture distribution of P_X and P_Y and we specify the number of samples from P_X with m, the number of samples from P_Y with n, and draw $m+n$ samples from Q. In this setting we use $\mathtt{R}_{m,n}$ to denote the sample rank where m, n refer to the number of samples from each distribution. Analogously the estimate $T_\varepsilon^{m,n}$ is referred to as the *entropic sample rank* and denoted by $\mathtt{R}_{m,n}^\varepsilon$.

We can now define the *sample rank energy* and *sample soft rank energy*.

Definition III.6. Given two sets of samples $X_1, \ldots, X_m \overset{i.i.d.}{\sim} P_X$ and $Y_1, \ldots, Y_n \overset{i.i.d.}{\sim} P_Y$, define the empirical mixture of the two sets of samples $P^{m+n} = \frac{1}{m+n}\left(\sum_{i=1}^m \delta_{X_i} + \sum_{j=1}^n \delta_{Y_j}\right)$. Let $Q^{m+n} = \frac{1}{m+n}\sum_{i=1}^{n+m}\delta_{U_i}$ where $U_i \overset{i.i.d}{\sim} \mathrm{Unif}([0,1]^d)$. Let $\mathtt{R}_{m,n}$ be the sample rank of P^{m+n} to Q^{m+n}. The *sample rank energy* is given by

$$\begin{aligned}\mathtt{RE}_{m,n}(P_X,P_Y)^2 \triangleq\; & \frac{2}{mn}\sum_{i,j=1}^{m,n}\|\mathtt{R}_{m,n}(X_i) - \mathtt{R}_{m,n}(Y_j)\| \\ & - \frac{1}{m^2}\sum_{i,j=1}^m \|\mathtt{R}_{m,n}(X_i) - \mathtt{R}_{m,n}(X_j)\| \\ & - \frac{1}{n^2}\sum_{i,j=1}^n \|\mathtt{R}_{m,n}(Y_i) - \mathtt{R}_{m,n}(Y_j)\|.\end{aligned}$$

Let $\mathtt{R}_{m,n}^\varepsilon$ be the entropic sample rank of P^{m+n} to Q^{m+n} defined via the plug-in estimate of the entropic map. The *sample soft rank energy* is given by

$$\begin{aligned}\texttt{sRE}^{\varepsilon}_{m,n}(P_X,P_Y)^2 &\triangleq \frac{2}{mn}\sum_{i,j=1}^{m,n}\|\texttt{R}^{\varepsilon}_{m,n}(X_i)-\texttt{R}^{\varepsilon}_{m,n}(Y_j)\|\\ &-\frac{1}{m^2}\sum_{i,j=1}^{m}\|\texttt{R}^{\varepsilon}_{m,n}(X_i)-\texttt{R}^{\varepsilon}_{m,n}(X_j)\|\\ &-\frac{1}{n^2}\sum_{i,j=1}^{n}\|\texttt{R}^{\varepsilon}_{m,n}(Y_i)-\texttt{R}^{\varepsilon}_{m,n}(Y_j)\|.\end{aligned}$$

IV. Utilizing $\texttt{sRE}^{\varepsilon}_{\lambda}$ for Change Point Detection

The primary application of this paper is the use of $\texttt{sRE}^{\varepsilon}_{\lambda}$ as a means to solve the CPD problem introduced in Section II by using it as a GoF statistic (see Algorithm 1 and Figure 1). While it is now well-established that entropic OT maps can be computed much faster with practical methods that are parallelizable [50], computation of OT maps that require solving a linear program still does not scale well in practice [51], thereby putting use of RE at a computational disadvantage compared to sRE. In this Section, we argue that soft rank energy has several more important advantages over rank energy for CPD.

A. Wasserstein Continuity Properties of Rank and Soft Rank Energy

In [10] it was shown that the rank energy is distribution-free under the null hypothesis that $P_X = P_Y$. Given that the soft rank energy is "close to" the rank energy (as quantified by Theorem IV.4), it is reasonable to hope that it should retain this property in an approximate sense. While the rank energy enjoys this important theoretical property, it poses issues for CPD beyond its considerable computational and statistical burdens [11]. In particular the rank energy can be highly unstable to small Wasserstein perturbations when measured according to W_1, which is defined[1]

$$W_1(P,Q)\triangleq \min_{\pi\in\Pi(P,Q)}\int\|x-y\|d\pi(x,y).$$

This is made precise in the following theorem.

Theorem IV.1. *For any $\lambda\in(0,1)$ and any $\epsilon,\delta>0$ there exists a pair of measures P_X,P_Y with $W_1(P_X,P_Y)<\delta$ and*

$$\texttt{RE}_{\lambda}(P_X,P_Y)\geq \sup_{Q_X,Q_Y\in\mathcal{P}_{ac}(\Omega)}\texttt{RE}_{\lambda}(Q_X,Q_Y)-\epsilon.$$

The proof is deferred to Section A1, and relies on an invariance to dilations of the OT map. This result shows that the rank energy strongly distorts the Wasserstein-1 metric, in the sense that there are no universal constants $0<\alpha\leq\beta<\infty$ such that

$$\alpha W_1(P_X,P_Y)\leq \texttt{RE}_{\lambda}(P_X,P_Y)\leq\beta W_1(P_X,P_Y)$$

for any pair P_X,P_Y. The nonexistence of β follows immediately from Theorem IV.1. The nonexistence of α follows by taking a sequence $\{(P^i_X,P^i_Y)\}_{i=1}^{\infty}$ so that $W_1(P^i_X,P^i_Y)\to\infty$ and noting that by definition, for any P^i_X,P^i_Y, $\texttt{RE}_{\lambda}(P^i_X,P^i_Y)^2\leq 2\sqrt{d}$. At this level, the rank energy fails to properly capture a standard notion of distance between measures, and can either greatly inflate or diminish relative to Wasserstein-1 metric.

We posit that the lack of an upper bound makes the rank energy overly sensitive and leads to a high false alarm rate in the CPD problem. This is because in practice there is an implicit, application-dependent threshold of distributional change which should be tolerated and not be flagged as a change point. In contrast, the rank energy aims to capture *any* change, no matter how subtle, which leads to the identification of change points below the implicit threshold. The proof of Theorem IV.1 also suggests that the rank energy is unstable when working with distributions that are much more concentrated than Unif$([0,1]^d)$.

In contrast, the soft rank energy enjoys a stability property with respect to W_1, which suggests that it is robust to small Wasserstein perturbations and may not raise a false alarm in these circumstances.

Theorem IV.2. *For any $\lambda\in(0,1)$ and $P_X,P_Y\in\mathcal{P}_{ac}(\mathbb{R}^d)$ it holds*

$$\texttt{sRE}^{\varepsilon}_{\lambda}(P_X,P_Y)^2\leq\frac{2d}{\varepsilon}W_1(P_X,P_Y).$$

The proof is deferred to Section A2 and relies crucially on the Lipschitz continuity of the entropic map (see Lemma A.1).

Remark IV.3. In Theorem IV.2, the factor $2d$ in the bound is an artifact of using $Q=\text{Unif}([0,1]^d)$. If instead one chooses $Q=\text{Unif}(B^d_2(u,1))$ for any $u\in\mathbb{R}^d$, then the factor of $2d$ in the bound above can be replaced by a dimension-free 8. Additionally, since $W_1(P_X,P_Y)\leq p^{1/p}W_p(P_X,P_Y)$ for all $p\geq1$ the conclusion also holds for these variants of the Wasserstein distance. We state it in terms of W_1 since up to an absolute constant it is the strongest bound of this form.

Comparing Theorem IV.1 to Theorem IV.2, there is a clear qualitative difference between the rank energy and the soft rank energy. This sensitivity also appears empirically and is demonstrated in Figure 2. In this figure when the samples are highly concentrated the rank energy suffers from large fluctuations while the soft rank energy remains comparatively smooth. This leads the RE to produce a few false positives while the sRE shows stability against those fluctuations. Theorem IV.2 also suggests the role of ε may act as a sensitivity knob with small ε leading to a highly sensitive signal while a large ε is more stable against perturbations.

B. Convergence of $\texttt{sRE}^{\varepsilon}_{\lambda}$ to $\texttt{RE}_{\lambda}$

While Theorems IV.1 and IV.2 suggest that there may be some fundamental differences between the soft rank energy and the rank energy, it is still possible to derive convergence results between them. This is to be expected since the optimal entropic coupling, π_ε between $P\in\mathcal{P}_{ac}(\mathbb{R}^d)$ and Q is known to converge weakly to the unregularized coupling $\pi=[\text{Id}\otimes T]\#P$ (where T is the OT map from P to Q) as $\varepsilon\to0^+$ (See [52] Proposition 3.2). If one imposes the condition that the OT map is Lipschitz, one may use the recent results from [53] to arrive at a *quantitative* estimate of the difference between $\texttt{sRE}^{\varepsilon}_{\lambda}$ and $\texttt{RE}_{\lambda}$.

[1] We use the convention that $W_p^p(P,Q)\triangleq\min_{\pi\in\Pi(P,Q)}\int\frac{1}{p}\|x-y\|^pd\pi(x,y)$ for all $p\geq1$.

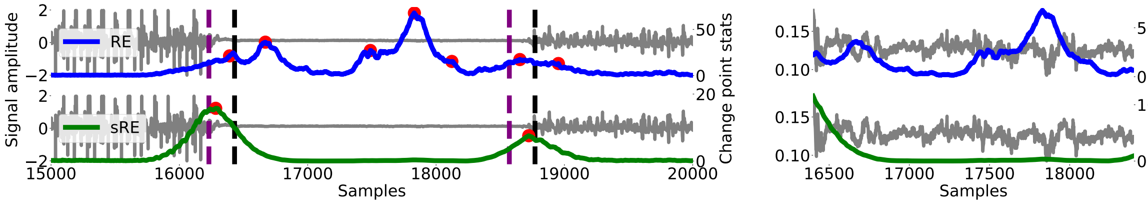

Fig. 2. The left plot shows a sequence of HASC-PAC2016 dataset with the detected change points marked by red dots. The right plot is a zoomed-in version of a short segment. Both RE and sRE can detect the true changes (black dashed line) within a certain margin (dashed purple), but RE also produces false positives due to sensitivity to small signal fluctuations. On the other hand, sRE displays greater stability in this aspect, leading to superior performance as seen in Table II.

Theorem IV.4. *Let $P_X, P_Y \in \mathcal{P}_{ac}(B(0,r))$ for some $r < \infty$. If $\mathtt{R}_\lambda$ is L-Lipschitz then it holds that $|\mathtt{sRE}^{\varepsilon}_{\lambda}(P_X,P_Y)^2 - \mathtt{RE}_\lambda(P_X,P_Y)^2| \leq C\sqrt{Ld\varepsilon\log(1/\varepsilon)} + O(\varepsilon)$, for some constant C.*

The proof is deferred to Section A3. Theorem IV.4 implies that for small ε the soft rank energy is a close approximation of the rank energy. This is important because the rank energy is distribution-free under the null hypothesis that $P = Q$. It is reasonable to expect that the soft rank energy with small ε approximately inherits this property; this is empirically observed in Figure 3, where one can see that the soft rank energy is nearly distribution-free under the null. An asymptotic variant of this result appears in [[11] Proposition 16], however they require that P_X and P_Y have compact support, $\mathtt{R}_\lambda$ is Lipschitz, the inverse map $\mathtt{R}_\lambda^{-1}$ be Hölder smooth, and that the Jacobian of $\mathtt{R}_\lambda$ be strictly positive definite. Theorem IV.4 implies their result by taking the limit as $\varepsilon \to 0^+$ and we obtain it with only the conditions that P_X and P_Y have compact support and $\mathtt{R}_\lambda$ is Lipschitz. In addition one can only derive a slower rate of convergence from an analysis of their proof and the leading coefficient depends on the integrated Fisher information [54].

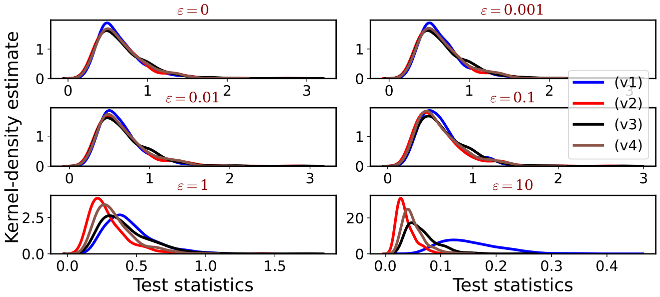

Fig. 3. Kernel density estimates of RE ($\varepsilon = 0$) and sRE under the null for v1 (Cauchy), v2 (multivariate Gaussian), v3 (Uniform), and v4 (Laplace) distributional settings scaled by a factor of $mn/(m+n)$. RE exhibits distribution-free behavior under the null, and sRE shows qualitatively similar behavior for small values of ε. However, for larger values of ε, the density curves of sRE deviate from this pattern, indicating a loss of the distribution-freeness property. Here $m = n = 200$ and the statistics are plotted using 1000 random draws.

Remark IV.5. The convergence of $\mathtt{sRE}^{\varepsilon}_{\lambda}$ to $\mathtt{RE}_\lambda$ in the presence of Theorems IV.1 and IV.2 may seem surprising since these results suggest that $\mathtt{RE}$ and $\mathtt{sRE}$ are fundamentally different. This can be reconciled by noting the bound $\mathtt{sRE}^{\varepsilon}_{\lambda}(P_X,P_Y)^2 \leq 2\sqrt{d}$ for any P_X, P_Y and ε and by choosing ε small enough relative to $W_1(P_X,P_Y)$ it will hold that $W_1(P_X,P_Y)/\varepsilon > 2\sqrt{d}$, which leads to the bound in Theorem IV.2 becoming vacuous.

C. Statistical Properties

It is now well-known that the plug-in estimate of the optimal map suffers from the curse of dimensionality. In fact, [55] show that for *any (measurable) estimator* T_0 of the OT map T there exists a measure P with

$$\mathbb{E}\int_{\mathbb{R}^d} \left\|T_0(x) - T(x)\right\|^2 dP(x) \gtrsim n^{-2/d}.$$

This says that estimation of the OT maps suffer from the curse of dimensionality unless further assumptions such as higher order smoothness of the densities are placed on the measures P and Q. Since estimation of rank energy requires estimation of the OT map, this result implies that rank energy also suffers from the curse of dimensionality in statistical estimation from samples, that is, $\mathtt{RE}_{m,n}(P_X,P_Y)^2$ will converge slowly to $\mathtt{RE}_\lambda(P_X,P_Y)^2$.

In contrast, once entropy regularization is introduced the high-dimensional statistical issues are largely avoided. In [11] it is shown that when P is subgaussian and Q has bounded support then a certain canonical extension of $T_\varepsilon^{n,n}$ (which we will also denote by $T_\varepsilon^{n,n}$) enjoys the following bound:

$$\mathbb{E}||T_\varepsilon^{n,n} - T_\varepsilon||^2_{L^2(P)} \lesssim n^{-1/2}.$$

When both P and Q have bounded support it is shown in [56] that

$$\mathbb{E}||T_\varepsilon^{n,n} - T_\varepsilon||^2_{L^2(P)} \lesssim n^{-1}.$$

Using this canonical extension and decoupling the estimation of the map from the estimation of the soft rank energy statistic, in [11] fast dimension-independent rates of convergence of soft rank energy are established. In this paper, we do not employ the canonical extension and avoid decoupling the estimates of the map and the statistic while obtaining faster rates of convergence, albeit under the assumption that measures are compactly supported. This is stated in the following result.

Theorem IV.6. *Let $P_X, P_Y \in \mathcal{P}(B(0,r))$ and let $r_0 = \max(r, \sqrt{d})$. Let $X_1, ..., X_n \overset{i.i.d.}{\sim} P_X$ and $Y_1, ..., Y_n \overset{i.i.d.}{\sim} P_Y$ be jointly independent. Then*

$$\mathbb{E}|\mathtt{sRE}^{\varepsilon}_{n,n}(P_X, P_Y)^2 - \mathtt{sRE}^{\varepsilon}_{1/2}(P_X, P_Y)^2| \\ \leq \frac{24r_0\sqrt{1+\varepsilon^2}}{\sqrt{2n}} \exp(22r_0^2/\varepsilon) + 6\sqrt{\frac{d\pi}{n}}.$$

The proof is deferred to Section A4. The first step in proving this result is to introduce an intermediary term which approximates $\mathtt{sRE}^{\varepsilon}_{1/2}$ by a discrete sum evaluated at the sample points $X_1, ..., X_n, Y_1, ..., Y_n$ and then applies the triangle inequality. This breaks the estimate into two terms, the first a Monte Carlo estimate of an expectation (which is easily controlled), the second measuring how closely $\mathtt{R}^{\varepsilon}_{n,n}$ approximates $\mathtt{R}^{\varepsilon}_{1/2}$ on the sampled points. Controlling the second term is substantially complicated by two things. First, the distribution of $(X_1, ..., X_n, Y_1, ..., Y_n)$ is *not the same* as $(P_{1/2})^{n+n}$. Second there is a dependence between the sample entropic rank $\mathtt{R}^{\varepsilon}_{n,n}$ and the points used to estimate it, and one must ensure that on these points the soft rank map is well-behaved. In [11] these issues are handled via resampling ideas in which two batches of independent samples, $(X_1, ..., X_n, Y_1, ..., Y_n)$ and $(X'_1, ..., X'_n, Y'_1, ..., Y'_n)$ are used. The first batch estimates a suitably extendable version of $\mathtt{R}^{\varepsilon}_{n,n}$ and the second batch is used to actually compute $\mathtt{sRE}^{\varepsilon}_{n,n}$ but not estimate $\mathtt{R}^{\varepsilon}_{n,n}$. This approach requires extra samples, requires artificially sampling from $P_{1/2}$ instead of P_X and P_Y separately, and requires an out-of-sample extension of the soft rank map since $\mathtt{R}^{\varepsilon}_{n,n}$ is only defined at the sample points. These technical points differentiate the estimate in [11] from the one we consider which is the true plug-in estimate of sRE, is computationally simpler, and more efficient in its use of samples. In addition in the context of the sliding-window approach, it is not clear how one would partition samples to use the estimator in [11]. However our approach is limited to measures with compact support while [11] is able to cover subgaussian measures, a question we leave to future work.

V. Numerical Evaluation

Apart from the hyperparameters that are specific to the GoF statistic picked, the other main hyperparameters for Algorithm 1 are the window size n and threshold parameter η. For practically all methods, the choice of the window size n is primarily governed by the frequency of change points. In general, if there are more change points to be expected in a given window, n should be chosen small enough (so that at the true change point there is no contamination from other distributions) and vice-versa. The threshold parameter η is typically data specific and can be either heuristically picked depending on relative size of the peaks over all the data or can be selected based on a theoretically justified threshold if the limiting statistics under the null for the statistics employed are known. In our experiments, once we have calculated the change point statistics, we use a standard peak finding procedure[2] with thresholding to identify potential change points. Since the statistical guarantees of the various tests differ, we evaluate their performance using metrics that vary the threshold parameter η over all possible values[3]. One potential drawback of the peak search algorithm is that it may generate many small sub-peaks around the largest peaks. To prevent the detection of multiple successive change points when only one change point is present, we apply a minimal horizontal distance Δ in samples to ensure that every pair of predicted change points $\hat{\tau} \neq \hat{\tau}'$ are at least Δ samples apart. A comprehensive explanation of these hyperparameters is provided in Appendix B.

A. Evaluation Metrics

As noted in Section II we consider two widely used metrics in the CPD literature [5], [18] to evaluate the performance: (a) area under the precision-recall curve (AUC-PR) and (b) best F1-score across all detection thresholds.

To account for uncertainty in the exact annotation of true change points, we allow a margin of error ξ when declaring a point either as TP or FP or FN. A predicted change point $\hat{\tau}_k$ is considered a TP if it is within ξ of a true change point τ_j (i.e., $|\tau_j - \hat{\tau}_k| \leq \xi$), otherwise it is considered a FP. A true change point τ_j that does not have a detected change within ξ is considered a FN. The choice of ξ is important for proper performance assessment. A small ξ may increase the number of FPs, while a larger ξ may misleadingly improve performance by considering detected change points far from true change points as TPs. Additionally, multiple true change points in close proximity may increase ambiguity when using a larger ξ. To ensure fairness in comparison, we use the same ξ for all methods.

Apart from the window size n, threshold η, detection margin ξ, and minimal horizontal distance Δ, the regularization parameter ε for sRE is crucial to CPD performance. Section V-B provides a detailed exploration of its impact for various choices of window size.

B. Selection of Window Size n and ε for sRE

To gain insights into selecting appropriate values for n and ε, we evaluate the performance of sRE-based CPD on a synthetic dataset. The dataset we considered is comprised of 10 distinct segments with a total length of 3300 samples and 9 change points. Details about segment lengths and corresponding distributions are provided in Table I.

For window size $n = 25$, a comparison between RE and sRE with various choices of regularization on the synthetic dataset is shown in Figure 4. In this figure we see that increasing the regularization parameter ε of sRE does indeed produce a smoothing effect on the change point statistics in agreement with Theorem IV.2. For $\varepsilon = 0.1$ and $\varepsilon = 1$ we see that increasing ε leads to smaller fluctuations away from the true change points and leads to a better performance than RE which is highly oscillatory and has many false positives which we believe is because of its poor continuity properties

[2] We use scipy.signal.find_peaks from Python Scipy1.9.1.

[3] Code to reproduce results is available at https://github.com/ShoaibBinMasud/CPD-using-sRE

This article has been accepted for publication in IEEE Transactions on Information Theory. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TIT.2024.3367182

TABLE I
UNDERLYING DISTRIBUTION AND LENGTH OF EACH SEGMENT OF THE SYNTHETIC DATASET.

	seg:1	seg:2	seg:3	seg:4	seg:5	seg:6	seg:7	seg:8	seg:9	seg:10
Distribution	$\mathcal{N}(0_d, .001I_d)$	$\mathcal{N}(0_d, .01I_d)$	$\mathcal{N}(1_d, I_d)$	Laplace$(0_d, I_d)$	$\mathcal{N}(1_d, I_d)$	$\Gamma(2, 2)$	$\mathcal{N}(0_d, .1I_d)$	$\mathcal{N}(1_d, 0_d)$	$\mathcal{N}(0_d, .01I_d)$	$\mathcal{N}(0_d, .001I_d)$
Length	300	400	500	300	400	300	200	300	200	400

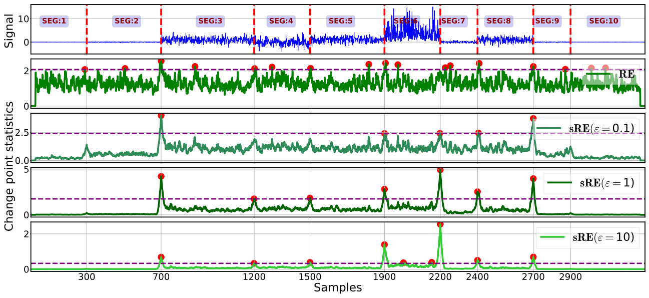

Fig. 4. **Top Row**: A single dimension of the synthetic data (top row) with true change points (vertical dotted red line). **Bottom Four Rows**: Change point statistics (with window size $n = 25$) using RE and sRE on synthetic dataset with threshold η (horizontal dashed purple) providing the best F1-score, the detected change points (red dot). We choose minimum horizontal distance $\Delta = n$. For every n, we use detection margin $\xi = 20$. The plot shows that sRE statistics become smoother as the value of ε increases.

as discussed in Theorem IV.1. However, *over-regularizing* leads to two problems. First, it causes missed changed points because the bound W_1/ε becomes small yet still dominates sRE, as is the case for the first and last change points in Figure 4. Second, in order to recover the true change points when the statistic is small one must choose a small threshold which can cause false positives elsewhere in the signal as is the case when $\varepsilon = 10$ in Figure 4. An additional plot (Figure 6) depicting the effect of ε for window size $n = 50$ is included in Appendix D.

Numerical evaluation across a range of window sizes n and regularization parameters ε is presented in Figure 5, revealing three distinct behaviors related to n. In the small n regime (e.g., $n \in \{10, 25\}$), where statistical noise in the estimate triggers numerous false alarms, moderately large regularization parameters (e.g., $\varepsilon \in \{0.5, 2\}$) are preferred, as suggested by the bounds in Theorem IV.6. For typical window sizes (e.g., $n \in \{50, 100\}$), a range of regularization parameters (e.g., $\varepsilon \in \{0.1, 1, 2, 5\}$) achieves a balance between sensitivity and robustness, with performance generally improving as n increases. In the large n regime (e.g., $n = 300$), where a single window encompasses multiple change points, localization becomes challenging, leading to performance degradation. Throughout most window size regimes, although it improves with increasing window size, small regularization parameters (e.g., $\varepsilon = 0.01$) tend to yield poor results and may suffer from numerical issues. Similarly, over-regularization (e.g., $\varepsilon = 10$) fails to achieve the best performance in most window regimes, although accuracy increases with n.

The optimal range of ε is influenced by the scale of the data (quantified by average pairwise distance, for example) and dimension, while the optimal range of n is typically governed by the frequency of the change points. However these two parameters also interact with each other in guaranteeing the quality of the estimate of the sRE statistic from samples as indicated by Theorem IV.6. While precise claims about the best choices are challenging, empirical observations suggest that setting these parameters to moderate values (e.g., $\varepsilon \in \{0.1, 1\}$ and $n \in \{50, 200\}$) tends to yield reasonable performance in various settings considered in this work.

Additional results, including a comparison with other GoF statistics-based CPD methods on the synthetic datasets, are provided in Appendix D.

C. Results on Real Data

In this study, we evaluate and compare the performance of the sRE with other GoF statistics for CPD on 5 real-world datasets including 4 time-series datasets and a hyperspectral image dataset. Detailed descriptions of these datasets as well as discussion of the various hyperparameters used in this study can be found in Appendix C.

a) HASC-PAC2016, HASC-2011: On HASC-PAC2016, sRE performs the best on all metrics. On HASC2011, sRE also performs better overall compared to most of the methods. In contrast, RE has the lowest overall performance on both datasets. This is because RE produces false positives in low amplitude regions between activities, called "rest," due to its sensitivity to any signal regardless of its amplitude (Figure 2). In contrast, sRE provides smoother statistics compared to RE which is validated by Theorem IV.2 and ignores changes in those regions, resulting in a significant improvement in performance.

b) Bee Dance: Beedance is a comparatively challenging dataset for CPD due it frequent fluctuations. Among the methods tested, SinkDiv and W1 demonstrated the best performance in terms of AUC-PR and F1-score. While sRE also performs well, it does not achieve the same level of success as SinkDiv and W1. In contrast, RE has the poorest overall performance, likely due to its tendency to respond to all fluctuations, including those that may not be considered as change points.

c) Salinas A: On this high-dimensional hyperspectral image dataset, sRE outperforms all other methods by a significant margin in AUC-PR score and a decent margin in the F1-score. To further investigate this we construct a lower-dimensional version of this dataset by performing PCA to reduce to five dimensions (which accounts for 99.79% of the variance). These findings are summarized in Table III. In this setting we see that sRE still performs competitively with all other methods. The relatively high jump in performance for the AUC-PR metric for sRE in going from low to high dimensions seems to be specific to this dataset. This is evidenced by the high dimensional synthetic data experiments in Appendix D in Table V, which indicates that while sRE remains competitive

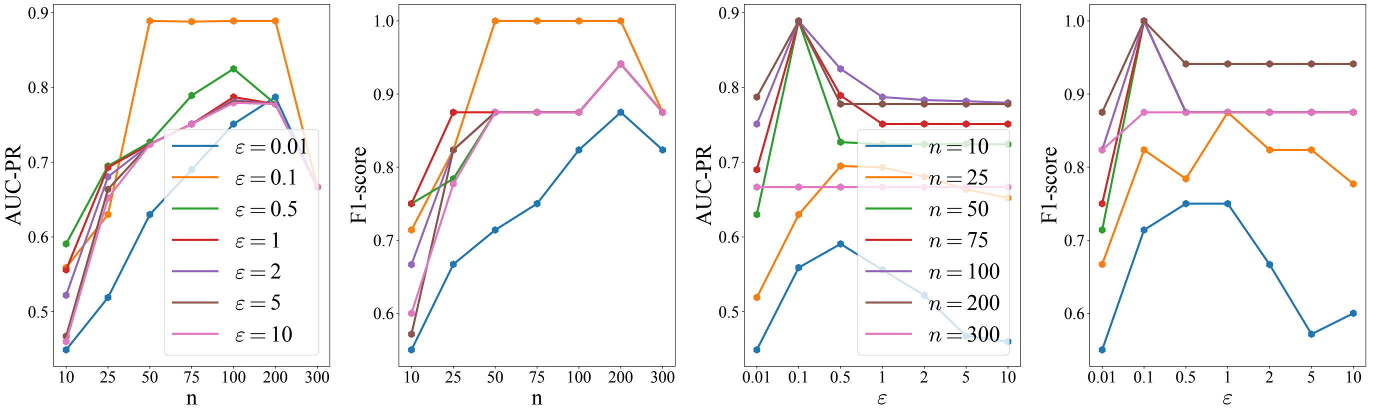

Fig. 5. Average AUC-PR and best F1-scores on synthetic dataset w.r.t. n and ε for different ε and n, respectively.

TABLE II
PERFORMANCE COMPARISON OF RE AND SRE WITH OTHER STATISTICS USED FOR CPD (BOLD: BEST).

Method	AUC-PR					Average	Best F1-score					Average
	HSAC PAC2016 (d=3)	HSAC 2011 (d=3)	Beedance (d=3)	Salinas (d=204)	ECG (d=1)		HSAC PAC2016 (d=3)	HSAC 2011 (d=3)	Beedance (d=3)	Salinas (d=204)	ECG (d=1)	
M-stat [17]	0.688	0.565	0.566	0.471	0.442	0.546	0.804	0.676	0.723	0.708	0.667	0.716
SinkDiv [16]	0.679	0.578	**0.764**	0.501	**0.487**	0.601	0.791	0.699	**0.823**	0.558	0.682	0.710
W1 [6]	0.678	**0.652**	0.763	0.252	0.441	0.557	0.806	0.702	0.820	0.525	0.682	0.707
WQT [5]	0.638	0.411	0.424	0.308	0.449	0.446	0.772	0.636	0.698	0.598	0.682	0.677
RE	0.596	0.382	0.367	0.312	0.482	0.427	0.779	0.641	0.646	0.523	0.684	0.654
sRE	**0.740**	0.598	0.687	**0.714**	0.473	**0.642**	**0.831**	**0.709**	0.801	**0.772**	0.682	**0.759**

TABLE III
TABLE INVESTIGATING THE HIGH DISCREPANCY IN AUC-PR FOR SALINAS A DATASET.

Method	AUC-PR		Best F1-score	
	PCA(5)	Original	PCA(5)	Original
M-stat [17]	0.475	0.471	0.710	0.708
SinkDiv [16]	0.307	0.501	0.613	0.558
W1 [6]	0.251	0.252	0.525	0.525
WQT [5]	0.478	0.308	0.710	0.598
RE	0.456	0.312	0.581	0.523
sRE	0.524	0.714	0.619	0.772

in high dimensions with its combined computational and statistical advantages, it may not necessarily outperform others by a large margin.

d) ECG: Despite being developed based on the concept of multivariate rank, both RE and sRE are effective at detecting change points when the signal is one-dimensional. As shown in Table II, all methods, including RE and sRE, perform similarly well on the univariate ECG signal.

e) Overall: The overall competitive performance of sRE compared to SinkDiv, W1, WQT, and M-stat for real-world data may be attributed to the following reasons. First, as shown in [10], RE being rank-based is *relatively insensitive to presence of outliers*. Real-world data is expected to contain outliers due to sensor motion and other extraneous factors. This robustness to outliers, combined with the stability to distributional perturbations (Theorems IV.1, IV.2), gives sRE an advantage over other statistics in practical cases. Second, we also note that compared with WQT, which is implicitly a univariate GoF adapted to high-dimensional data via projections, sRE is multivariate. Compared to W1, we note that in practical settings estimating the true W1 statistic from the data suffers from curse of dimensionality (with estimation error rate scaling as $O(n^{-1/d})$). On the other hand Theorem IV.6 shows that the estimation error rate for sRE is independent of the dimension and is parametric in n. We therefore expect that for moderate window sizes in practical settings sRE can perform better. We note that SinkDiv beats or comes close to sRE in performance on several data sets. We believe that this could be due to the fact that estimating Sinkhorn divergence is as sample efficient as sRE [57] and interpolates between MMD and W2 distances [15], which for some datasets may lead to improved performance. We would like to point out that there is no clear winner and we are reporting the results giving each method its best shot with comprehensive hyperparameter tuning.

VI. Conclusion and Future Work

We have established that soft rank energy enjoys efficient statistical and computational complexity, is Lipschitz with respect to Wasserstein-1, and performs well as a high-dimensional GoF measure on a range of real-world CPD problems. A problem left to future work is to extend Theorem IV.6 to measures with unbounded support under certain concentration assumptions, namely the subgaussian [58] or subexponential distributions, without the resampling and map extension tricks which are used in [11]. The important technical question of deriving theoretically optimal selection of thresholds guaranteeing a false alarm rate control is also left for future work. In this context, since sRE is a function of the entropy-regularized transport map, one can utilize recent work on limit theorems for these maps under the null [59] for establishing limit theorems for sRE. This in turn may be used to establish a theoretically optimal threshold for rejection under the null for a user specified false alarm rate.

In addition, while we have chosen the uniform distribution on the unit cube $[0,1]^d$ as the target measure in the definitions of the rank maps in this paper, it is of interest to consider the role of this distribution and if other distributions may lead to better theoretical guarantees (see Remark IV.3 for a specific example). Noting that the rank maps allow for comparisons of distributions vis-à-vis their transport maps to a specified target distribution, it is of interest to investigate the complementary picture, namely comparing distributions via their multivariate quantile maps [40], [41] and connections with the *linear optimal transport* framework [60], where one compares the distributions via the transport maps from a specific reference measure to these distributions as the target measures.

VII. Acknowledgements

This research was sponsored by the U.S. Army DEVCOM Soldier Center and was accomplished under Cooperative Agreement Number W911QY-19-2-0003 awarded to Tufts University. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. Army DEVCOM Soldier Center, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation hereon.

MW is supported by the NSF through grants CAREER CCF-1553075 and DMS-2309519. SM is supported by W911QY-19-2-0003. JMM acknowledges support from the NSF through grants DMS-1912737, DMS-1924513, DMS-2309519, and DMS-2318894 and the Camille & Henry Dreyfus Foundation Machine Learning in the Chemical Sciences and Engineering Award. SA acknowledges support by NSF grants CAREER CCF-1553075, DRL-1931978, and DMS-2309519. All authors acknowledge support through the Tufts TRIPODS Institute, supported by the NSF under grant CCF-1934553.

Appendix

A. Proofs from Section IV

1) Proof of Theorem IV.1:

Proof of Theorem IV.1. Without loss of generality we can assume that $\epsilon < \sup_{Q_X,Q_Y} \mathtt{RE}_\lambda(Q_X,Q_Y)$ since otherwise the claim holds by the positivity of $\mathtt{RE}_\lambda$ (See Section A3). Now let P'_X, P'_Y be absolutely continuous and such that

$$\mathtt{RE}_\lambda(P'_X,P'_Y) \geq \sup_{Q_X,Q_Y} \mathtt{RE}_\lambda(Q_X,Q_Y) - \epsilon.$$

Let $w = W_1(P'_X,P'_Y)$. Note that $\mathtt{RE}_\lambda(P'_X,P'_Y) > 0$ implies $w > 0$ so that the map $S : \mathbb{R}^d \to \mathbb{R}^d$ given by

$$S(x) = \frac{\delta}{w} x,$$

is well-defined. Let $P_X = S\#P'_X$ and $P_Y = S\#P'_Y$, and note that P_X, P_Y are also both absolutely continuous. We will show that $\mathtt{RE}_\lambda(P_X,P_Y) = \mathtt{RE}_\lambda(P'_X,P'_Y)$, which can be seen as a consequence of the fact that the optimal transport map has an invariance to scaling. Indeed let $\mathtt{R}'_\lambda$ denote the rank map of $P'_\lambda = \lambda P'_X + (1-\lambda)P'_Y$ and let $\mathtt{R}_\lambda$ denote the rank of $P_\lambda = \lambda P_X + (1-\lambda)P_Y$. We claim that $\mathtt{R}_\lambda = \mathtt{R}'_\lambda \circ S^{-1}$. To see that $\mathtt{R}'_\lambda \circ S^{-1}$ is a valid map, note that $S^{-1}\#P_\lambda = P'_\lambda$ and therefore

$$(\mathtt{R}'_\lambda \circ S^{-1})\#P_\lambda = \mathtt{R}'_\lambda\#(S^{-1}\#P_\lambda) = \mathtt{R}'_\lambda\#P'_\lambda = \mathrm{Unif}([0,1]^d).$$

To see that it is optimal, we can compute its gradient as

$$\begin{aligned}\nabla(\mathtt{R}'_\lambda \circ S^{-1})(x) =& \nabla S^{-1}(x)\nabla\mathtt{R}'_\lambda(S^{-1}(x))\\ =& \frac{w}{\delta}\mathbb{I}\nabla\mathtt{R}'_\lambda(S^{-1}(x))\\ =& \frac{w}{\delta}\nabla\mathtt{R}'_\lambda(S^{-1}(x)).\end{aligned}$$

Since $\mathtt{R}'_\lambda$ is the gradient of a convex function, $\nabla\mathtt{R}'_\lambda(S^{-1}(x))$ is a positive semi-definite matrix and since $w/\delta > 0$ it must be that $(w/\delta)\nabla\mathtt{R}'_\lambda(S^{-1}(x))$ is also positive semi-definite, which shows that $\mathtt{R}'_\lambda \circ S^{-1}$ is the gradient of a convex function. Recalling that P_X, P_Y are absolutely continuous and using Brenier's theorem, this shows that $\nabla\mathtt{R}'_\lambda(S^{-1}(x))$ is the unique optimal map. This confirms $\mathtt{R}_\lambda = \mathtt{R}'_\lambda \circ S^{-1}$.

In particular, this establishes

$$\begin{aligned}&\mathtt{RE}_\lambda(P_X,P_Y)^2\\ &= 2\mathbb{E}_{P_X,P_Y}\left\|\mathtt{R}_\lambda(X) - \mathtt{R}_\lambda(Y)\right\| - \mathbb{E}_{P_X}\left\|\mathtt{R}_\lambda(X) - \mathtt{R}_\lambda(X')\right\|\\ &\quad - \mathbb{E}_{P_Y}\left\|\mathtt{R}_\lambda(Y) - \mathtt{R}_\lambda(Y')\right\|\\ &= 2\mathbb{E}_{P_X,P_Y}\left\|\mathtt{R}'_\lambda(S^{-1}((X))) - \mathtt{R}'_\lambda(S^{-1}((Y)))\right\|\\ &\quad - \mathbb{E}_{P_X}\left\|\mathtt{R}'_\lambda(S^{-1}((X))) - \mathtt{R}'_\lambda(S^{-1}((X')))\right\|\\ &\quad - \mathbb{E}_{P_Y}\left\|\mathtt{R}'_\lambda(S^{-1}((Y))) - \mathtt{R}'_\lambda(S^{-1}((Y')))\right\|\\ &= 2\mathbb{E}_{P'_X P'_Y}\left\|\mathtt{R}'_\lambda(X) - \mathtt{R}'_\lambda(Y)\right\|\\ &\quad - \mathbb{E}_{P'_X}\left\|\mathtt{R}'_\lambda(X) - \mathtt{R}'_\lambda(X')\right\|\\ &\quad - \mathbb{E}_{P'_Y}\left\|\mathtt{R}'_\lambda(Y) - \mathtt{R}'_\lambda(Y')\right\|\\ &= \mathtt{RE}_\lambda(P'_X,P'_Y)^2\end{aligned}$$

Taking square roots and using the assumptions on P'_X and P'_Y shows

$$\mathtt{RE}_\lambda(P_X,P_Y) = \mathtt{RE}_\lambda(P'_X,P'_Y) \geq \sup_{Q_X,Q_Y} \mathtt{RE}_\lambda(Q_X,Q_Y) - \epsilon.$$

To conclude, let T' be the optimal map in terms of W_1 from P'_X to P'_Y. Then we have

$$\begin{aligned} W_1(P_X,P_Y) &\leq \int \big\|(\delta/w)T'((w/\delta)x)-x\big\|dP_X(x)\\ &= \frac{\delta}{w}\int \big\|T'((w/\delta)x)-(w/\delta)x\big\|dP_X(x)\\ &= \frac{\delta}{w}\int \big\|T'(x)-x\big\|dP'_X(x)\\ &= \frac{\delta}{w}W_1(P'_X,P'_Y) = \frac{\delta}{w}w=\delta \end{aligned}$$

where we have used the fact that $(\delta/w)T'((w/\delta)x)\#P_X = P_Y$, which can be verified in a similar way as above. This shows that the pair P_X, P_Y satisfy the two required properties. □

2) Proof of Theorem IV.2:

Before proving Theorem IV.2, we first establish the Lipschitz continuity of the entropic map.

Lemma A.1. *Suppose that* $\mathrm{Supp}(Q) \subseteq B_2^d(u,r)$ *for some* $u \in \mathbb{R}^d$, $r > 0$. *Then the entropic transport map* T_ε *from* P *to* Q *is* $(4r^2/\varepsilon)$*-Lipschitz continuous.*

For convenience we will introduce the notation $\Sigma_\varepsilon^x \triangleq \mathrm{Cov}_{Y\sim\pi_\varepsilon^x}(Y)$. We first recall a known result in the literature.

Lemma A.2 ([49] Lemma 1)**.** *Let* π_ε^x *denote the conditional distribution of* π_ε *given* $X = x$. *Then*

$$\nabla T_\varepsilon(x) = \frac{1}{\varepsilon}\mathrm{Cov}_{Y\sim\pi_\varepsilon^x}(Y) = \frac{1}{\varepsilon}\Sigma_\varepsilon^x.$$

Using that the Lipschitz constant of a vector-valued function is the supremum of the operator norm of its Jacobian, we have the following corollary.

Corollary A.3. *The entropic map is L-Lipschitz with respect to the Euclidean distance with*

$$L = \frac{1}{\varepsilon}\sup_{x\in\Omega}\big\|\Sigma_\varepsilon^x\big\|_{op}.$$

Proof of Lemma A.1. Note that for all x the support of π_ε^x is contained in $B_2^d(u,r)$. Let $\bar{Y} = \mathbb{E}_{Y\sim\pi_\varepsilon^x}[Y] \in B_2^d(u,r)$. Letting $Z = Y - \bar{Y}$ we have by the translation invariance of the covariance matrix and the fact that Z is mean-zero

$$\Sigma_\varepsilon^x = \mathrm{Cov}(Z) = \mathbb{E}ZZ^\top.$$

Note that

$$Z \in B_2^d(u,r) - \bar{Y} \subset (u + B_2^d(0,r)) - (u + B_2^d(0,r)) = B_2^d(0,2r)$$

and therefore for any unit $v \in \mathbb{R}^d$ with $\big\|v\big\| = 1$ we have

$$\begin{aligned} v^\top\Sigma_\varepsilon^x v &= v^\top\mathbb{E}[ZZ^\top]v\\ &= \mathbb{E}[(v^\top Z)(Z^\top v)]\\ &\leq \mathbb{E}[(\big\|v\big\|\cdot\big\|Z\big\|)(\big\|Z\big\|\cdot\big\|v\big\|)]\\ &\leq \mathbb{E}\big\|Z\big\|^2 \leq 4r^2. \end{aligned}$$

This implies that for all $x \in \Omega$ we have $\big\|\Sigma_\varepsilon^x\big\|_{op} \leq 4r^2$. Taking the supremum over x and applying Corollary A.3 we have $L = \frac{1}{\varepsilon}\sup_{x\in\Omega}\big\|\Sigma_\varepsilon^x\big\|_{op} \leq \frac{1}{\varepsilon}(4r^2)$ which proves the result. □

Proof of Theorem IV.2. First note that we are using $Q = \mathrm{Unif}([0,1]^d)$ and $\mathrm{Supp}(Q) \subset B_2^d((1/2)\mathbf{1}, \sqrt{d/4})$ where $\mathbf{1}$ denotes the all 1 vector in $\mathbb{R}^d$. Therefore by Lemma A.1 we have that soft rank map $\mathtt{R}_\lambda^\varepsilon$ from P_λ to Q is $(d/\varepsilon)-$Lipschitz.

In addition let T be a transport map from P_X to P_Y such that

$$\mathbb{E}_X\big\|T(X)-X\big\| = W_1(P_X,P_Y).$$

Now let $X, X' \sim P_X, Y, Y' \sim P_Y$ be independent of each other. Note from definition III.5 we have:

$$\begin{aligned} &\mathtt{sRE}_\lambda^\varepsilon(P_X,P_Y)^2\\ &= 2\mathbb{E}_{X,Y}\left[\big\|\mathtt{R}_\lambda^\varepsilon(X)-\mathtt{R}_\lambda^\varepsilon(Y)\big\|\right]\\ &\quad - \mathbb{E}_{X,X'}\left[\big\|\mathtt{R}_\lambda^\varepsilon(X)-\mathtt{R}_\lambda^\varepsilon(X')\big\|\right]\\ &\quad - \mathbb{E}_{Y,Y'}\left[\big\|\mathtt{R}_\lambda^\varepsilon(Y)-\mathtt{R}_\lambda^\varepsilon(Y')\big\|\right]\\ &= 2\mathbb{E}_{X,X'}\left[\big\|\mathtt{R}_\lambda^\varepsilon(X)-\mathtt{R}_\lambda^\varepsilon(T(X'))\big\|\right]\\ &\quad - \mathbb{E}_{X,X'}\left[\big\|\mathtt{R}_\lambda^\varepsilon(X)-\mathtt{R}_\lambda^\varepsilon(X')\big\|\right]\\ &\quad - \mathbb{E}_{X,X'}\left[\big\|\mathtt{R}_\lambda^\varepsilon(T(X))-\mathtt{R}_\lambda^\varepsilon(T(X'))\big\|\right]\\ &= \mathbb{E}_{X,X'}\Big[2\big\|\mathtt{R}_\lambda^\varepsilon(X)-\mathtt{R}_\lambda^\varepsilon(T(X'))\big\|\\ &\qquad\qquad - \big\|\mathtt{R}_\lambda^\varepsilon(X)-\mathtt{R}_\lambda^\varepsilon(X')\big\|\\ &\qquad\qquad - \big\|\mathtt{R}_\lambda^\varepsilon(T(X))-\mathtt{R}_\lambda^\varepsilon(T(X'))\big\|\Big]\\ &\leq \mathbb{E}_{X,X'}\Big|\big\|\mathtt{R}_\lambda^\varepsilon(X)-\mathtt{R}_\lambda^\varepsilon(T(X'))\big\| - \big\|\mathtt{R}_\lambda^\varepsilon(X)-\mathtt{R}_\lambda^\varepsilon(X')\big\|\Big|\\ &\quad + \mathbb{E}_{X,X'}\Big|\big\|\mathtt{R}_\lambda^\varepsilon(X)-\mathtt{R}_\lambda^\varepsilon(T(X'))\big\|\\ &\qquad\qquad - \big\|\mathtt{R}_\lambda^\varepsilon(T(X))-\mathtt{R}_\lambda^\varepsilon(T(X'))\big\|\Big|\\ &\leq \mathbb{E}_{X,X'}\big\|(\mathtt{R}_\lambda^\varepsilon(X)-\mathtt{R}_\lambda^\varepsilon(T(X'))) - (\mathtt{R}_\lambda^\varepsilon(X)-\mathtt{R}_\lambda^\varepsilon(X'))\big\|\\ &\quad + \mathbb{E}_{X,X'}\big\|(\mathtt{R}_\lambda^\varepsilon(X)-\mathtt{R}_\lambda^\varepsilon(T(X')))\\ &\qquad\qquad - (\mathtt{R}_\lambda^\varepsilon(T(X))-\mathtt{R}_\lambda^\varepsilon(T(X')))\big\|\\ &= \mathbb{E}_{X,X'}\left[\big\|\mathtt{R}_\lambda^\varepsilon(T(X'))-\mathtt{R}_\lambda^\varepsilon(X')\big\| + \big\|\mathtt{R}_\lambda^\varepsilon(X)-\mathtt{R}_\lambda^\varepsilon(T(X))\big\|\right]\\ &= 2\mathbb{E}_X\left[\big\|\mathtt{R}_\lambda^\varepsilon(T(X))-\mathtt{R}_\lambda^\varepsilon(X)\big\|\right]\\ &\leq 2\mathbb{E}_X\left[\frac{d}{\varepsilon}\big\|T(X)-X\big\|\right] = \frac{2d}{\varepsilon}W_1(P_X,P_Y). \end{aligned}$$

On the third line we have used that since T transports P_X to P_Y that $T(X') \sim P_Y$. In the sixth line we have used the reverse-triangle inequality. The eighth line uses the fact that X, X' are i.i.d. The ninth uses the fact that T_ε^λ is (d/ε)-Lipschitz and the last line is by the assumption on T. □

3) Proof of Theorem IV.4:

Proof of Theorem IV.4. We begin by recalling that $\mathtt{RE}_\lambda$ and $\mathtt{sRE}_\lambda^\varepsilon$ have an equivalent formulation [11]

$$\mathtt{sRE}_\lambda^\varepsilon(P_X,P_Y)^2 =$$
$$C_d\int_{\mathcal{S}^{d-1}}\int_{\mathbb{R}}\left(\mathbb{P}(a^\top\mathtt{R}_\lambda^\varepsilon(X)\leq t) - \mathbb{P}(a^\top\mathtt{R}_\lambda^\varepsilon(Y)\leq t)\right)^2 dt d\kappa(a)$$

$$\mathtt{RE}_\lambda(P_X,P_Y)^2 =$$
$$C_d\int_{\mathcal{S}^{d-1}}\int_{\mathbb{R}}\left(\mathbb{P}(a^\top\mathtt{R}_\lambda(X)\leq t) - \mathbb{P}(a^\top\mathtt{R}_\lambda(Y)\leq t)\right)^2 dt d\kappa(a)$$

where $\mathcal{S}^{d-1}$ is the unit sphere in $\mathbb{R}^d$ and $d\kappa(a)$ is integration over its surface and $C_d = (2\Gamma(d/2))^{-1}\sqrt{\pi}(d-1)\Gamma\big((d-1)/2\big)$ is an appropriate normalizing constant. Let

$$\begin{aligned}u_{a,t} &\triangleq \mathbb{P}\big(a^\top \mathtt{R}^\varepsilon_\lambda(X) \le t\big) - \mathbb{P}\big(a^\top \mathtt{R}^\varepsilon_\lambda(Y) \le t\big),\\ v_{a,t} &\triangleq \mathbb{P}\big(a^\top \mathtt{R}_\lambda(X) \le t\big) - \mathbb{P}\big(a^\top \mathtt{R}_\lambda(Y) \le t\big).\end{aligned}$$

Then, it follows that

$$\begin{aligned}&|\mathtt{sRE}^\varepsilon_\lambda(P_X,P_Y)^2 - \mathtt{RE}_\lambda(P_X,P_Y)|^2\\ &= C_d\left|\int_{\mathcal{S}^{d-1}}\int_{\mathbb{R}} u^2_{a,t} - v^2_{a,t}\,dtd\kappa(a)\right|\\ &\le C_d\int_{\mathcal{S}^{d-1}}\int_{\mathbb{R}} |u^2_{a,t} - v^2_{a,t}|dtd\kappa(a)\\ &= C_d\int_{\mathcal{S}^{d-1}}\int_{\mathbb{R}} |(u_{a,t} - v_{a,t})(u_{a,t}+v_{a,t})|dtd\kappa(a)\\ &\le 2C_d\int_{\mathcal{S}^{d-1}}\int_{\mathbb{R}} |u_{a,t} - v_{a,t}|dtd\kappa(a).\end{aligned}$$

We can further simplify the last integral as

$$\begin{aligned}&\int_{\mathbb{R}} |u_{a,t} - v_{a,t}|dt\\ &= \int_{\mathbb{R}}\Big|\mathbb{P}\left(a^\top\mathtt{R}^\varepsilon_\lambda(X) \le t\right) - \mathbb{P}\left(a^\top\mathtt{R}^\varepsilon_\lambda(Y) \le t\right)\\ &\qquad - \mathbb{P}\left(a^\top\mathtt{R}_\lambda(X) \le t\right) + \mathbb{P}\left(a^\top\mathtt{R}_\lambda(Y) \le t\right)\Big|dt\\ &\le \int_{\mathbb{R}}\left|\mathbb{P}\left(a^\top\mathtt{R}^\varepsilon_\lambda(X) \le t\right) - \mathbb{P}\left(a^\top\mathtt{R}_\lambda(X) \le t\right)\right|dt\\ &\quad + \int_{\mathbb{R}}\left|\mathbb{P}\left(a^\top\mathtt{R}^\varepsilon_\lambda(Y) \le t\right) - \mathbb{P}\left(a^\top\mathtt{R}_\lambda(Y) \le t\right)\right|.\end{aligned}$$

Let $X^\varepsilon_a = a^\top\mathtt{R}^\varepsilon_\lambda(X)$, $X_a = a^\top\mathtt{R}_\lambda(X)$ and $P_{X^\varepsilon_a}, P_{X_a}$ be their laws respectively. By the formula of Wasserstein-1 distance in dimension 1,

$$\begin{aligned}&\int_{\mathbb{R}}\left|\mathbb{P}\left(a^\top\mathtt{R}^\varepsilon_\lambda(X) \le t\right) - \mathbb{P}\left(a^\top\mathtt{R}_\lambda(X) \le t\right)\right|dt\\ &= \int_{\mathbb{R}}\left|\mathbb{P}\left(X^\varepsilon_a \le t\right) - \mathbb{P}\left(X_a \le t\right)\right|dt\\ &= W_1(P_{X^\varepsilon_a}, P_{X_a})\\ &= W_1((a^\top\mathtt{R}^\varepsilon_\lambda)\#P_X, (a^\top\mathtt{R}_\lambda)\#P_X)\\ &\le W_2((a^\top\mathtt{R}^\varepsilon_\lambda)\#P_X, (a^\top\mathtt{R}_\lambda)\#P_X)\\ &\le \sqrt{\mathbb{E}_X|a^\top\mathtt{R}^\varepsilon_\lambda(X) - a^\top\mathtt{R}_\lambda(X)|^2} \qquad (5)\\ &\le \sqrt{\mathbb{E}_X\left\|\mathtt{R}^\varepsilon_\lambda(X) - \mathtt{R}_\lambda(X)\right\|^2}.\end{aligned}$$

In equation (5) we have used the sub-optimal coupling $(a^\top\mathtt{R}^\varepsilon_\lambda(\cdot)\otimes a^\top\mathtt{R}_\lambda(\cdot))\#P_X$. We have also used the fact that $W_1 \le W_2$, Cauchy-Schwartz and the fact that $\|a\| = 1$. By an analogous computation we also have

$$\begin{aligned}&\int_{\mathbb{R}}\left|\mathbb{P}\left(a^\top\mathtt{R}^\varepsilon_\lambda(Y) \le t\right) - \mathbb{P}\left(a^\top\mathtt{R}_\lambda(Y) \le t\right)\right|dt\\ &\le \sqrt{\mathbb{E}_Y\left\|\mathtt{R}^\varepsilon_\lambda(Y) - \mathtt{R}_\lambda(Y)\right\|^2}.\end{aligned}$$

Now under the assumption that $\mathtt{R}_\lambda$ is L Lipschitz and that P_λ is supported on a bounded domain, we note from [[53], Proposition 4.5] the following bound

$$\|\mathtt{R}^\varepsilon_\lambda - \mathtt{R}_\lambda\|^2_{L^2(P_\lambda)} \le L\varepsilon\log(1/\varepsilon) + O(\varepsilon) \triangleq g(\varepsilon).$$

Now note that

$$\begin{aligned}&\|\mathtt{R}^\varepsilon_\lambda - \mathtt{R}_\lambda\|^2_{L^2(P_\lambda)}\\ &= \lambda\|\mathtt{R}^\varepsilon_\lambda(X) - \mathtt{R}_\lambda(X)\|^2_{L^2(P_X)}\\ &\quad + (1-\lambda)\|\mathtt{R}^\varepsilon_\lambda(Y) - \mathtt{R}_\lambda(Y)\|^2_{L^2(P_Y)}\\ &\le g(\varepsilon).\end{aligned}$$

This implies both

$$\begin{aligned}\|\mathtt{R}^\varepsilon_\lambda(X) - \mathtt{R}_\lambda(X)\|^2_{L^2(P_X)} &\le \frac{1}{\lambda}g(\varepsilon)\\ \|\mathtt{R}^\varepsilon_\lambda(Y) - \mathtt{R}_\lambda(Y)\|^2_{L^2(P_Y)} &\le \frac{1}{1-\lambda}g(\varepsilon).\end{aligned}$$

Collecting the computations above we have,

$$\begin{aligned}&|\mathtt{sRE}^\varepsilon_\lambda(P_X,P_Y)^2 - \mathtt{RE}_\lambda(P_X,P_Y)^2|\\ &\le 2C_d\int_{\mathcal{S}^{d-1}}\sqrt{\mathbb{E}_X\left\|\mathtt{R}^\varepsilon_\lambda(X) - \mathtt{R}_\lambda(X)\right\|^2}\\ &\quad + \sqrt{\mathbb{E}_Y\left\|\mathtt{R}^\varepsilon_\lambda(Y) - \mathtt{R}_\lambda(Y)\right\|^2}d\kappa(a)\\ &= 2C_d\gamma_d\sqrt{\mathbb{E}_X\left\|\mathtt{R}^\varepsilon_\lambda(X) - \mathtt{R}_\lambda(X)\right\|^2}\\ &\quad + 2C_d\gamma_d\mathbb{E}\sqrt{\mathbb{E}_Y\left\|\mathtt{R}^\varepsilon_{m,n}(Y) - \mathtt{R}_\lambda(Y)\right\|^2}\\ &\le 2C_d\gamma_d\left(\frac{1}{\sqrt{\lambda}} + \frac{1}{\sqrt{1-\lambda}}\right)\sqrt{g(\varepsilon)},\end{aligned}$$

where γ_d is the surface area of the unit sphere in $\mathbb{R}^d$. □

4) *Proof of Theorem IV.6:*
In order to state the proof of this result, we must introduce an additional piece of notation:

$$\begin{aligned}\mathtt{sRE}^{\varepsilon,S}_{n,n}(P_X,P_Y)^2 \triangleq \frac{1}{n^2}\sum_{i,j=1}^n\Big(&2\left\|\mathtt{R}^\varepsilon_\lambda(X_i) - \mathtt{R}^\varepsilon_\lambda(Y_j)\right\|\\ &- \left\|\mathtt{R}^\varepsilon_\lambda(X_i) - \mathtt{R}^\varepsilon_\lambda(X_j)\right\| - \left\|\mathtt{R}^\varepsilon_\lambda(Y_i) - \mathtt{R}^\varepsilon_\lambda(Y_j)\right\|\Big).\end{aligned}$$

This is a mixture of both $\mathtt{sRE}^\varepsilon_\lambda(P_X,P_Y)^2$ and $\mathtt{sRE}^\varepsilon_{m,n}(P_X,P_Y)^2$ since it takes the a finite sum when computing the integral and uses the population soft rank map. This one point of difference between both $\mathtt{sRE}^\varepsilon_\lambda(P_X,P_Y)^2$ (using the finite sum) and $\mathtt{sRE}^\varepsilon_{m,n}(P_X,P_Y)^2$ (using the population map) makes it a natural intermediate step between the two terms. The choice of superscript S is to indicate that it is a summation version of the sample soft rank energy.

This article has been accepted for publication in IEEE Transactions on Information Theory. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TIT.2024.3367182

Proof of Theorem IV.6.
Adding and subtracting $\mathtt{sRE}^{\varepsilon,S}_{n,n}(P_X,P_Y)^2$ and using the triangle inequality we have

$$\begin{aligned}
&\mathbb{E}\left|\mathtt{sRE}^{\varepsilon}_{n,n}(P_X,P_Y)^2-\mathtt{sRE}^{\varepsilon}_{1/2}(P_X,P_Y)^2\right|\\
&\le \mathbb{E}\left|\mathtt{sRE}^{\varepsilon}_{n,n}(P_X,P_Y)^2-\mathtt{sRE}^{\varepsilon,S}_{n,n}(P_X,P_Y)^2\right|\\
&\quad+\mathbb{E}\left|\mathtt{sRE}^{\varepsilon,S}_{n,n}(P_X,P_Y)^2-\mathtt{sRE}^{\varepsilon}_{1/2}(P_X,P_Y)^2\right|\\
&\le \frac{24r_0\sqrt{1+\varepsilon^2}}{\sqrt{2n}}\exp(22r_0^2/\varepsilon)+6\sqrt{\frac{d\pi}{n}}
\end{aligned}$$

where the last inequality applies Lemmas A.4 and A.5. □

The proof of Theorem IV.6 requires two technical lemmas involving $\mathtt{sRE}^{\varepsilon,S}_{n,n}(P_X,P_Y)^2$. The first handles error incurred by the map estimate.

Lemma A.4. *With the notation defined above it holds that*

$$\begin{aligned}
&\mathbb{E}\left|\mathtt{sRE}^{\varepsilon}_{n,n}(P_X,P_Y)^2-\mathtt{sRE}^{\varepsilon,S}_{n,n}(P_X,P_Y)^2\right|\\
&\qquad\le \frac{24r_0\sqrt{1+\varepsilon^2}}{\sqrt{2n}}\exp(22r_0^2/\varepsilon).
\end{aligned}$$

Proof. Through several applications of the triangle and reverse triangle inequalities one can show the first line of the following chain. The rest is using that $L^1\le L^2$ followed by the bound in Lemma A.14:

$$\begin{aligned}
&\left|\mathtt{sRE}^{\varepsilon}_{n,n}(P_X,P_Y)^2-\mathtt{sRE}^{\varepsilon,S}_{n,n}(P_X,P_Y)^2\right|\\
&=\Bigg|\frac{1}{n^2}\sum_{i,j=1}^{n}\Big(2\big\|\mathtt{R}^{\varepsilon}_{n,n}(X_i)-\mathtt{R}^{\varepsilon}_{n,n}(Y_j)\big\|\\
&\quad-\big\|\mathtt{R}^{\varepsilon}_{n,n}(X_i)-\mathtt{R}^{\varepsilon}_{n,n}(X_j)\big\|-\big\|\mathtt{R}^{\varepsilon}_{n,n}(Y_i)-\mathtt{R}^{\varepsilon}_{n,n}(Y_j)\big\|\Big)\\
&\quad-\frac{1}{n^2}\sum_{i,j=1}^{n}\Big(2\big\|\mathtt{sR}^{\varepsilon}_{1/2}(X_i)-\mathtt{sR}^{\varepsilon}_{1/2}(Y_j)\big\|\\
&\quad-\big\|\mathtt{sR}^{\varepsilon}_{1/2}(X_i)-\mathtt{sR}^{\varepsilon}_{1/2}(X_j)\big\|-\big\|\mathtt{sR}^{\varepsilon}_{1/2}(Y_i)-\mathtt{sR}^{\varepsilon}_{1/2}(Y_j)\big\|\Big)\Bigg|\\
&\le\frac{1}{n^2}\sum_{i,j=1}^{n}2\Big|\big\|\mathtt{R}^{\varepsilon}_{n,n}(X_i)-\mathtt{R}^{\varepsilon}_{n,n}(Y_j)\big\|\\
&\qquad\qquad-\big\|\mathtt{sR}^{\varepsilon}_{1/2}(X_i)-\mathtt{sR}^{\varepsilon}_{1/2}(Y_j)\big\|\Big|\\
&\quad+\Big|\big\|\mathtt{R}^{\varepsilon}_{n,n}(X_i)-\mathtt{R}^{\varepsilon}_{n,n}(X_j)\big\|-\big\|\mathtt{sR}^{\varepsilon}_{1/2}(X_i)-\mathtt{sR}^{\varepsilon}_{1/2}(X_j)\big\|\Big|\\
&\quad+\Big|\big\|\mathtt{R}^{\varepsilon}_{n,n}(Y_i)-\mathtt{R}^{\varepsilon}_{n,n}(Y_j)\big\|-\big\|\mathtt{sR}^{\varepsilon}_{1/2}(Y_i)-\mathtt{sR}^{\varepsilon}_{1/2}(Y_j)\big\|\Big|\\
&\le\frac{1}{n^2}\sum_{i,j=1}^{n}2\big\|(\mathtt{R}^{\varepsilon}_{n,n}(X_i)-\mathtt{R}^{\varepsilon}_{n,n}(Y_j))\\
&\qquad\qquad-(\mathtt{sR}^{\varepsilon}_{1/2}(X_i)-\mathtt{sR}^{\varepsilon}_{1/2}(Y_j))\big\|\\
&\quad+\big\|(\mathtt{R}^{\varepsilon}_{n,n}(X_i)-\mathtt{R}^{\varepsilon}_{n,n}(X_j))-(\mathtt{sR}^{\varepsilon}_{1/2}(X_i)-\mathtt{sR}^{\varepsilon}_{1/2}(X_j))\big\|\\
&\quad+\big\|(\mathtt{R}^{\varepsilon}_{n,n}(Y_i)-\mathtt{R}^{\varepsilon}_{n,n}(Y_j))-(\mathtt{sR}^{\varepsilon}_{1/2}(Y_i)-\mathtt{sR}^{\varepsilon}_{1/2}(Y_j))\big\|\\
&\le\frac{1}{n^2}\sum_{i,j=1}^{n}2\big\|\mathtt{R}^{\varepsilon}_{n,n}(X_i)-\mathtt{sR}^{\varepsilon}_{1/2}(X_i)\big\|\\
&\quad+2\big\|\mathtt{R}^{\varepsilon}_{n,n}(Y_j)-\mathtt{sR}^{\varepsilon}_{1/2}(Y_j)\big\|+\big\|\mathtt{R}^{\varepsilon}_{n,n}(X_i)-\mathtt{sR}^{\varepsilon}_{1/2}(X_i)\big\|\\
&\quad+\big\|\mathtt{R}^{\varepsilon}_{n,n}(X_j)-\mathtt{sR}^{\varepsilon}_{1/2}(X_j)\big\|+\big\|\mathtt{R}^{\varepsilon}_{n,n}(Y_i)-\mathtt{sR}^{\varepsilon}_{1/2}(Y_i)\big\|\\
&\quad+\big\|\mathtt{R}^{\varepsilon}_{n,n}(Y_j)-\mathtt{sR}^{\varepsilon}_{1/2}(Y_j)\big\|\\
&=\frac{4}{n}\sum_{i=1}^{n}\big\|\mathtt{R}^{\varepsilon}_{n,n}(X_i)-\mathtt{sR}^{\varepsilon}_{1/2}(X_i)\big\|+\big\|\mathtt{R}^{\varepsilon}_{n,n}(Y_i)-\mathtt{sR}^{\varepsilon}_{1/2}(Y_i)\big\|\\
&=8\big\|\mathtt{R}^{\varepsilon}_{n,n}-\mathtt{sR}^{\varepsilon}_{1/2}\big\|_{L^1((P_X^n+P_Y^n)/2}\\
&\le 8\big\|\mathtt{R}^{\varepsilon}_{n,n}-\mathtt{sR}^{\varepsilon}_{1/2}\big\|_{L^2((P_X^n+P_Y^n)/2}
\end{aligned}$$

Taking expectations on both sides and applying Jensen's inequality followed by Lemma A.14 (with r_0 so that P_X,P_Y and $Q=\mathrm{Unif}([0,1]^d)$ all have support in $B(0,r_0)$) we have

$$\begin{aligned}
&\mathbb{E}\left|\mathtt{sRE}^{\varepsilon}_{n,n}(P_X,P_Y)^2-\mathtt{sRE}^{\varepsilon,S}_{n,n}(P_X,P_Y)^2\right|\\
&\le \mathbb{E}8\big\|\mathtt{R}^{\varepsilon}_{n,n}-\mathtt{sR}^{\varepsilon}_{1/2}\big\|_{L^2((P_X^n+P_Y^n)/2)}\\
&\le 8\sqrt{\mathbb{E}\big\|\mathtt{R}^{\varepsilon}_{n,n}-\mathtt{sR}^{\varepsilon}_{1/2}\big\|^2_{L^2((P_X^n+P_Y^n)/2)}}\\
&\le \frac{24r_0\sqrt{1+\varepsilon^2}}{\sqrt{2n}}\exp(22r_0^2/\varepsilon).
\end{aligned}$$

□

The second lemma handles the error incurred by using a discrete sum instead of an integration.

Lemma A.5. *With the notation defined above it holds that*

$$\mathbb{E}\left|\mathtt{sRE}^{\varepsilon,S}_{n,n}(P_X,P_Y)^2 - \mathtt{sRE}^{\varepsilon}_{1/2}(P_X,P_Y)^2\right| \le 6\sqrt{\frac{d\pi}{n}}$$

Proof. In the setting of Lemma A.15 let $h(x,y) = \left\|\mathtt{sR}^{\varepsilon}_{1/2}(x) - \mathtt{sR}^{\varepsilon}_{1/2}(y)\right\|$. Then $\left\|h\right\|_\infty \le \sqrt{d}$. It can also be seen from the definitions that

$$\mathtt{sRE}^{\varepsilon,S}_{n,n}(P_X,P_Y)^2 = \frac{2}{n^2}\sum_{i,j=1}^{n} h(X_i,Y_j)$$
$$-\frac{1}{n^2}\sum_{i,j=1}^{n} h(X_i,X_j) - \frac{1}{n^2}\sum_{i,j=1}^{n} h(Y_i,Y_j),$$
$$\begin{aligned}\mathbb{E}\mathtt{sRE}^{\varepsilon,S}_{n,n}(P_X,P_Y)^2 &= \mathbb{E}[2h(X,Y) - h(X,X') - h(Y,Y')] \\ &= \mathtt{sRE}^{\varepsilon}_{1/2}(P_X,P_Y)^2.\end{aligned}$$

Therefore by Lemma A.15 we have

$$\mathbb{E}\left|\mathtt{sRE}^{\varepsilon,S}_{n,n}(P_X,P_Y)^2 - \mathtt{sRE}^{\varepsilon}_{1/2}(P_X,P_Y)^2\right| \le 6\sqrt{\frac{d\pi}{n}}.$$

□

B. Descriptions of Hyperparameters

In the evaluation we require several hyperparameters which have been mentioned in the main text. The complete list of the hyperparameters as well as their effects are summarized below

- Window Size n: The number of samples seen in each one of the windows. Increasing n typically leads to smoother changes in the GoF statistic over time and a gain in performance. However if the window size is too large multiple change points may fall within the same half of a window which may be problematic.

- Threshold η: The minimum value the GoF statistic must take in order to be registered as a change point. Since the GoF statistic computed on samples is very rarely a constant 0 it is useful to choose a small threshold below which no change points can be predicted. This prevents the peak finding procedure from proposing change points in regions where there are clearly no changes. Increasing η leads to discarding more and more proposed changed points. However a value of η which is too large may lead to missing subtle change points.

- Horizontal Displacement Δ: The minimum distance apart which two predicted change points must be, that is $|\hat{\tau}_j - \hat{\tau}_k| \ge \Delta$ for every j,k. Using this prevents the prediction of several change points in rapid succession due to small sub-peaks near a single large peak. The larger the setting of Δ is taken the more spaced out the predicted change points must be. Taking Δ too large relative to the frequency of the true change points may be problematic as it can force true change points to be ignored because of the horizontal displacement constraint.

- Margin of Error ξ: The maximum allowable distance which a predicted change point $\hat{\tau}_k$ can be from a true change point τ_j while still being considered correct. If $|\hat{\tau}_k - \tau_j| \le \xi$ than it is considered to have correctly identified τ_j. This only impacts the numerical evaluation of the methods and scores will increase as the margin of error increases. The choice of ξ should depend on the quality of the annotated change points which can be noisy. A large ξ may inflate the performance of the methods. A ξ which is too small may lead to poor scores for methods which perform well but do not consistently place the precise change points in a small target, especially if there is ambiguity in the proper placement of the annotations,

C. Datasets and Hyperparameters in Section V-C

(a) HASC-PAC2016: A human activity recognition dataset consists of over 700 three-axis accelerometer sequences sampled at 100Hz where the subjects perform six different actions, ‘stay’, ‘walk’, ‘jog’, ‘skip’, ‘stairs up’, and ‘stairs down’ ($d = 3$). Time points that exhibits changes in activity are annotated as ground truth. To evaluate the performance of the CPD methods on this dataset, we consider the 20 longest sequences which have an average length of 17,000 samples and 15 change points. We choose $n = 500, \xi = 200, \Delta = 250$ for this dataset. We use $\varepsilon = 0.1$ to compute sRE.

(b) HASC-2011: Another human activity recognition dataset where people perform six different actions, ‘stay’, ‘walk’, ‘escalator up’, ‘elevator up’, ‘stairs up’ and ‘stairs down’ and an accelerometer takes three-dimensional data ($d = 3$). Change points are annotated in the same way as in HASC-PAC2016 dataset. This dataset consists of 2 sequences, which have an average length of 37000 samples and 46 change points. We select $n = 500, \xi = 200, \Delta = 250$ and $\varepsilon = 0.1$ to compute sRE for this dataset.

(c) Bee Dance: A dataset containing 6 three-dimensional sequences each collected from the movement of dancing honeybees communicating through three actions: ‘turn right’, ‘turn left’ and ‘waggle’ ($d = 3$). The first two dimensions correspond to the spatial x and y coordinates and the third one is the heading angle of the bees captured via video tracking. Each sequence has an average length of 790 samples and 19 change points. For this dataset, we choose $n = 20, \xi = 10, \Delta = 10$ and $\varepsilon = 1$ for sRE.

(d) Salinas A Hyperspectral Image: A high-dimensional image consisting of 83×86 pixels, each a $d = 204$ dimensional vector of spectral reflectances. The data was recorded by the Airborne Visible/Infrared Imaging Spectrometer (AVIRIS) sensor over farmland in Salinas Valley, California, USA in 1998 at a spatial resolution of 1.3 m. Spectral signatures, ranging in recorded wavelength from 380 nm to 2500 nm across 204 spectral bands, were recorded. For this data, each pixel is treated as a sample. Samples are divided into six different classes, corresponding to the material classes of the pixels (e.g., broccoli greens, corn greens, lettuce of different ages). While not a time series, we can read the pixels row-

TABLE IV
AVERAGE AUC-PR AND BEST F1-SCORES ON SYNTHETIC DATASET (TAKEN OVER 25 INDEPENDENT INSTANCES) W.R.T. WINDOW SIZE n (BOLD: BEST, ITALIC: SECOND BEST). REPORTED SRE IS CALCULATED USING THE BEST REGULARIZATION PARAMETER $\varepsilon = 0.1$. AS n INCREASES, EACH METHOD GAINS A SIGNIFICANT IMPROVEMENT IN BOTH METRICS.

Method	AUC-PR				Best F1-score			
	$n = 25$	$n = 50$	$n = 100$	$n = 200$	$n = 25$	$n = 50$	$n = 100$	$n = 200$
M-stat [17]	0.442	0.665	0.879	0.879	0.625	0.737	**1.0**	**1.0**
SinkDiv [16]	0.329	0.425	0.598	0.657	0.510	0.695	0.834	0.875
W1 [5]	0.259	0.462	0.563	0.756	0.459	0.747	0.825	0.889
WQT [6]	**0.867**	0.879	**0.887**	0.882	**0.947**	**1.0**	**1.0**	**1.0**
RE	0.377	0.717	0.746	0.767	0.538	*0.875*	*0.875*	0.875
sRE	0.631	**0.882**	0.885	**0.886**	0.724	**1.0**	**1.0**	**1.0**

wise and annotate a pixel as a change point if it has a different labeled class from the previous pixel. For this paper, we consider first 500 samples which contain 54 change points. We choose $n = 10, \xi = 2, \Delta = 2$. To compute sRE, we use $\varepsilon = 1$.

(e) ECG: A one dimensional ($d = 1$) dataset consisting of a single sequence having a length of 8600 samples and 89 change points, where each change point represent an abnormal heartbeat. For this dataset, we choose $n = 50, \xi = 20, \Delta = 25$ and $\varepsilon = 0.1$ to compute sRE.

D. *Results on Synthetic Data*

Table IV provides a numerical comparison of sRE with other GoF statistics for various window sizes n on the synthetic dataset (Table I). For a relatively smaller window size $n = 25$, WQT outperforms all other methods, with sRE exhibiting the second-best performance. As the window size increases to $n = 50$, sRE demonstrates a comparable performance to WQT. While other statistics like M-stats, RE, SinkDiv, and W1 all show improved performance with larger window sizes, they require even larger window sizes for comparable performance.

We note that WQT is very competitive with and outperforms other methods in a small window regime. This is due to our choice of synthetic data (Table I) where we picked the distributions to be isotropic and hence each projection in WQT is as informative as working in high dimensions and may in fact be better, since estimation of WQT in 1-dimension is very sample efficient. To assess the generalizability of this scenario, we conduct further evaluations on another synthetic dataset, denoted as “synthetic data II” featuring non-isotropic distributions. The details of this new synthetic dataset are outlined below.

Synthetic data II: This is a $d = 10$ dimensional dataset comprising 31 segments and 30 change points, with each segment consisting of 50 samples. In each segment, samples are drawn from a multivariate Gaussian distribution with a mean of 0 and a randomly generated covariance matrix Σ_i. Specifically, $X_1, \ldots, X_{\tau_1} \sim \mathcal{N}(0, \Sigma_1)$, $X_{\tau_1+1}, \ldots, X_{\tau_2} \sim \mathcal{N}(0, \Sigma_2)$, and so forth. Each covariance matrix Σ_i is generated by first sampling a square matrix $\tilde{\Sigma}_i$ with i.i.d. entries sampled uniformly from $[0, 1]$. We then set $\Sigma_i = \tilde{\Sigma}_i^T \tilde{\Sigma}_i$.

The comparison on this *synthetic data II* is presented in Table V. Interestingly, WQT is no longer the top performer, and the proposed sRE is the second best method among the ones compared. This also indicates that it is not necessarily the case that sRE will outperform other methods, but is a sample and computationally efficient option even for high-dimensional datasets.

E. *Background on Entropic Optimal Transport*

1) Dual Optimality Conditions: In this section of the appendix we review the essentials of entropic optimal transport, mainly following the work in [56]. The most important fact is the dual of the EOT stated below.

Theorem A.6. *Let P, Q be distributions on $\mathbb{R}^d$ with bounded support and let $\varepsilon > 0$. Then*

$$S_\varepsilon(P, Q) = \sup_{(f,g) \in L^\infty(P) \times L^\infty(Q)} \int f dP + \int g dQ + \varepsilon$$
$$-\varepsilon \iint \exp\left(\frac{1}{\varepsilon}\left[f(x) + g(y) - \frac{1}{2}\|x - y\|^2\right]\right) dP(x) dQ(y).$$

The supremum is attained at a pair $(f_0, g_0) \in L^\infty(P) \times L^\infty(Q)$ of dual potentials, which are unique up to the translation $(f_0, g_0) \mapsto (f_0 + c, g_0 - c)$ for $c \in \mathbb{R}$.

Moreover, primal and dual solutions are linked via the following relationships. For any pair $(f, g) \in L^\infty(P) \times L^\infty(Q)$, let π be the measure with density

$$\frac{d\pi}{d(P \otimes Q)}(x, y) = \exp\left(\frac{1}{\varepsilon}\left[f(x) + g(y) - \frac{1}{2}\|x - y\|^2\right]\right). \tag{6}$$

Then the pair (f, g) is optimal for $S_\varepsilon(P, Q)$ if and only if π is a coupling of P and Q and π is optimal for $S_\varepsilon(P, Q)$.

For proof and discussion of this result see [61]. We will let (f_*, g_*) denote the optimal dual potentials in $S_\varepsilon(P, Q)$ which satisfy $\int g_* dQ = 0$. Given a set of samples $Y_1, \ldots, Y_n \overset{i.i.d.}{\sim} Q$, we define another pair of optimal dual potentials $(\overline{f}_*, \overline{g}_*)$ by

$$\overline{f}_* \triangleq f_* + \frac{1}{n} \sum_{i=1}^{n} g_*(Y_i), \qquad \overline{g}_* \triangleq g_* - \frac{1}{n} \sum_{i=1}^{n} g_*(Y_i).$$

In addition, let (f_n, g_n) be the unique optimizers for $S_\varepsilon(P^n, Q^n)$ such that $\frac{1}{n} \sum_{i=1}^{n} g_n(Y_i) = 0$. Using these definitions, the optimality condition (6) for (f_*, g_*) and can be re-written as

$$f_*(x) = -\varepsilon \ln\left(\int \exp\left(\varepsilon\left[g_*(y) - \frac{1}{2}\|x - y\|^2\right]\right) dQ(y)\right),$$
$$g_*(y) = -\varepsilon \ln\left(\int \exp\left(\varepsilon\left[f_*(x) - \frac{1}{2}\|x - y\|^2\right]\right) dP(x)\right),$$

TABLE V
AVERAGE AUC-PR AND F1-SCORE ARE COMPUTED OVER 10 INDEPENDENT INSTANCES OF SYNTHETIC DATASET II, UTILIZING A WINDOW SIZE OF $n = 25$ FOR ALL METHODS.

Method	AUC-PR	Best F1-score
M-stat [17]	0.479	0.648
SinkDiv [16]	0.572	0.615
W1 [5]	0.622	0.677
WQT [6]	0.476	0.562
RE	0.464	0.559
sRE	0.619	0.668

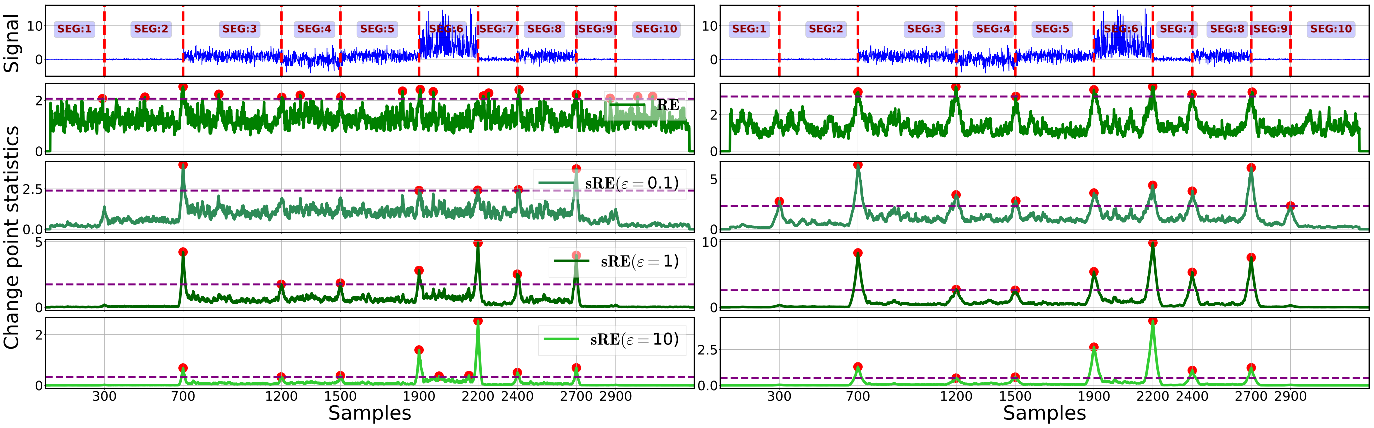

Fig. 6. Top Row: A single dimension of the synthetic data (top row) with true change points (vertical dotted red line). **Bottom Four Rows**: Change point statistics (with window size, left: $n = 25$, right: $n = 50$) using RE and sRE on synthetic dataset with threshold η (horizontal dashed purple) providing the best F1-score and the detected change points (red dot). The plot shows that sRE statistics become smoother as the value of ε increases.

which holds for P a.e. x and Q a.e. y. Similarly for (f_n, g_n) the optimality condition can be restated as

$$f_n(x) = -\varepsilon \ln\left(\int \exp\left(\frac{1}{\varepsilon}\left[g_n(y) - \frac{1}{2}\|x-y\|^2\right]\right) dQ^n(y)\right),$$
$$g_n(y) = -\varepsilon \ln\left(\int \exp\left(\frac{1}{\varepsilon}\left[f_n(x) - \frac{1}{2}\|x-y\|^2\right]\right) dP^n(x)\right),$$

which must hold for every $x \in \{X_1, ..., X_n\}$ and every $y \in \{Y_1, ..., Y_n\}$. Finally we will define the optimal relative densities as

$$\begin{aligned} p_*(x,y) &\triangleq \frac{d\pi}{d(P\otimes Q)}(x,y) \\ &= \exp\left(\frac{1}{\varepsilon}\left[f_*(x) + g_*(y) - \frac{1}{2}\|x-y\|^2\right]\right), \\ p_n(x,y) &\triangleq \frac{d\pi_n}{d(P^n\otimes Q^n)}(x,y) \\ &= \exp\left(\frac{1}{\varepsilon}\left[f_n(x) + g_n(y) - \frac{1}{2}\|x-y\|^2\right]\right), \end{aligned}$$

where the latter is only defined on the support of $P^n \otimes Q^n$. From the optimality conditions for (f_*, g_*) and the definition of $p_*(x, y)$ it follows that for P a.e. x and Q a.e. y

$$\mathbb{E}_{Y\sim Q}[p_*(x,Y)] = \int p_*(x,y')dQ(y') = 1 \qquad (7)$$
$$\mathbb{E}_{X\sim P}[p_*(y,X)] = \int p_*(x',y)dP(x') = 1. \qquad (8)$$

From these equations we can express the population and sample entropic maps as

$$T_\varepsilon(x) = \int y p_*(x,y)dQ(y),$$
$$T_\varepsilon^n(x) = \frac{1}{n}\sum_{i=1}^n Y_i p_n(x, Y_i),$$

where the latter is only defined for $x \in \{X_1, ..., X_n\}$.

2) Dual Results:

Following [56] we denote the objective in the dual problem by

$$\Phi(f,g) \triangleq \int f dP + \int g dQ + \varepsilon$$
$$-\varepsilon \iint \exp\left(\frac{1}{\varepsilon}\left[f(x)+g(y)-\frac{1}{2}\|x-y\|^2\right]\right) dP(x)dQ(y).$$

Furthermore, denote the empirical dual objective Φ_n by

$$\Phi_n(f,g) \triangleq \frac{1}{n}\sum_{i=1}^n [f(X_i) + g(Y_i)] + \varepsilon$$
$$-\frac{\varepsilon}{n^2}\sum_{i,j=1}^n \exp\left(\frac{1}{\varepsilon}\left[f(X_i)+g(Y_j)-\frac{1}{2}\|X_i - Y_j\|^2\right]\right).$$

One can calculate the partial derivatives of the empirical dual as

$$\frac{\partial \Phi_n(f,g)}{\partial f(X_i)} = \frac{1}{n} - \frac{1}{n^2}\sum_{j=1}^{n} p(X_i,Y_j)$$
$$\frac{\partial \Phi_n(f,g)}{\partial g(Y_j)} = \frac{1}{n} - \frac{1}{n^2}\sum_{i=1}^{n} p(X_i,Y_j)$$

where

$$p(x,y) = \exp\left(\frac{1}{\varepsilon}\left[f(x)+g(y)-\frac{1}{2}\left\|x-y\right\|^2\right]\right).$$

For what follows it will be useful to introduce the space $L^2(P^n)\times L^2(Q^n)$. This is the Cartesian product of the spaces $L^2(P^n)$ and $L^2(Q^n)$ and as such it has the following inner product

$$\begin{aligned}&\langle (a,b),(u,v)\rangle_{L^2(P^n)\times L^2(Q^n)}\\ &= \langle a,u\rangle_{L^2(P^n)} + \langle b,v\rangle_{L^2(Q^n)}\\ &= \frac{1}{n}\sum_{i=1}^{n} a(X_i)u(X_i) + \frac{1}{n}\sum_{j=1}^{n} b(Y_j)v(Y_j)\end{aligned}$$

for the functions $a,u\in L^2(P^n), b,v\in L^2(Q^n)$. This space is a Hilbert space and its norm is given by

$$\left\|(a,b)\right\|^2_{L^2(P^n)\times L^2(Q^n)} = \langle (a,b),(a,b)\rangle_{L^2(P^n)\times L^2(Q^n)}.$$

In addition, we can identify the gradient $\nabla\Phi_n(f,g)$ as an element of (the dual of) $L^2(P^n)\times L^2(Q^n)$, which is given by

$$\begin{aligned}&\nabla\Phi_n(f,g)\\ &= n\cdot\left(\left[\frac{\partial \Phi_n(f,g)}{\partial f(X_1)},...,\frac{\partial \Phi_n(f,g)}{\partial f(X_n)}\right],\right.\\ &\qquad\left.\left[\frac{\partial \Phi_n(f,g)}{\partial g(Y_1)},...,\frac{\partial \Phi_n(f,g)}{\partial g(Y_n)}\right]\right)\\ &= \left(\left[1-\frac{1}{n}\sum_{j=1}^{n}p(X_1,Y_j),...,1-\frac{1}{n}\sum_{j=1}^{n}p(X_n,Y_j)\right],\right.\\ &\qquad\left.\left[1-\frac{1}{n}\sum_{i=1}^{n}p(X_i,Y_1),...,1-\frac{1}{n}\sum_{i=1}^{n}p(X_i,Y_n)\right]\right).\end{aligned}$$

Note that because we are treating $\nabla\Phi_n(f,g)$ as an element of $L^2(P^n)\times L^2(Q^n)$ we must multiply it by a factor of n because it must be the unique element which satisfies

$$\begin{aligned}&\lim_{h\to 0}\frac{\Phi_n((f,g)+h(a,b))-\Phi_n(f,g)}{h}\\ &\quad= \langle \nabla\Phi_n(f,g),(a,b)\rangle_{L^2(P^n)\times L^2(Q^n)}\end{aligned}$$

and the extra factor of n is required to cancel the factor of $\frac{1}{n}$ which appears in the definition of $\langle\cdot,\cdot,\rangle_{L^2(P^n)\times L^2(Q^n)}$.

Using the formulas above we can now calculate

$$\begin{aligned}&\left\|\nabla\Phi_n(f,g)\right\|^2_{L^2(P^n)\times L^2(Q^n)}\\ &= \frac{1}{n}\sum_{i=1}^{n}\left(1-\frac{1}{n}\sum_{j=1}^{n}p(X_i,Y_j)\right)^2 \qquad (9)\\ &\quad+\frac{1}{n}\sum_{j=1}^{n}\left(1-\frac{1}{n}\sum_{i=1}^{n}p(X_i,Y_j)\right)^2.\end{aligned}$$

Note that by the translational invariance of the dual potentials we have for every $c\in\mathbb{R}$ that

$$\begin{aligned}\Phi(f,g) &= \Phi(f+c,g-c),\\ \Phi_n(f,g) &= \Phi_n(f+c,g-c),\\ \nabla\Phi_n(f,g) &= \nabla\Phi_n(f+c,g-c).\end{aligned}$$

We now move onto a first basic structural result bounding the optimal dual potentials above and below. This result is essentially contained in both [56], [57] and we include it here only to make the constants explicit in our case.

Lemma A.7. *Let $P,Q\in\mathcal{P}(B(0,r))$ and let $(f_*,g_*),(f_n,g_n)$ be the optimal dual potentials as above for $S_\varepsilon(P,Q)$ and $S_\varepsilon(P^n,Q^n)$ respectively. Then*

$$\begin{aligned}\left\|f_n\right\|_{L^\infty(P^n)},\left\|g_n\right\|_{L^\infty(Q^n)} &\le 2r^2,\\ \left\|f_*\right\|_{L^\infty(P)},\left\|g_*\right\|_{L^\infty(Q)} &\le 2r^2.\end{aligned}$$

As a consequence it also holds that

$$\left\|\overline{f}_*\right\|_{L^\infty(P)},\left\|\overline{g}_*\right\|_{L^\infty(Q)}\le 4r^2.$$

In particular for $(P\otimes Q)$-a.e. (x,y) and every $(x,y)\in Supp(P^n\otimes Q^n)$,

$$\begin{aligned}\exp\left(-\frac{6r^2}{\varepsilon}\right)&\le p_*(x,y)\le\exp\left(\frac{4r^2}{\varepsilon}\right),\\ \exp\left(-\frac{6r^2}{\varepsilon}\right)&\le p_n(x,y)\le\exp\left(\frac{4r^2}{\varepsilon}\right).\end{aligned}$$

Proof. The optimality condition (6) and marginal constraints on π_* imply for P-a.e. x that

$$\begin{aligned}1 &= \int\exp\left(\frac{1}{\varepsilon}\left[f_*(x)+g_*(y)-\frac{1}{2}\left\|x-y\right\|^2\right]\right)dQ(y)\\ &\ge \exp\left(\frac{1}{\varepsilon}\left[f_*(x)-\frac{1}{2}(2r)^2\right]\right)\int\exp\left(\frac{1}{\varepsilon}g_*(y)\right)dQ(y)\\ &\ge \exp\left(\frac{1}{\varepsilon}\left[f_*(x)-2r^2\right]\right)\end{aligned}$$

where the first inequality uses the fact that $x,y\in B(0,r)$ which implies $\left\|x-y\right\|\le 2r$ and the last inequality uses Jensen's inequality and the assumption that $\int g_*dQ = 0$. Taking logs on both sides and rearranging we see that for P a.e. x

$$f_*(x)\le 2r^2. \qquad (10)$$

Using this with the optimality and the marginal constraints on π_* we have

$$\begin{aligned}1 &= \int\exp\left(\frac{1}{\varepsilon}\left[f_*(x)+g_*(y)-\frac{1}{2}\left\|x-y\right\|^2\right]\right)dP(x)\\ &\le \int\exp\left(\frac{1}{\varepsilon}\left[2r^2+g_*(y)\right]\right)dP(x)\\ &= \exp\left(\frac{1}{\varepsilon}\left[2r^2+g_*(y)\right]\right).\end{aligned}$$

Taking logarithms on both sides we have

$$-2r^2\le g_*(y).$$

We next claim that $\int f_*(x)dP(x) \geq 0$ which can be seen from the fact that

$$\begin{aligned}
&0 \leq S_\varepsilon(P,Q) \\
&= \int f_* dP + \int g_* dQ + \varepsilon - \\
&\quad \varepsilon \iint \exp\left(\frac{1}{\varepsilon}\left[f_*(x)+g_*(y)-\frac{1}{2}\left\|x-y\right\|^2\right]\right) dP(x)dQ(y) \\
&= \int f_* dP + 0 - \varepsilon \int d\pi + \varepsilon \\
&= \int f_* dP
\end{aligned}$$

where we have used the optimality condition on g_* to remove one integral and (6) to simplify the other. With this established we can repeat the proof above swapping the roles of (x, f_*, P) with (y, g_*, Q) respectively with the only important note being that

$$\int \exp\left(\frac{1}{\varepsilon} f_*\right) dP \geq \exp\left(\frac{1}{\varepsilon}\int f_* dP\right) \geq \exp\left(\frac{1}{\varepsilon}\cdot 0\right) = 1$$

which is enough to show

$$\begin{aligned}
\exp\left(\frac{1}{\varepsilon}\left[g_*(y) - \frac{1}{2}(2r)^2\right]\right) \int \exp\left(\frac{1}{\varepsilon} f_*(x)\right) dP(x) \\
\geq \exp\left(\frac{1}{\varepsilon}\left[g_*(y) - 2r^2\right]\right)
\end{aligned}$$

which mirrors the calculation above.

The proof for f_n and g_n is completely analogous, just replacing the integrals with summations as needed.

The bounds on $\overline{f}_*$ and $\overline{g}_*$ follow from

$$\begin{aligned}
|\overline{f}_*(X_i)| &= \left| f_*(X_i) + \frac{1}{2n}\sum_{j=1}^{2n} g_n(Z_j) \right| \\
&\leq |f_*(X_i)| + \frac{1}{2n}\sum_{j=1}^{2n} |g_n(Z_j)| \\
&\leq 2r^2 + 2r^2 = 4r^2,
\end{aligned}$$

and an identical calculation for $\overline{g}_*$.

The bounds on p_* and p_n follow from

$$\begin{aligned}
p_*(x,y) &= \exp\left(\frac{1}{\varepsilon}\left[f_*(x) + g_*(y) - \frac{1}{2}\left\|x-y\right\|^2\right]\right) \\
&\leq \exp\left(\frac{1}{\varepsilon}\left[2r^2 + 2r^2 - 0\right]\right) \\
&= \exp\left(\frac{4r^2}{\varepsilon}\right), \\
p_*(x,y) &= \exp\left(\frac{1}{\varepsilon}\left[f_*(x) + g_*(y) - \frac{1}{2}\left\|x-y\right\|^2\right]\right) \\
&\geq \exp\left(\frac{1}{\varepsilon}\left[-2r^2 - 2r^2 - 2r^2\right]\right) \\
&= \exp\left(-\frac{6r^2}{\varepsilon}\right),
\end{aligned}$$

and an analogous calculation for p_n. □

For convenience we introduce the set

$$\begin{aligned}
\mathcal{S}_L \triangleq \Big\{ &(f,g) \in L^\infty(P^n) \times L^\infty(Q^n) : \\
&\left\|f\right\|_{L^\infty(P^n)}, \left\|g\right\|_{L^\infty(Q^n)} \leq L, \int g dQ^n = 0 \Big\}.
\end{aligned}$$

With this set defined we have the following two results, which are slight generalizations of existing results in [56] in that they use a general r instead of $r = 1/2$.

Lemma A.8. *Let $P, Q \in \mathcal{P}(B(0,r))$. Then for each L, Φ_n is δ-strongly concave with respect to the norm $\left\|\cdot\right\|_{L^2(P^n)\times L^2(Q^n)}$ on $\mathcal{S}_L$ for $\delta = \exp(-[2L + 2r^2]/\varepsilon)/\varepsilon$ in the sense that for any $(f,g),(f',g') \in \mathcal{S}_L$ we have with probability 1*

$$\begin{aligned}
&\Phi_n(f,g) - \Phi_n(f',g') \\
&\geq \langle \nabla\Phi_n(f,g), (f,g) - (f',g')\rangle_{L^2(P^n)\times L^2(Q^n)} \\
&\quad + \frac{\delta}{2}\left\|(f,g) - (f',g')\right\|^2_{L^2(P^n)\times L^2(Q^n)}.
\end{aligned}$$

Proof. Fix $(f,g),(f',g') \in \mathcal{S}_L$ and define the function $h : [0,1] \to \mathbb{R}$ by

$$h(t) \triangleq \Phi_n((1-t)f + tf', (1-t)g + tg').$$

Then it suffices to show that h satisfies

$$h''(t) \leq -\delta\left\|(f,g) - (f',g')\right\|^2_{L^2(P^n)\times L^2(Q^n)},$$

for all $t \in [0,1]$. Fix $t \in [0,1]$. A direct calculation shows

$$\begin{aligned}
&h''(t) \\
&= -\frac{1}{\varepsilon n^2}\sum_{i,j=1}^{n}(f(X_i) - f'(X_i) + g(Y_j) - g'(Y_j))^2 \\
&\quad \times \exp\left(\frac{1}{\varepsilon}\left[(1-t)(f(X_i) + g(Y_j))\right.\right. \\
&\qquad \left.\left. + t(f'(X_i) + g'(Y_k)) - \frac{1}{2}\left\|X_i - Y_j\right\|^2\right]\right).
\end{aligned}$$

By the bounds $|f|, |f'|, |g|, |g'| \leq L$ and $\left\|X_i - Y_j\right\| \leq 2r$, we have

$$\begin{aligned}
h''(t) \leq -&\frac{1}{\varepsilon n^2}\exp(-[2L + 2r^2]/\varepsilon) \\
&\times \sum_{i,j=1}^{n}(f(X_i) - f'(X_i) + g(Y_j) - g'(Y_j))^2.
\end{aligned}$$

The last sum can be re-factored as

$$\begin{aligned}
&\frac{1}{n^2}\sum_{i,j=1}^{n}(f(X_i) - f'(X_i) + g(Y_j) - g'(Y_j))^2 \\
&= \left\|(f,g) - (f',g')\right\|^2_{L^2(P^n)\times L^2(Q^n)} \\
&\quad + \frac{2}{n^2}\left[\sum_{i=1}^{n} f(X_i) - f'(X_i)\right]\left[\sum_{j=1}^{n} g(Y_j) - g'(Y_j)\right] \\
&= \left\|(f,g) - (f',g')\right\|^2_{L^2(P^n)\times L^2(Q^n)}.
\end{aligned}$$

On the second line the latter term is zero because $\int g dQ^n = \int g' dQ^n = 0$ since $g, g' \in \mathcal{S}_L$. □

A direct consequence of strong concavity is the following result known as the Polyak-Łojasiewicz (PL) inequality [62].

Lemma A.9. *Let $P, Q \in \mathcal{P}(B(0,r))$. Let $L > 0$ be such that $(f_n, g_n) \in \mathcal{S}_L$. Then for any $f, g \in \mathcal{S}_L$,*

$$\begin{aligned}&\Phi_n(f_n, g_n) - \Phi_n(f, g)\\ &\leq \frac{\varepsilon}{2} e^{[2L+2r^2]/\varepsilon} \big\| \nabla \Phi_n(f, g) \big\|^2_{L^2(P^n) \times L^2(Q^n)}.\end{aligned}$$

Combining the two preceding lemmas we have the following result known as the "error bound".

Lemma A.10. *Let $P, Q \in \mathcal{P}(B(0,r))$. Let $L > 0$ be such that $(f_n, g_n) \in \mathcal{S}_L$. The for any $f, g \in \mathcal{S}_L$,*

$$\begin{aligned}&\big\| (f_n, g_n) - (f, g) \big\|^2_{L^2(P^n) \times L^2(Q^n)}\\ &\quad \leq \varepsilon^2 e^{4[L+r^2]/\varepsilon} \big\| \nabla \Phi_n(f, g) \big\|^2_{L^2(P^n) \times L^2(Q^n)}.\end{aligned} \tag{11}$$

Proof. Note that since (f_n, g_n) is optimal for Φ_n we have $\nabla \Phi_n(f_n, g_n) = 0$. Therefore by Lemmas A.8 and A.9 we have

$$\begin{aligned}&\frac{\exp(-[2L+2r^2]/\varepsilon)}{2\varepsilon} \big\| (f_n, g_n) - (f, g) \big\|^2_{L^2(P^n) \times L^2(Q^n)}\\ &\leq \Phi_n(f_n, g_n) - \Phi_n(f, g)\\ &\leq \frac{\varepsilon}{2} e^{[2L+2r^2]/\varepsilon} \big\| \nabla \Phi_n(f, g) \big\|^2_{L^2(P^n) \times L^2(Q^n)}.\end{aligned}$$

Multiplying the first and last by $2\varepsilon \exp([2L+2r^2]/\varepsilon)$ gives

$$\begin{aligned}&\big\| (f_n, g_n) - (f, g) \big\|^2_{L^2(P^n) \times L^2(Q^n)}\\ &\leq \varepsilon^2 e^{4[L+r^2]/\varepsilon} \big\| \nabla \Phi_n(f, g) \big\|^2_{L^2(P^n) \times L^2(Q^n)}.\end{aligned}$$

□

We now proceed to an upper bound in expectation of the quantity on the right hand side of (11). This is the first point at which sampling patterns play any role as well as the first time we derive a novel result which is not essentially contained in [56].

Lemma A.11. *Suppose that $P_X, P_Y, Q \in \mathcal{P}(B(0,r))$. Let $X_1, \ldots X_n \sim P_X, Y_1, \ldots, Y_n \sim P_Y$ and $Z_1, \ldots, Z_{2n} \sim Q$ be jointly independent samples. Let $P_{1/2} = \frac{1}{2}P_X + \frac{1}{2}P_Y$. Let Φ_{2n} denote the dual objective between $\frac{1}{2}(P_X^n + P_Y^n)$ and Q^{2n} and let f_*, g_* be the optimal entropic potentials between $P_{1/2}$ and Q with $\int g_* dQ = 0$. Then*

$$\mathbb{E} \big\| \nabla \Phi_{2n}(f_*, g_*) \big\|^2_{L^2((P_X^n + P_Y^n)/2) \times L^2(Q^{2n})} \leq \frac{9 \exp(8r^2/\varepsilon)}{8n}$$

where the expectation is with respect to the samples $X_1, \ldots, X_n, Y_1, \ldots, Y_n, Z_1, \ldots, Z_{2n}$.

Proof. Using (9), taking expectations and applying linearity we have

$$\begin{aligned}&\mathbb{E} \big\| \nabla \Phi_{2n}(f_*, g_*) \big\|^2_{L^2((P_X^n + P_Y^n)/2) \times L^2(Q^{2n})}\\ &= \mathbb{E}\left[\frac{1}{2n} \sum_{i=1}^{n} \left(1 - \frac{1}{2n} \sum_{j=1}^{2n} p_*(X_i, Z_j) \right)^2 \right]\\ &+ \mathbb{E}\left[\frac{1}{2n} \sum_{i=1}^{n} \left(1 - \frac{1}{2n} \sum_{j=1}^{2n} p_*(Y_i, Z_j) \right)^2 \right]\\ &+ \mathbb{E}\left[\frac{1}{2n} \sum_{j=1}^{2n} \left(1 - \frac{1}{2n} \sum_{i=1}^{n} p_*(X_i, Z_j) - \frac{1}{2n} \sum_{i=1}^{n} p_*(Y_i, Z_j) \right)^2 \right]\end{aligned}$$

We will handle the three terms separately. For the first term we have

$$\begin{aligned}&\mathbb{E}\left[\frac{1}{2n} \sum_{i=1}^{n} \left(1 - \frac{1}{2n} \sum_{j=1}^{2n} p_*(X_i, Z_j) \right)^2 \right]\\ &= \frac{1}{2} \mathbb{E}\left[\left(\frac{1}{2n} \sum_{j=1}^{2n} (1 - p_*(X_1, Z_j)) \right)^2 \right]\\ &= \frac{1}{2} \frac{1}{4n^2} \sum_{j,k=1}^{2n} \mathbb{E}\left[(1 - p_*(X_1, Z_j))(1 - p_*(X_1, Z_k)) \right]\\ &= \frac{1}{8n^2} \sum_{j=1}^{2n} \mathbb{E}\left[(1 - p_*(X_1, Z_j))^2 \right]\\ &= \frac{1}{4n} \mathrm{Var}(p_*(X_1, Z_j))\\ &\leq \frac{1}{4n} \frac{\exp(8r^2/\varepsilon)}{4}\\ &= \frac{\exp(8r^2/\varepsilon)}{16n},\end{aligned}$$

where the first line uses the fact that the X_i are identically distributed, the second is just factoring and linearity of expectation, the third uses the fact that $\mathbb{E}[(1 - p_*(X_1, Z_j))(1 - p_*(X_1, Z_k))] = 0$ if $j \neq k$ (see below), the fourth uses that $\mathbb{E} p_*(X_1, Z_j) = 1$, and the fifth uses that by Lemma A.7 that $p_* \in [0, \exp(4r^2/\varepsilon)]$ and Popoviciu's inequality. The zero-mean formula is verified as follows

$$\begin{aligned}&\mathbb{E}[(1 - p_*(X_1, Z_j))(1 - p_*(X_1, Z_k))]\\ &= \mathbb{E}_{X_1}\left[\mathbb{E}_{Z_j, Z_k}\left[(1 - p_*(X_1, Z_j))(1 - p_*(X_1, Z_k)) \middle| X_1 \right] \right]\\ &= \mathbb{E}_{X_1}\left[\mathbb{E}_{Z_j}\left[1 - p_*(X_1, Z_j) \middle| X_1 \right] \mathbb{E}_{Z_k}\left[1 - p_*(X_1, Z_k) \middle| X_1 \right] \right]\\ &= \mathbb{E}_{X_1}[(0)(0)] = 0\end{aligned}$$

where we have used that $1 - p_*(X_1, Z_j)$ is conditionally independent of $1 - p_*(X_1, Z_k)$ given X_1, followed by (7).

Replacing X with Y in the calculations above one can handle the second term in an identical way and show

$$\mathbb{E}\left[\frac{1}{2n}\sum_{i=1}^{n}\left(1-\frac{1}{2n}\sum_{j=1}^{2n}p_*(Y_i,Z_j)\right)^2\right] \le \frac{\exp(8r^2/\varepsilon)}{16n}.$$

This handles the first two terms which must be controlled. We now turn our focus to the last term which is the most challenging to handle.

$$\begin{aligned}
&\mathbb{E}\left[\frac{1}{2n}\sum_{j=1}^{2n}\left(1-\frac{1}{2n}\sum_{i=1}^{n}p_*(X_i,Z_j)-\frac{1}{2n}\sum_{i=1}^{n}p_*(Y_i,Z_j)\right)^2\right]\\
&=\mathbb{E}\left[\left(\frac{1}{2n}\sum_{i=1}^{n}(1-p_*(X_i,Z_1))+\frac{1}{2n}\sum_{i=1}^{n}(1-p_*(Y_i,Z_1))\right)^2\right]\\
&=\mathbb{E}\Bigg[\frac{1}{4n^2}\sum_{i,k=1}^{n}(1-p_*(X_i,Z_1))(1-p_*(X_k,Z_1))\\
&\qquad+2(1-p_*(X_i,Z_1))(1-p_*(Y_k,Z_1))\\
&\qquad+(1-p_*(Y_i,Z_1))(1-p_*(Y_k,Z_1))\Bigg]\\
&=\frac{1}{4n^2}\sum_{i=1}^{n}\mathbb{E}\Big[(1-p_*(X_i,Z_1))^2\\
&\quad+2(1-p_*(X_i,Z_1))(1-p_*(Y_i,Z_1))+(1-p_*(Y_i,Z_i))^2\Big]\\
&\le\frac{1}{4n}4\exp(8r^2/\varepsilon)=\frac{\exp(8r^2/\varepsilon)}{n}
\end{aligned}$$

where the second line is just linearity, i.i.d. assumptions, and refactoring. The third expanding the squares. The fourth uses a fact shown below, and the fifth uses the uniform bounds on p_* from Lemma A.7 which in turn implies an upper bound on $1-p_*$. The fact that we must show is that for $i \neq k$

$$\begin{aligned}
\mathbb{E}\Big[&(1-p_*(X_i,Z_1))(1-p_*(X_k,Z_1))\\
&+2(1-p_*(X_i,Z_1))(1-p_*(Y_i,Z_1))\\
&\quad+(1-p_*(Y_i,Z_1))(1-p_*(Y_k,Z_1))\Big]=0.
\end{aligned}$$

Note that by (8)

$$\begin{aligned}
1&=\mathbb{E}_{W\sim P_{1/2}}\left[p_*(W,Z)\right]\\
&=\mathbb{E}_{X,Y}\left[(1/2)p_*(X,Z)+(1/2)p_*(Y,Z)\right]\\
&=\frac{1}{2}\mathbb{E}_X\left[p_*(X,Z)\right]+\frac{1}{2}\mathbb{E}_Y\left[p_*(Y,Z)\right].
\end{aligned}$$

Rearranging the first and last equalities shows

$$\mathbb{E}_X\left[1-p_*(X,Z)\right]=\mathbb{E}_Y\left[p_*(Y,Z)-1\right]=-\mathbb{E}_Y\left[1-p_*(Y,Z)\right]$$

which is the crucial identity that we require. Using this we have conditioned on Z_1 that

$$\begin{aligned}
&\mathbb{E}[(1-p_*(X_i,Z_1))(1-p_*(X_k,Z_1))\\
&\qquad+(1-p_*(X_i,Z_1))(1-p_*(Y_k,Z_1))]\\
&=\mathbb{E}[(1-p_*(X_i,Z_1))]\mathbb{E}[(1-p_*(X_k,Z_1))]\\
&\qquad+\mathbb{E}[(1-p_*(X_i,Z_1))]\mathbb{E}[(1-p_*(Y_k,Z_1))]\\
&=\mathbb{E}[(1-p_*(X_i,Z_1))]\mathbb{E}[(1-p_*(X_k,Z_1))]\\
&\qquad+\mathbb{E}[(1-p_*(X_i,Z_1))]\left(-\mathbb{E}[(1-p_*(X_k,Z_1))]\right)\\
&=0.
\end{aligned}$$

Similarly,

$$\begin{aligned}
&\mathbb{E}[(1-p_*(Y_i,Z_1))(1-p_*(Y_k,Z_1))\\
&\qquad+(1-p_*(X_i,Z_1))(1-p_*(Y_k,Z_1))]\\
&=\mathbb{E}[(1-p_*(Y_i,Z_1))]\mathbb{E}[(1-p_*(Y_k,Z_1))]\\
&\qquad+\mathbb{E}[(1-p_*(X_i,Z_1))]\mathbb{E}[(1-p_*(Y_k,Z_1))]\\
&=\mathbb{E}[(1-p_*(Y_i,Z_1))]\mathbb{E}[(1-p_*(Y_k,Z_1))]\\
&\qquad+\left(-\mathbb{E}[(1-p_*(Y_i,Z_1))]\right)\mathbb{E}[(1-p_*(Y_k,Z_1))]\\
&=0.
\end{aligned}$$

Adding these expressions proves the required result.

Tracking back the bounds above we have shown

$$\begin{aligned}
&\mathbb{E}\left\|\nabla\Phi_n(f_*,g_*)\right\|^2_{L^2((P_X^n+P_Y^n)/2)\times L^2(Q^{2n})}\\
&\le\frac{\exp(8r^2/\varepsilon)}{16n}+\frac{\exp(8r^2/\varepsilon)}{16n}+\frac{\exp(8r^2/\varepsilon)}{n}\\
&=\frac{9\exp(8r^2/\varepsilon)}{8n}.
\end{aligned}$$

□

By combining Lemmas A.7, A.10, and A.11 we achieve the following bound in the deviation of the potentials. This bound compares the estimated potentials to the true potentials once they have been appropriately shifted to account for the fact that the optimal potentials are only unique up to an additive constant.

Lemma A.12. *Consider the setting of Lemma A.11. Then $\overline{f_*},\overline{g_*}$ satisfy*

$$\begin{aligned}
&\mathbb{E}\left\|(f_n,g_n)-(\overline{f}_*,\overline{g}_*)\right\|^2_{L^2((P_X^n+P_Y^n)/2)\times L^2(Q^{2n})}\\
&\qquad\le\frac{9\varepsilon^2}{8n}\exp(28r^2/\varepsilon).
\end{aligned}$$

Proof. Let L be such that $f_n,g_n,\overline{f}_*,\overline{g}_*$ are with probability 1 contained in $\mathcal{S}_L$. Taking expectations in Lemma A.10 and then applying Lemma A.11 we have

$$\begin{aligned}
&\mathbb{E}\left\|(f_n,g_n)-(\overline{f}_*,\overline{g}_*)\right\|^2_{L^2((P_X^n+P_Y^n)/2)\times L^2(Q^{2n})}\\
&\le\varepsilon^2e^{4[L+r^2]/\varepsilon}\mathbb{E}\left\|\nabla\Phi_{2n}(\overline{f}_*,\overline{g}_*)\right\|^2_{L^2((P_X^n+P_Y^n)/2)\times L^2(Q^{2n})}\\
&=\varepsilon^2e^{4[L+r^2]/\varepsilon}\mathbb{E}\left\|\nabla\Phi_{2n}(f_*,g_*)\right\|^2_{L^2((P_X^n+P_Y^n)/2)\times L^2(Q^{2n})}\\
&\le\varepsilon^2e^{4[L+r^2]/\varepsilon}\frac{9\exp(8r^2/\varepsilon)}{8n}.
\end{aligned}$$

Lemma A.7 ensures that $f_n, g_n, \overline{f}_*, \overline{g}_* \in \mathcal{S}_{4r^2}$. Using this setting of L in the bound above we have

$$\begin{aligned}
&\mathbb{E}\big\|(f_n, g_n) - (\overline{f}_*, \overline{g}_*)\big\|^2_{L^2((P_X^n+P_Y^n)/2)\times L^2(Q^{2n})} \\
&\leq \varepsilon^2 e^{4[4r^2+r^2]/\varepsilon} \frac{9\exp(8r^2/\varepsilon)}{8n} \\
&= \frac{9\varepsilon^2}{8n}\exp(28r^2/\varepsilon).
\end{aligned}$$

□

By observing that the relative densities p_* and p_n are determined by the dual potentials (f_*, g_*) and (f_n, g_n) one can immediately convert the result above into a bound on the relative densities as follows.

Lemma A.13. *Consider the setting of Lemma A.11. Then the relative density p_n satisfies*

$$\mathbb{E}\big\|p_n - p_*\big\|^2_{L^2((P_X^n+P_Y^n)/2\otimes Q^{2n})} \leq \frac{9\varepsilon^2}{4n}\exp(44r^2/\varepsilon).$$

Proof. To start note that for every $(w, z) \in (\{X_1, ..., X_n\} \cup \{Y_1, ..., Y_n\}) \times \{Z_1, ..., Z_{2n}\}$ we have

$$\begin{aligned}
&|p_n(w,z) - p_*(w,z)| \\
&= \bigg|\exp\left(\frac{1}{\varepsilon}\left[f_n(w) + g_n(z) - \frac{1}{2}\|w-z\|^2\right]\right) \\
&\qquad - \exp\left(\frac{1}{\varepsilon}\left[f_*(w) + g_*(z) - \frac{1}{2}\|w-z\|^2\right]\right)\bigg| \\
&= \bigg|\exp\left(\frac{1}{\varepsilon}\left[f_n(w) + g_n(z) - \frac{1}{2}\|w-z\|^2\right]\right) \\
&\qquad - \exp\left(\frac{1}{\varepsilon}\left[\overline{f}_*(w) + \overline{g}_*(z) - \frac{1}{2}\|w-z\|^2\right]\right)\bigg| \\
&\leq e^{\frac{1}{\varepsilon}8r^2}\left|f_n(w) - \overline{f}_*(w) + g_n(z) - \overline{g}_*(z)\right| \\
&\leq e^{\frac{1}{\varepsilon}8r^2}\left|f_n(w) - \overline{f}_*(w)\right| + e^{\frac{1}{\varepsilon}8r^2}\left|g_n(z) - \overline{g}_*(z)\right|
\end{aligned}$$

where we have used the fact that $f_n, g_n, f_*, g_* \in S_{4r^2}$ by Lemma A.7 followed by the fact that e^t is e^C-Lipschitz over $(-\infty, C]$.

From here we can compute

$$\begin{aligned}
&\big\|p_n - p_*\big\|^2_{L^2((P_X^n+P_Y^n)/2\otimes Q^{2n})} \\
&= \frac{1}{4n^2}\sum_{i=1}^{n}\sum_{j=1}^{2n}|p_n(X_i,Z_j) - p_*(X_i,Z_j)|^2 \\
&\quad + |p_n(Y_i,Z_j) - p_*(Y_i,Z_j)|^2 \\
&\leq \frac{1}{4n^2}\sum_{i=1}^{n}\sum_{j=1}^{2n} 2e^{16r^2/\varepsilon}|f_n(X_i) - \overline{f}_*(X_i)| \\
&\quad + 2e^{16r^2/\varepsilon}|f_n(Y_i) - \overline{f}_*(Y_i)| + 4e^{16r^2/\varepsilon}|g_n(Z_j) - \overline{g}_*(Z_j)| \\
&= 2e^{16r^2/\varepsilon}\big\|f_n - \overline{f}_*\big\|^2_{L^2((P_X^n+P_Y^n)/2)} + 2e^{16r^2/\varepsilon}\big\|g_n - \overline{g}_*\big\|^2_{L^2(Q^{2n})} \\
&= 2e^{16r^2/\varepsilon}\big\|(f_n,g_n) - (\overline{f}_*, \overline{g}_*)\big\|^2_{L^2((P_X^n+P_Y^n)/2)\times L^2(Q^{2n})}.
\end{aligned}$$

Taking expectations on the first and last and applying Lemma A.12 we have

$$\begin{aligned}
&\mathbb{E}\big\|p_n - p_*\big\|^2_{L^2((P_X^n+P_Y^n)/2\otimes Q^{2n})} \\
&\leq 2e^{16r^2/\varepsilon}\mathbb{E}\big\|(f_n,g_n) - (\overline{f}_*, \overline{g}_*)\big\|^2_{L^2((P_X^n+P_Y^n)/2)\times L^2(Q^{2n})} \\
&\leq \frac{9\varepsilon^2}{4n}\exp(44r^2/\varepsilon).
\end{aligned}$$

□

Now we can proceed to bounding the deviation of the entropic map on the samples.

Lemma A.14. *Consider the setting of Lemma A.11. Then the entropic map T_ε^n satisfies*

$$\mathbb{E}\big\|T_\varepsilon^n - T_\varepsilon\big\|^2_{L^2((P_X^n+P_Y^n)/2)} \leq \frac{9r^2(1+\varepsilon^2)}{2n}\exp(44r^2/\varepsilon)$$

Proof. For each sample $w \in \{X_1, ..., X_n\} \cup \{Y_1, ..., Y_n\}$ we have the bound

$$\begin{aligned}
&\big\|T_\varepsilon^n(w) - T_\varepsilon(w)\big\|^2 \\
&= \bigg\|\frac{1}{2n}\sum_{j=1}^{2n} p_n(w,Z_j)Z_j - \int p_*(w,z)z\,dQ(z)\bigg\|^2 \\
&\leq 2\bigg\|\frac{1}{2n}\sum_{j=1}^{2n}(p_n(w,Z_j) - p_*(w,Z_j))Z_j\bigg\|^2 \\
&\quad + 2\bigg\|\frac{1}{2n}\sum_{j=1}^{2n} p_*(w,Z_j)Z_j - \int p_*(w,z)z\,dQ(z)\bigg\|^2
\end{aligned}$$

We will handle these two terms separately. For the first we have by Jensen's inequality followed by the boundedness of Q

$$\begin{aligned}
&\bigg\|\frac{1}{2n}\sum_{j=1}^{2n}(p_n(w,Z_j) - p_*(w,Z_j))Z_j\bigg\|^2 \\
&\leq \frac{1}{2n}\sum_{j=1}^{2n}\big\|(p_n(x,Z_j) - p_*(w,Z_j))Z_j\big\|^2 \\
&\leq \frac{1}{2n}\sum_{j=1}^{2n} r^2\left(p_n(w,Z_j) - p_*(w,Z_j)\right)^2.
\end{aligned}$$

For the second term we can expand the square and take expectation over the Z_j to obtain

$$\begin{aligned}
&\mathbb{E}\bigg\|\frac{1}{2n}\sum_{j=1}^{2n} p_*(w,Z_j)Z_j - \int p_*(w,z)z\,dQ(z)\bigg\|^2 \\
&= \frac{1}{4n^2}\sum_{j,k=1}^{2n}\mathbb{E}\Big\langle p_*(w,Z_j)Z_j - \int p_*(w,z)z\,dQ(z), \\
&\qquad\qquad p_*(w,Z_k)Z_k - \int p_*(w,z)z\,dQ(z)\Big\rangle \\
&= \frac{1}{2n}\mathbb{E}\Big\|p_*(w,Z_1)Z_1 - \int p_*(w,z)z\,dQ(z)\Big\|^2
\end{aligned}$$

where we have used that for fixed x that $p_*(x,Z_j)Z_j - \int p_*(x,z)z\,dQ(z)$ and $p_*(x,Z_k)Z_k - \int p_*(x,z)z\,dQ(z)$ are zero-mean and independent for all $j \neq k$, which implies that

This article has been accepted for publication in IEEE Transactions on Information Theory. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TIT.2024.3367182

the cross terms cancel. We can further bound this by using Lemma A.7

$$\begin{aligned}&\mathbb{E}\big\|p_*(x,Z_1)Z_1 - \int p_*(x,z)zdQ(z)\big\|^2\\ &\le \mathbb{E}\big\|p_*(x,Z_1)Z_1\big\|^2\\ &\le \mathbb{E}\big\|p_*\big\|_\infty^2 \cdot \big\|Z_1\big\|^2\\ &\le r^2\exp(8r^2/\varepsilon)\end{aligned}$$

where the first inequality is an application of the variational formula for the variance and the fact that $\mathbb{E}[p_*(x,Z_1)Z_1] = \int p_*(x,z)zdQ(z)$.

Combining the inequalities derived we have

$$\begin{aligned}&\mathbb{E}\big\|T_\varepsilon^n - T_\varepsilon\big\|^2_{L^2((P_X^n+P_Y^n)/2)}\\ &= \mathbb{E}\frac{1}{2n}\sum_{i=1}^n \big\|T_\varepsilon^n(X_i) - T_\varepsilon(X_i)\big\|^2 + \big\|T_\varepsilon^n(Y_i) - T_\varepsilon(Y)\big\|^2\\ &\le \mathbb{E}\frac{1}{n}\sum_{i=1}^n\Bigg[\frac{r^2}{2n}\sum_{j=1}^{2n}(p_n(X_i,Z_j) - p_*(X_i,Z_j))^2\\ &\qquad\qquad + (p_n(Y_i,Z_j) - p_*(Y_i,Z_j))^2\Bigg]\\ &\quad + \mathbb{E}\frac{1}{n}\sum_{i=1}^n\left[\frac{1}{2n}r^2\exp(8r^2/\varepsilon) + \frac{1}{2n}r^2\exp(8r^2/\varepsilon)\right]\\ &= 2r^2\mathbb{E}\big\|p_n - p_*\big\|^2_{L^2((P_X^n+P_Y^n)/2\otimes Q^{2n})} + \frac{r^2}{n}\exp(8r^2/\varepsilon)\\ &\le \frac{9r^2\varepsilon^2}{2n}\exp(44r^2/\varepsilon) + \frac{r^2}{n}\exp(8r^2/\varepsilon)\\ &\le \frac{9r^2(1+\varepsilon^2)}{2n}\exp(44r^2/\varepsilon)\end{aligned}$$

where the second to last inequality follows from Lemma A.13. □

F. Additional Technical Results

Lemma A.15. *Let P, Q be any probability measures and let $X, X', X_1, ...X_n \sim P$ and $Y, Y', Y_1, ..., Y_n \sim Q$ be jointly independent. Let $h : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ such that $h(x,x) = 0$ for every $x \in \mathbb{R}^d$. Then*

$$\begin{aligned}&\mathbb{E}\Bigg|\frac{1}{n^2}\sum_{i,j=1}^n[2h(X_i,Y_j) - h(X_i,X_j) - h(Y_i,Y_j)]\\ &\qquad\qquad - \mathbb{E}[2h(X,Y) - h(X,X') - h(Y,Y')]\Bigg|\\ &\le 6\big\|h\big\|_\infty\sqrt{\frac{\pi}{n}}.\end{aligned}$$

Proof. The proof leverages the bounded differences inequality [63] for the function

$$\begin{aligned}&H_h(x_1,...,x_n,y_1,...,y_n)\\ &= \frac{1}{n^2}\sum_{i,j=1}^n 2h(x_i,y_j) - h(x_i,x_j) - h(y_i,y_j).\end{aligned}$$

This function satisfies the bounded differences property for each variable x_i:

$$\begin{aligned}&|H_h(x_1,x_2,...,x_n,y_1,...,y_n) - H_h(x_1',x_2,...,x_n,y_1,...,y_n)|\\ &= \Bigg|\frac{2}{n^2}\sum_{i=1}^n[h(x_1,y_i) - h(x_1',y_i)]\\ &\qquad\qquad - \frac{1}{n^2}\sum_{i=2}^n[h(x_1,x_i) - h(x_1',x_i)]\Bigg|\\ &\le \frac{2}{n^2}\sum_{i=1}^n[|h(x_1,y_i)| + |h(x_1',y_i)|]\\ &+ \frac{1}{n^2}\sum_{i=2}^n[|h(x_1,x_i)| + |h(x_1',x_i)|]\\ &\le \frac{2}{n^2}n2\big\|h\big\|_\infty + \frac{1}{n^2}(n-1)2\big\|h\big\|_\infty\\ &\le \frac{6}{n}\big\|h\big\|_\infty\end{aligned}$$

where we have used that $h(x_1,x_1) = h(x_1',x_1') = 0$ by the assumptions on h. An analogous computation holds for every other x_i and y_i. Next note that

$$\begin{aligned}&\mathbb{E}H_h(X_1,...,X_n,Y_1,...,Y_n)\\ &= \mathbb{E}\left[\frac{1}{n^2}\sum_{i,j=1}^n 2h(X_i,Y_j) - h(X_i,X_j) - h(Y_i,Y_j)\right]\\ &= \mathbb{E}\left[2h(X,Y) - h(X,X') - h(Y,Y')\right]\end{aligned}$$

Therefore by the bounded differences inequality we have

$$\begin{aligned}&\mathbb{P}\Bigg\{\Bigg|\frac{1}{n^2}\sum_{i,j=1}^n[2h(X_i,Y_j) - h(X_i,X_j) - h(Y_i,Y_j)]\\ &\qquad - \mathbb{E}[2h(X,Y) - h(X,X') - h(Y,Y')]\Bigg| > t\Bigg\}\\ &= \mathbb{P}\Bigg\{\Bigg|H_h(X_1,...,X_n,Y_1,...,Y_n)\\ &\qquad - \mathbb{E}H_h(X_1,...,X_n,Y_1,...,Y_n)\Bigg| > t\Bigg\}\\ &\le 2\exp\left(\frac{-2t^2}{2n(6\big\|h\big\|_\infty/n)^2}\right)\\ &= 2\exp\left(\frac{-nt^2}{36\big\|h\big\|_\infty^2}\right)\end{aligned}$$

Now using the tail bound form of expectation we have

$$\begin{aligned}
&\mathbb{E}\bigg|\frac{1}{n^2}\sum_{i,j=1}^{n}[2h(X_i,Y_j)-h(X_i,X_j)-h(Y_i,Y_j)] \\
&\qquad -\mathbb{E}[2h(X,Y)-h(X,X')-h(Y,Y')]\bigg| \\
&=\int_0^\infty \mathbb{P}\bigg\{\bigg|\frac{1}{n^2}\sum_{i,j=1}^{n}[2h(X_i,Y_j)-h(X_i,X_j)-h(Y_i,Y_j)] \\
&\qquad -\mathbb{E}[2h(X,Y)-h(X,X')-h(Y,Y')]\bigg|>t\bigg\}dt \\
&\leq 2\int_0^\infty \exp\left(\frac{-nt^2}{36\|h\|_\infty^2}\right)dt \\
&=6\|h\|_\infty\sqrt{\frac{\pi}{n}}.
\end{aligned}$$

□

References

[1] T. He, S. Ben-David, and L. Tong, "Nonparametric change detection and estimation in large-scale sensor networks," *IEEE Transactions on Signal Processing*, vol. 54, no. 4, pp. 1204–1217, 2006.

[2] J. Reeves, J. Chen, X. L. Wang, R. Lund, and Q. Q. Lu, "A review and comparison of changepoint detection techniques for climate data," *Journal of applied meteorology and climatology*, vol. 46, no. 6, pp. 900–915, 2007.

[3] R. P. Adams and D. J. MacKay, "Bayesian online changepoint detection," *arXiv preprint arXiv:0710.3742*, 2007.

[4] J.-P. Qi, Q. Zhang, Y. Zhu, and J. Qi, "A novel method for fast change-point detection on simulated time series and electrocardiogram data," *PloS one*, vol. 9, no. 4, p. e93365, 2014.

[5] K. C. Cheng, S. Aeron, M. C. Hughes, E. Hussey, and E. L. Miller, "Optimal transport based change point detection and time series segment clustering," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2020, pp. 6034–6038.

[6] K. C. Cheng, E. L. Miller, M. C. Hughes, and S. Aeron, "On matched filtering for statistical change point detection," *IEEE Open Journal of Signal Processing*, vol. 1, pp. 159–176, 2020.

[7] J. Damjanovic, J. M. Murphy, and Y.-S. Lin, "Catboss: Cluster analysis of trajectories based on segment splitting," *Journal of Chemical Information and Modeling*, vol. 61, no. 10, pp. 5066–5081, 2021.

[8] J. Damjanovic, Y.-S. Lin, and J. M. Murphy, "Modeling changes in molecular dynamics time series as Wasserstein barycentric interpolations," in *2023 International Conference on Sampling Theory and Applications (SampTA)*. IEEE, 2023, pp. 1–7.

[9] M. Hallin, "Measure transportation and statistical decision theory," *Annual Review of Statistics and Its Application*, vol. 9, no. 1, pp. 401–424, 2022.

[10] N. Deb and B. Sen, "Multivariate rank-based distribution-free nonparametric testing using measure transportation," *Journal of the American Statistical Association*, pp. 1–16, 2021.

[11] S. B. Masud, M. Werenski, J. M. Murphy, and S. Aeron, "Multivariate soft rank via entropy-regularized optimal transport: Sample efficiency and generative modeling," *Journal of Machine Learning Research*, vol. 24, no. 160, pp. 1–65, 2023.

[12] N. Deb, B. B. Bhattacharya, and B. Sen, "Efficiency lower bounds for distribution-free hotelling-type two-sample tests based on optimal transport," *arXiv preprint arXiv:2104.01986*, 2021.

[13] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola, "A kernel two-sample test," *The Journal of Machine Learning Research*, vol. 13, no. 1, pp. 723–773, 2012.

[14] A. Ramdas, N. G. Trillos, and M. Cuturi, "On Wasserstein two-sample testing and related families of nonparametric tests," *Entropy*, vol. 19, no. 2, p. 47, 2017.

[15] J. Feydy, T. Séjourné, F.-X. Vialard, S.-i. Amari, A. Trouvé, and G. Peyré, "Interpolating between optimal transport and MMD using Sinkhorn divergences," in *The 22nd International Conference on Artificial Intelligence and Statistics*. PMLR, 2019, pp. 2681–2690.

[16] N. Ahad, E. L. Dyer, K. B. Hengen, Y. Xie, and M. A. Davenport, "Learning Sinkhorn divergences for supervised change point detection," *arXiv preprint arXiv:2202.04000*, 2022.

[17] S. Li, Y. Xie, H. Dai, and L. Song, "Scan b-statistic for kernel change-point detection," *arXiv preprint arXiv:1507.01279*, 2015.

[18] S. Aminikhanghahi and D. J. Cook, "A survey of methods for time series change point detection," *Knowledge and information systems*, vol. 51, no. 2, pp. 339–367, 2017.

[19] C. Truong, L. Oudre, and N. Vayatis, "Selective review of offline change point detection methods," *Signal Processing*, vol. 167, p. 107299, 2020.

[20] K. Cheng, S. Aeron, M. C. Hughes, and E. L. Miller, "Dynamical Wasserstein barycenters for time-series modeling," *Advances in Neural Information Processing Systems*, vol. 34, pp. 27 991–28 003, 2021.

[21] W.-C. Chang, C.-L. Li, Y. Yang, and B. Póczos, "Kernel change-point detection with auxiliary deep generative models," *arXiv preprint arXiv:1901.06077*, 2019.

[22] J. Bai, R. L. Lumsdaine, and J. H. Stock, "Testing for and dating common breaks in multivariate time series," *The Review of Economic Studies*, vol. 65, no. 3, pp. 395–432, 1998.

[23] J. Bai, "Likelihood ratio tests for multiple structural changes," *Journal of econometrics*, vol. 91, no. 2, pp. 299–323, 1999.

[24] D. Wang, Y. Yu, and A. Rinaldo, "Univariate mean change point detection: Penalization, cusum and optimality," *Electronic Journal of Statistics*, vol. 14, no. 1, 2020.

[25] D. S. Matteson and N. A. James, "A nonparametric approach for multiple change point analysis of multivariate data," *Journal of the American Statistical Association*, vol. 109, no. 505, pp. 334–345, 2014.

[26] F. Chamroukhi, S. Mohammed, D. Trabelsi, L. Oukhellou, and Y. Amirat, "Joint segmentation of multivariate time series with hidden process regression for human activity recognition," *Neurocomputing*, vol. 120, pp. 633–644, 2013.

[27] W.-H. Lee, J. Ortiz, B. Ko, and R. Lee, "Time series segmentation through automatic feature learning," *arXiv preprint arXiv:1801.05394*, 2018.

[28] X. Yu and Y. Cheng, "A comprehensive review and comparison of cusum and change-point-analysis methods to detect test speededness," *Multivariate Behavioral Research*, vol. 57, no. 1, pp. 112–133, 2022.

[29] D. Siegmund and E. Venkatraman, "Using the generalized likelihood ratio statistic for sequential detection of a change-point," *The Annals of Statistics*, pp. 255–271, 1995.

[30] P. Fryzlewicz, "Wild binary segmentation for multiple change-point detection," *The Annals of Statistics*, vol. 42, no. 6, pp. 2243–2281, 2014.

[31] A. Kolmogorov, "Sulla determinazione empirica di una lgge di distribuzione," *Inst. Ital. Attuari, Giorn.*, vol. 4, pp. 83–91, 1933.

[32] N. V. Smirnov, "On the estimation of the discrepancy between empirical curves of distribution for two independent samples," *Bull. Math. Univ. Moscou*, vol. 2, no. 2, pp. 3–14, 1939.

[33] F. J. Massey Jr, "The Kolmogorov-Smirnov test for goodness of fit," *Journal of the American statistical Association*, vol. 46, no. 253, pp. 68–78, 1951.

[34] H. Cramér, *On the composition of elementary errors: Statistical applications*. Almqvist and Wiksell, 1928.

[35] R. v. Mises, *Probability, Statistics, and Truth*. Springer-Verlag, 2013, vol. 7.

[36] T. W. Anderson, "On the distribution of the two-sample Cramer-von Mises criterion," *The Annals of Mathematical Statistics*, pp. 1148–1159, 1962.

[37] D. M. Hawkins and Q. Deng, "A nonparametric change-point control chart," *Journal of Quality Technology*, vol. 42, no. 2, pp. 165–173, 2010.

[38] B. K. Sriperumbudur, K. Fukumizu, A. Gretton, B. Schölkopf, and G. R. Lanckriet, "On the empirical estimation of integral probability metrics," *Electronic Journal of Statistics*, vol. 6, pp. 1550–1599, 2012.

[39] F. Santambrogio, "Optimal transport for applied mathematicians," *Birkäuser, NY*, vol. 55, no. 58-63, p. 94, 2015.

[40] M. Hallin, "On distribution and quantile functions, ranks and signs in $\mathbb{R}^d$," *ECARES Working Papers*, 2017.

[41] V. Chernozhukov, A. Galichon, M. Hallin, and M. Henry, "Monge–Kantorovich depth, quantiles, ranks and signs," *Annals of Statistics*, vol. 45, no. 1, pp. 223–256, 2017.

[42] Y. Brenier, "Polar factorization and monotone rearrangement of vector-valued functions," *Communications on Pure and Applied Mathematics*, vol. 44, no. 4, pp. 375–417, 1991.

[43] R. J. McCann, "Existence and uniqueness of monotone measure-preserving maps," *Duke Mathematical Journal*, vol. 80, no. 2, pp. 309–324, 1995.

[44] R. T. Rockafellar, *Convex analysis*. Princeton university press, 1970, vol. 18.

[45] M. Cuturi, "Sinkhorn distances: Lightspeed computation of optimal transport," *Advances in neural information processing systems*, vol. 26, pp. 2292–2300, 2013.

[46] G. Peyré and M. Cuturi, "Computational optimal transport: With applications to data science," *Foundations and Trends® in Machine Learning*, vol. 11, no. 5-6, pp. 355–607, 2019.

[47] A. Genevay, M. Cuturi, G. Peyré, and F. Bach, "Stochastic optimization for large-scale optimal transport," *Advances in neural information processing systems*, vol. 29, 2016.

[48] M. Cuturi, O. Teboul, and J.-P. Vert, "Differentiable ranks and sorting using optimal transport," in *Advances in neural information processing systems*, vol. 32, 2019.

[49] S. Chewi and A.-A. Pooladian, "An entropic generalization of caffarelli's contraction theorem via covariance inequalities," *Comptes Rendus. Mathématique*, vol. 361, no. G9, pp. 1471–1482, 2023.

[50] J. Altschuler, J. Niles-Weed, and P. Rigollet, "Near-linear time approximation algorithms for optimal transport via Sinkhorn iteration," in *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[51] T. Lin, N. Ho, and M. I. Jordan, "On the efficiency of entropic regularized algorithms for optimal transport," *Journal of Machine Learning Research*, vol. 23, no. 137, pp. 1–42, 2022.

[52] E. Bernton, P. Ghosal, and M. Nutz, "Entropic optimal transport: Geometry and large deviations," *Duke Mathematical Journal*, vol. 171, no. 16, pp. 3363–3400, 2022.

[53] G. Carlier, P. Pegon, and L. Tamanini, "Convergence rate of general entropic optimal transport costs," *Calculus of Variations and Partial Differential Equations*, vol. 62, no. 4, p. 116, 2023.

[54] A.-A. Pooladian and J. Niles-Weed, "Entropic estimation of optimal transport maps," *arXiv preprint arXiv:2109.12004*, 2021.

[55] J.-C. Hütter and P. Rigollet, "Minimax estimation of smooth optimal transport maps," *The Annals of Statistics*, vol. 49, no. 2, pp. 1166 – 1194, 2021.

[56] P. Rigollet and A. J. Stromme, "On the sample complexity of entropic optimal transport," *arXiv preprint arXiv:2206.13472*, 2022.

[57] G. Mena and J. Niles-Weed, "Statistical bounds for entropic optimal transport: Sample complexity and the central limit theorem," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[58] M. Werenski, J. M. Murphy, and S. Aeron, "Estimation of entropy-regularized optimal transport maps between non-compactly supported measures," *arXiv preprint arXiv:2311.11934*, 2023.

[59] Z. Goldfeld, K. Kato, G. Rioux, and R. Sadhu, "Limit theorems for entropic optimal transport maps and the Sinkhorn divergence," *arXiv preprint arXiv:2207.08683*, 2022.

[60] W. Wang, D. Slepčev, S. Basu, J. A. Ozolek, and G. K. Rohde, "A linear optimal transportation framework for quantifying and visualizing variations in sets of images," *International journal of computer vision*, vol. 101, no. 2, pp. 254–269, 2013.

[61] S. D. Marino and A. Gerolin, "An optimal transport approach for the schrödinger bridge problem and convergence of Sinkhorn algorithm," *Journal of Scientific Computing*, vol. 85, no. 2, pp. 1–28, 2020.

[62] H. Karimi, J. Nutini, and M. Schmidt, "Linear convergence of gradient and proximal-gradient methods under the Polyak-łojasiewicz condition," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 2016, pp. 795–811.

[63] C. McDiarmid, "On the method of bounded differences," *Surveys in Combinatorics*, vol. 141, no. 1, pp. 148–188, 1989.

Matthew Werenski received the B.S degree in computer science and mathematics from the University of Edinburgh in 2019, an M.S. degree in Computer Science from Tufts University in 2021. He has been working as a graduate research assistant in the computer science department at Tufts University since 2019. His research interests include machine learning, high-dimensional probability, empirical process theory, optimal transport, and statistics.

Shoaib Bin Masud received the B.S. degree from Bangladesh University of Engineering and Technology in 2016, the M.S. degree from Tufts University in 2021, and the Ph.D. degree from Tufts University in 2023, all in electrical and computer engineering. From 2016-2018 he worked as an assistant engineer in Bangladesh. From 2018-2023 he worked as a graduate research assistant at Tufts University. His research interests include generative AI, time-series analysis, feature selection, machine learning, and signal processing.

James M. Murphy received the B.S. degree in mathematics from the University of Chicago in 2011 and the Ph.D. degree in mathematics from the University of Maryland, College Park in 2015. He was an assistant research professor with the Department of Mathematics at Duke University from 2015 to 2016 and was an assistant research scientist and a senior lecturer with the Department of Mathematics at Johns Hopkins University from 2016 to 2018. He is currently an assistant professor in the Department of Mathematics at Tufts University, with secondary appointments in the Departments of Computer Science and Electrical and Computer Engineering. His research interests include applied harmonic analysis, machine & statistical learning, high-dimensional statistics, and image, signal, & network processing.

Shuchin Aeron is an associate professor in the Department of Electrical and Computer Engineering at Tufts School of Engineering. He received his Ph.D. from Boston University in 2009 where he received the best thesis awards from both the department of ECE and from the School of Engineering. He is a recipient of Dean's fellowship and a Schlumberger-Doll research grant in support of his PhD research. From 2009-2022 he was a postdoctoral research fellow at Schlumberger-Doll Research (SDR), where he worked on signal processing solution products for borehole acoustics leading to multiple patents. In 2016, he received the NSF CAREER award supporting his research in tensor algebraic methods for data analytics. He was a visiting faculty at Mitsubishi Electric Research Labs (MERL) in 2019. He is currently a senior member of the Institute of Electrical and Electronics Engineers (IEEE), a TC member of IEEE MLSP, and was an associate editor for IEEE Transactions on Geosciences and Remote Sensing (TGRS) from 2018-2023. His research interests are in statistical signal processing, information theory, tensor data analytics, high-dimensional probability, statistical learning, and optimal transport.