# I-SEA: Importance Sampling and Expected Alignment-Based Deep Distance Metric Learning for Time Series Analysis and Embedding

**Sirisha Rambhatla**[1], **Zhengping Che**[2], **Yan Liu**[3]

[1] Management Sciences Department, University of Waterloo, Waterloo, ON, Canada
[2] AI Innovation Center, Midea Group, Beijing, China
[3] Computer Science Department, University of Southern California, Los Angeles, CA, U.S.A.
sirisha.rambhatla@uwaterloo.ca, chezp@midea.com, yanliu.cs@usc.edu

**Abstract**

Learning effective embeddings for potentially irregularly sampled time-series, evolving at different time scales, is fundamental for machine learning tasks such as classification and clustering. Task-dependent embeddings rely on similarities between data samples to learn effective geometries. However, many popular time-series similarity measures are not valid distance metrics, and as a result they do not reliably capture the intricate relationships between the multi-variate time-series data samples for learning effective embeddings. One of the primary ways to formulate an accurate distance metric is by forming distance estimates via Monte-Carlo-based expectation evaluations. However, the high-dimensionality of the underlying distribution, and the inability to sample from it, pose significant challenges. To this end, we develop an Importance Sampling based distance metric – I-SEA – which enjoys the properties of a metric while consistently achieving superior performance for machine learning tasks such as classification and representation learning. I-SEA leverages Importance Sampling and Non-parametric Density Estimation to adaptively estimate distances, enabling implicit estimation from the underlying high-dimensional distribution, resulting in improved accuracy and reduced variance. We theoretically establish the properties of I-SEA and demonstrate its capabilities via experimental evaluations on real-world healthcare datasets.

## 1 Introduction

Learning to embed time-series is at the heart of a number of machine learning tasks such as classification (Hayashi, Mizuhara, and Suematsu 2005), clustering (Ma et al. 2019), forecasting (Murray 1993), recommendation, search and retrieval (McFee, Barrington, and Lanckriet 2012; Oord, Dieleman, and Schrauwen 2013). Further, representations which encode the geometry of the data are also fundamental for other data mining and information retrieval tasks with applications in healthcare (Yang and Shahabi 2004; Xiong and Chen 2006; Saigo, Vert, and Akutsu 2006; Yassine, Singh, and Alamri 2017), music retrieval (McFee, Barrington, and Lanckriet 2012; Oord, Dieleman, and Schrauwen 2013), speech processing (Sakoe and Chiba 1978; Myers, Rabiner, and Rosenberg 1980; Chorowski et al. 2015), human activity understanding (Tran and Sorokin 2008; Jiang, Jr, and Gonzalez 2012), meteorology and climate (Lhermitte et al. 2011; Baranowski et al. 2015).

Measuring distances between data samples is key for faithfully encoding the geometries for embedding. To this end, time-series are compared based on their similarity under a certain monotonic and non-decreasing arrangement, or *alignment* (Sakoe and Chiba 1978). Notwithstanding their success, popular methods rely on an optimal alignment, which prevents them from constituting a *valid distance metric* (Müller 2007; Mei et al. 2015; Cuturi and Blondel 2017), an essential property for reliably capturing the geometry via pair-wise distances (Cover and Hart 1967; Cox and Cox 2008).

Moreover, *time-series metric learning*, which learns supervised task-dependent embeddings via linear or non-linear (deep learning-based) transformations to capture complex temporal relationships among the features in multi-variate time-series (Xing et al. 2002; Salakhutdinov and Hinton 2007; Weinberger and Saul 2009; Hoffer and Ailon 2015), also critically relies on effective pair-wise distance comparisons to learn embeddings such that the distances in the transformed space reflect the nearest neighbor properties imposed by the supervision (Shalev-Shwartz, Singer, and Ng 2004). However, since popular ways to compare time-series do not constitute a *distance metric*, the learned representations or embeddings based on such measures also do not encode the complex relationships between the data samples (Cover and Hart 1967; Cox and Cox 2008). This situation is further exacerbated by irregular sampling, missing entries and other non-idealities. As a result, developing reliable distance metrics for time-series remains a challenging problem.

One way to form a valid distance metric is by averaging distances over all possible *alignment paths* (Cuturi et al. 2007; Che et al. 2017). However, these averages (*Expected Alignments*) are difficult to compute accurately since the computations involve expectation evaluations w.r.t a high-dimensional distribution over *all* alignment paths. This problem is further compounded by a) the inability to sample from this distribution due to the combinatorial nature of the problem (even when *known*), and b) the distribution being over rare events (since good alignments are rare).

To address these challenges, we propose I-SEA: Importance Sampling and Expected Alignment-based distance metric for comparing time-series, which leverages a) deep learning-based representations to capture complex temporal

feature dependencies, and b) Non-parametric Density Estimation and Weighted Importance Sampling for accurate distance estimation, to learn effective embeddings of multivariate time-series. Our specific contributions are as follows:

- **Importance Sampling-based data-driven distance metric for time-series**. We develop a deep learning and Importance Sampling-based distance metric which learns a task-dependent metric using a large margin-based triplet loss metric learning approach. We establish the theoretical properties of I-SEA, showing that it is a valid distance metric for comparing time-series.
- **Improved distance estimation via Importance Sampling**. Adopting a rare event distribution view of the similarity distribution over alignment paths, we develop a Weighted Importance Sampling-based approach, which utilizes Adaptive Non-parametric Density Estimation and Rejection Sampling to enable implicit estimation from an inherently high-dimensional distribution. The resulting metrics are accurate and exhibit reduced variance properties across different datasets.
- **Learning faithful embeddings**. Our neural network-based representations and distance estimation effectively encode the relationships between time-series. Furthermore, our estimation procedure shows low variance, while conventional Importance Sampling estimators are known to result in high variance if the distribution over desired region has a small support (Precup 2000).

A key contribution of our work is to enable accurate distance computations by implicit estimation from a distribution over all alignment paths using Importance Sampling. The primary challenge here, in addition to the high-dimension of the distribution, is that as opposed to conventional Importance Sampling, sampling over time-series involves two distributions – one over the alignment path lengths, and other over all alignment paths of a specific length. Although both are *a priori* unknown, our main result leverages the fundamental differences between these two to develop Importance Sampling-based metrics for time-series. As a result, our contributions here provide, to the best of our knowledge, the first work to tackle and leverage these high-dimensional distributions to develop a metric for time-series, and the techniques developed here can be of independent interest.

### 1.1 Related Works

**Comparing Time-series**. Classical similarity measures such as dynamic time warping (DTW) (Müller 2007) and multiple sequence alignment (MSA) (Hogeweg and Hesper 1984), rely on an alignment step before comparing the time-series, independent of the data. Global Alignment Kernel-based (GAK) methods also belong to this class albeit leverage the kernel-trick to compute alignment over all paths to develop a metric (Cuturi et al. 2007; Cuturi 2011). Application-specific measures are a popular way to compare time-series (Qiu et al. 2019), but these often do not constitute a metric. On the other hand, recent optimal transport-based metrics do not consider the sequence order (Huang et al. 2016). To mitigate this, Su and Hua (2018); Su and Wu (2019) develop a locally order-preserving variant of the Wasserstein metric. However, these works do not consider the intra-sequence relationships nor show if they constitute a valid metric; see also Shanmugam (2018).

**Deep Metric Learning**. Task-dependent metrics learn an embedding by transforming the data either linearly or non-linearly (Xing et al. 2002). These rely on alignment computations on the transformed data to develop a similarity measure. Since linear transformation-based methods (Lajugie et al. 2014; Mei et al. 2015) fail to capture the complex dependencies across features in multivariate time-series metric learning, deep learning-based metric learning has gained popularity (Hoffer and Ailon 2015), leading to gradient training amenable loss function for end-to-end training (Cuturi and Blondel 2017), and alignment-independent techniques (Mueller and Thyagarajan 2016). As a result, metric learning inherently relies on faithfully computing pairwise distances to learn an embedding (Weinberger and Saul 2009). To this end, valid distance metrics (which follow *non-negativity*, *symmetry*, and *triangle inequality*) are critical for learning effective embeddings since they can encode the relative geometry (Shalev-Shwartz, Singer, and Ng 2004). For this, recent methods average distances over all possible alignment paths (Expected Alignment) to form a valid distance metric (Che et al. 2017), treating all alignment paths as equals, while favorable alignments are rare. Instead, we leverage Weighted Importance Sampling and Non-parametric density estimation to weigh alignment paths based on their favorability to develop valid distance metrics for time series.

**Importance Sampling and Rare Event Distributions**. Importance sampling is a Monte-Carlo variance reduction method, also used to estimate expectations w.r.t. a distribution while drawing samples from a different one (Precup 2000; Rubinstein and Kroese 2016). Since choosing a sampling distribution is critical for controlling estimation error (specifically for rare event distributions), adaptive strategies leverage Monte-Carlo sampling to simultaneously estimate sampling and the target expectation for both parametric (Karamchandani, Bjerager, and Cornell 1989) and non-parametric (Zhang 1996) densities; see also Glynn and Iglehart (1989).

## 2 I-SEA: Methodology

I-SEA has two main components to address the two key challenges to develop a valid metric for multi-variate time-series: a) accurate distance computation between time-series, and b) capturing complex feature dependence structure using data-driven representations. We now detail each of these, with our overall metric learning-based architecture shown in Fig. 1 and notations summarized below[1].

[1]**Notation**: We denote vectors and matrices by bold lower $\mathbf{x}$ and capital $\mathbf{X}$ case letters, respectively. $\mathbb{E}[\cdot]$ denotes the expectation operator. $\|\cdot\|$ denotes the 2-norm. $\mathcal{U}\{\cdot\}$ denotes the discrete uniform distribution. Let $\mathbf{X} \in \mathbb{R}^{n \times T_X}$ be a multi-variate time-series where $n$ denotes the number of variates and $T_X$ denotes the number of time steps. For a time-series $\mathbf{X}$, $\mathbf{X}_{\boldsymbol{\alpha}_U} \in \mathbb{R}^{n \times U}$ denotes the arrangement of columns of $\mathbf{X}$ according to indices of an integer-valued vector $\boldsymbol{\alpha}_U \in \mathbb{R}^U$ (which can have repeated entries). We use $\mathbf{X}_{\boldsymbol{\alpha}_U}(t) \in \mathbb{R}^n$ for $t$-th column of $\mathbf{X}_{\boldsymbol{\alpha}_U}$. For a set of samples $\mathcal{S}$, we use $\mathcal{S}_i^+$ to denote the set of all in-class samples, and $\mathcal{S}_i^-$ to denote all out-of-class

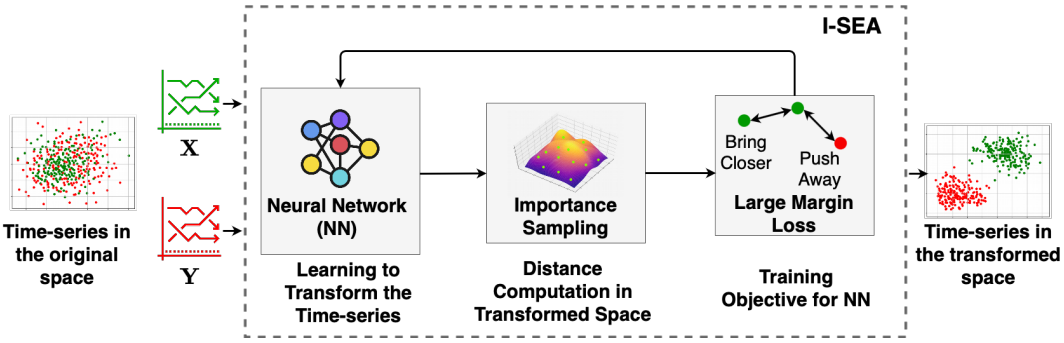



Figure 1: I-SEA: Importance Sampling and Expected Alignment-based Deep Metric Learning. We adopt a Triplet Loss-based large margin approach to train the neural network. We accomplish distance computations via Importance Sampling-based Expected Alignment for the training loss, resulting in an effective metric learning framework for time-series.

## 2.1 Distance computations via Importance Sampling and Expected Alignment

Time-series evolving at different time scales are compared via a process called *alignment* (Sakoe and Chiba 1978). Alignment finds correspondences between two time-series by selecting the entries of one w.r.t. another in a monotonically non-decreasing fashion. Formally, an alignment path $A_U := (\boldsymbol{\alpha}_U, \boldsymbol{\beta}_U)$ for multi-variate time-series $\mathbf{X} \in \mathbb{R}^{n \times T_X}$ and $\mathbf{Y} \in \mathbb{R}^{n \times T_Y}$ is defined as follows.

**Definition 2.1.** *An alignment path $A_U$ between two time-series is defined as a pair of monotonically non-decreasing sequences $(\boldsymbol{\alpha}_U, \boldsymbol{\beta}_U)$, where $\boldsymbol{\alpha}_U, \boldsymbol{\beta}_U \in \mathbb{R}^U$.*

Here, the sequences $\boldsymbol{\alpha}_U$ and $\boldsymbol{\beta}_U$ denote the indices chosen from the time-series $\mathbf{X}$ and $\mathbf{Y}$, respectively. With this, the distance over an alignment path between two multi-variate time-series $\mathbf{X} \in \mathbb{R}^{n \times T_X}$ and $\mathbf{Y} \in \mathbb{R}^{n \times T_Y}$ for an alignment path $A_U := (\boldsymbol{\alpha}_U, \boldsymbol{\beta}_U)$ of length $U$ is formalized as follows.

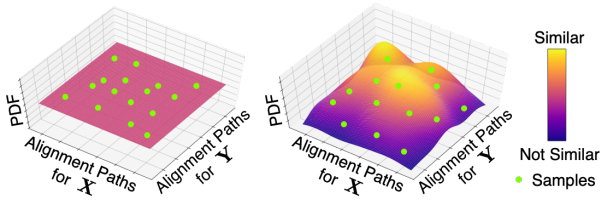
**Definition 2.2.** *For a distance metric $d(\cdot)$, the distance $D_{A_U}^{(\mathbf{X},\mathbf{Y})}$ between multivariate time-series $\mathbf{X} \in \mathbb{R}^{n \times T_X}$ and $\mathbf{Y} \in \mathbb{R}^{n \times T_Y}$ under the alignment path $A_U$ is defined as*

$$D_{A_U}^{(\mathbf{X},\mathbf{Y})} = D_{\boldsymbol{\alpha}_U, \boldsymbol{\beta}_U}^{(\mathbf{X},\mathbf{Y})} := \sum_{t=1}^{U} d(\mathbf{X}_{\boldsymbol{\alpha}_U}(t), \mathbf{Y}_{\boldsymbol{\beta}_U}(t)). \quad (1)$$

Here, we use the distance metric $d(\cdot)$ (say Euclidean distance) to compute the *local* distance between vectors $\mathbf{X}_{\boldsymbol{\alpha}_U}(t)$ and $\mathbf{Y}_{\boldsymbol{\beta}_U}(t)$ in $\mathbb{R}^n$ for a given alignment path $A_U$.

**Expectation w.r.t a High-dimensional Distribution**. Equipped with a metric over an alignment path, averaging across distances over alignment paths (*Expected Alignment*) leads to a valid distance metric (Cuturi et al. 2007; Cuturi 2011; Che et al. 2017), as opposed to that over a single path (Sakoe and Chiba 1978). Our crucial observation is that a naive averaging by considering all alignment paths to be of equal importance may not be accurate. In other words, two time-series may have higher similarity only along a few alignment paths, i.e. they may be *rare*. As a result, a naive averaging may lead to inaccurate estimates by being agnostic to the underlying similarity structure, as shown in Fig. 2(a).

samples w.r.t $i$. We use $\mathrm{Tr}(\cdot)$ to denote the trace operator and $\|\cdot\|_{\mathrm{F}}$ for the Frobenius norm.




(a) (b)

Figure 2: Sampling Alignment Paths. Panel (a) and (b) show path-agnostic, and Importance Sampling-based Expected Alignment (I-SEA), respectively. Here, I-SEA leverages similarity between sampled paths for an accurate distance metric.

**Importance Sampling**. Formally, let $p(\cdot)$ denote the distribution over all alignment paths $\mathcal{A}_U$ of length $U$, where the distribution has higher weight for *better* alignment paths, shown in Fig. 2(b). It would indeed be ideal if we could sample from a distribution $p(\cdot)$ which reflects the similarity structure (or distances) over the alignment paths between the time-series. However, since these alignment paths and the corresponding distances are unknown *a priori*, (and $p(\cdot)$ is unknown, in general) we cannot sample directly from $p(\cdot)$. We can, however, sample uniformly over all alignment paths, say a distribution $q(\cdot)$, and assess the *goodness* of a path $A_U$ *after* observing it, and use these *scores* to adjust the distance estimate. To this end, we leverage Importance Sampling (Precup 2000; Rubinstein and Kroese 2016) to evaluate the expectation w.r.t. $p(\cdot)$, while drawing samples using $q(\cdot)$.

**Significance**. I-SEA can leverage any type of Importance Sampling approach depending upon the application. We here present a Weighted Importance Sampling and adaptive Non-parametric Density Estimation-based approach specifically suited for the case when favorable alignments are rare. Our analysis can be of independent interest for Importance Sampling for time-series data. In our discussion, we use cosine similarity-based scores $s(A_U) \in [0, 1]$ to assess an alignment path (shown below).

$$s(A_U) := \tfrac{(1+\cos\theta)}{2}, \text{ where } \cos\theta = \tfrac{\mathrm{Tr}(\mathbf{Y}_{\boldsymbol{\beta}_U}^\top \mathbf{X}_{\boldsymbol{\alpha}_U})}{\|\mathbf{X}_{\boldsymbol{\alpha}_U}\|_{\mathrm{F}} \|\mathbf{Y}_{\boldsymbol{\beta}_U}\|_{\mathrm{F}}}. \quad (2)$$

Cosine-similarity is a popular choice for computing pairwise similarity between matrices. It is independent of the length of the time series since it focuses on the angle between

two vectors (Schütze, Manning, and Raghavan 2008). Based on the time-series, other choices may include linear, polynomial, sigmoid (and soft-max), Radial Basis Function (RBF), and Laplacian scoring functions (Zhang et al. 2007).

**Non-parametric Density Estimation and Weighted Importance Sampling**. The primary challenge here is the high-dimensionality of $p(\cdot)$, which is a distribution over all alignment paths between two multi-variate time-series. There are two sources of randomness here: the choice over a) the alignment path lengths $U$, and b) all alignment paths of length $U$, denoted by $\mathcal{A}_U$. To this end, we leverage Non-parametric Adaptive Kernel Density Estimation to estimate the distribution over alignment lengths $U \sim f(U)$, and Weighted Importance Sampling strategy for the distribution over alignment paths. Overall, the I-SEA $D_{\text{I-SEA}}$ is defined as follows.

**Definition 2.3.** *For alignment path length $U \sim f(U)$ supported on $\{U_l, U_h\}$, and a corresponding alignment path $A_U \sim p(A_U)$, for a distribution $p(\cdot)$ over alignment paths $\mathcal{A}_U$ of length $U$, the Importance Sampling and Expected Alignment-based distance metric, $D_{\text{I-SEA}}$ is defined as*

$$D_{\text{I-SEA}}(\mathbf{X}, \mathbf{Y}) := \mathbb{E}_U \left[ \mathbb{E}_{A_U} \left[ D_{A_U}^{(\mathbf{X},\mathbf{Y})} \right] \right] = \sum_{U \in \{U_l, U_h\}} f(U) \sum_{A_U \in \mathcal{A}_U} D_{A_U}^{(\mathbf{X},\mathbf{Y})} p(A_U), \quad (3)$$

*where $D_{A_U}^{(\mathbf{X},\mathbf{Y})}$ denotes the distance between $\mathbf{X}$ and $\mathbf{Y}$ under the alignment path $A_U$ (Def. 2.2).*

Next, Thm. 1 establishes the validity of $D_{\text{I-SEA}}$ as a distance metric, as follows (proof in Supp. A.1).

**Theorem 1.** *For a distance metric $d(\mathbf{x}, \mathbf{y})$ for $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$, the distance $D_{\text{I-SEA}}(\mathbf{X}, \mathbf{Y})$ between two multi-variate time-series $\mathbf{X} \in \mathbb{R}^{n \times T_X}$ and $\mathbf{Y} \in \mathbb{R}^{n \times T_Y}$ defined in Eq.* (3) *is a valid distance metric. Namely, it satisfies the following:*

1. *Non-negativity: $D_{\text{I-SEA}}(\mathbf{X}, \mathbf{Y}) \geq 0$,*
2. *Symmetry: $D_{\text{I-SEA}}(\mathbf{X}, \mathbf{Y}) = D_{\text{I-SEA}}(\mathbf{Y}, \mathbf{X})$, and*
3. *Triangle Inequality: $D_{\text{I-SEA}}(\mathbf{X}, \mathbf{Z}) \leq D_{\text{I-SEA}}(\mathbf{X}, \mathbf{Y}) + D_{\text{I-SEA}}(\mathbf{Y}, \mathbf{Z})$.*

**Estimating I-SEA**. Alg. 1 details the overall Importance Sampling-based procedure to form I-SEA for given time-series $\mathbf{X}$ and $\mathbf{Y}$. Although I-SEA yields a valid metric as per Thm. 1, the distributions $f(\cdot)$ and $p(\cdot)$ are both unknown *a priori*. Our key observation is that there is a difference between the properties of these two distributions. Since $f(\cdot)$ is a distribution over alignment lengths, it is a one-dimensional distribution and if it were indeed known, we can sample from it using Monte Carlo sampling techniques. However, since $p(\cdot)$ is a distribution over *all* alignment paths of a given length, it is potentially a very high-dimensional distribution, and even if it were known, we may not be able to sample from it directly. We leverage this difference to decouple these distributions to formulate our estimator as discussed below.

**Non-parametric Adaptive Density Estimation for $f(\cdot)$.** . We pose $f(\cdot)$ estimation as a Non-parametric Density Estimation problem, that we accomplish via weighted Kernel Density Estimation (KDE) by adaptively improving the estimate using $p(\cdot)$ as weights for each path with kernel $\kappa(\cdot)$ (Zhang 1996). The number of adaptive steps $k$ to update $f(\cdot)$ can be chosen based on the computation budget. To sample path length $U$ using the KDE estimate $\hat{f}_k(\cdot)$ of $f(\cdot)$, we use Rejection Sampling (Shapiro 2003) at each step.

Algorithm 1: Distance Computation for I-SEA

**Input:** Time-series $\mathbf{X} \in \mathbb{R}^{n \times T_X}$, $\mathbf{Y} \in \mathbb{R}^{n \times T_Y}$, parameters $U_l, U_h, k, m$, Kernel $\kappa(\cdot)$ with KDE bandwidth $h$, evaluated using Silverman (2018).

**Initialize:** $\hat{f}_0(\cdot) = U \sim \mathcal{U}\{U_l, U_h\}$, $q(\cdot) = \frac{1}{m} \;\; \forall i \in [m]$ (uniformly distributed).

**Output:** I-SEA distance $\hat{D}_{\text{I-SEA}}^{\text{weighted}}$ between $\mathbf{X}$ and $\mathbf{Y}$.

**for** $k = 0, 1, 2, \ldots$ **do**

**1**: Sample $m$ alignment path lengths $\{U_i\}_{i=1}^m$ from $\hat{f}_k(\cdot)$ using Rejection Sampling.

**2**: For each path length $\{U_i\}_{i=1}^m$, sample corresponding alignment paths $\{A\}_{i=1}^m$ using $q(\cdot)$.

**3**: For each alignment path $\{A\}_{i=1}^m$, calculate score using $s(\cdot)$ in Eq. (2), and set $p(A_i) = \frac{s(A_i)}{\sum_{j=1}^m s(A_j)}$.

**4**: Compute the Weighted Importance Sampling-based estimator ($D_{A_i}^{(\mathbf{X},\mathbf{Y})}$ as per Def. 2.2):

$$\hat{D}_{\text{I-SEA}}^{\text{weighted}} := \frac{1}{\sum_{i=1}^m \frac{p(A_i)}{q(A_i)}} \sum_{i=1}^m D_{A_i}^{(\mathbf{X},\mathbf{Y})} \frac{p(A_i)}{q(A_i)}. \quad (4)$$

**5**: For each $u \in \{U_l, U_h\}$, update $\hat{f}_{k+1}(\cdot)$ using Weighted KDE from samples $\{U_i\}_{i=1}^m$ with $\{p(A_i)\}_{i=1}^m$ as weights by $\hat{f}_{k+1}(u) = \frac{1}{m \cdot h} \sum_i p(A_i) \kappa\left(\frac{U_i - u}{h}\right)$.

**end for**

**Weighted Importance Sampling for $p(\cdot)$.** We leverage Importance Sampling[2] that allows us to sample the alignment paths $A_i$ according to a distribution $q(A_i)$ to form an Importance Sampling estimator $\hat{D}_{\text{I-SEA}}$ of $D_{\text{I-SEA}}$ w.r.t the target distribution $p(\cdot)$ using the estimator

$$\hat{D}_{\text{I-SEA}} := \frac{1}{m} \sum_{i=1}^m D_{A_i}^{(\mathbf{X},\mathbf{Y})} \frac{p(A_i)}{q(A_i)}. \quad (5)$$

This importance-based weighting is effective in forming an estimate where the important region is relatively small (Precup 2000). For $q(\cdot)$, we utilize an algorithm for uniformly sampling over the alignment paths presented in Alg. B.1 in Supp. B. In general, $q(\cdot)$ can be any tractable distribution. The estimate $\hat{D}_{\text{I-SEA}}$ is indeed an unbiased estimator of $D_{\text{I-SEA}}$ as established by the following lemma (proof in Supp. A.2).

**Lemma 2.** *If the alignment paths are sampled as per a distribution $q(\cdot)$, then for $U \sim f(U)$ supported on $\{U_l, U_h\}$, and $A_U \sim q(A_U)$, $D_{\text{I-SEA}}$ can be estimated as*

$$D_{\text{I-SEA}}(\mathbf{X}, \mathbf{Y}) = \mathbb{E}_U \left[ \mathbb{E}_{A_U} \left[ D_{A_U}^{(\mathbf{X},\mathbf{Y})} \frac{p(A_U)}{q(A_U)} \right] \right]. \quad (6)$$

To mitigate the case where we only sample *bad* alignment paths, the following result establishes the number of samples

[2] In practice, the *Weighted* Importance Sampling estimator shown in Eq. (4) is used for its superior stability properties (Precup 2000). Theoretical properties follow directly.

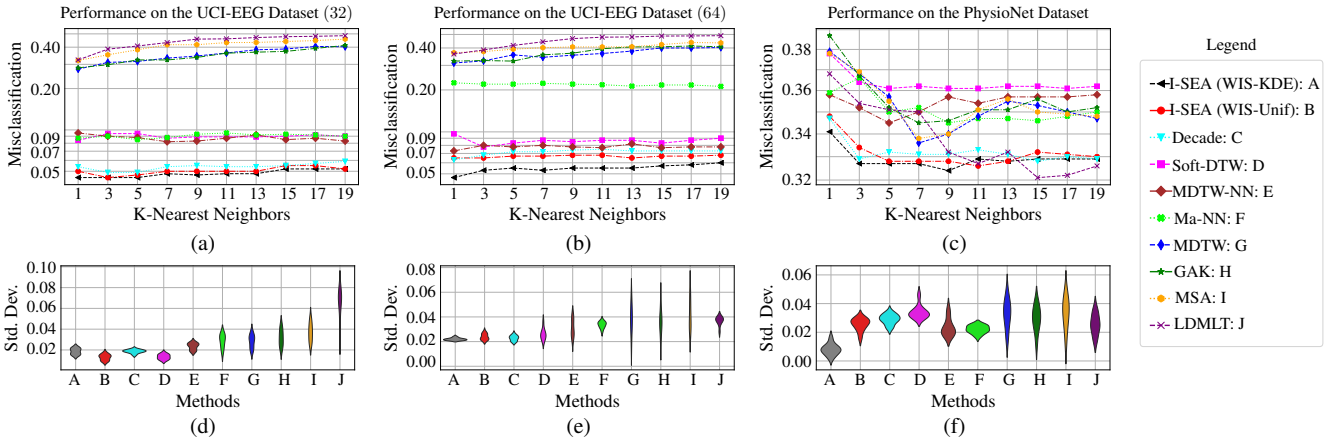



Figure 3: K-Nearest Neighbor Classification Performance. Panels (a) – (c) and (d) – (f) show the performance of I-SEA as compared to competing methods in terms of mean Misclassification ($1 - \texttt{Accuracy}$) and distribution of standard deviation across $K$ (violin plots), respectively. I-SEA consistently performs better both in terms of accuracy and lower variance. The colors in the legend also correspond to the violin plot, with I-SEA (WIS-KDE) denoted by gray. (Best viewed in color.)

required to faithfully assess the distance, by establishing that the empirical estimate $\hat{D}_{\text{I-SEA}}$ indeed concentrates around its mean $D_{\text{I-SEA}}$ (proof in Supp. A.3).

**Lemma 3.** *For any multi-variate time-series $\mathbf{X} \in \mathbb{R}^{n \times T_X}$ and $\mathbf{Y} \in \mathbb{R}^{n \times T_Y}$, where the local distance $d(\cdot)$ between the time-series under any alignment path $A_U := (\boldsymbol{\alpha}_U, \boldsymbol{\beta}_U)$ is bounded by 1, i.e.,*

$$d(\mathbf{X}_{\boldsymbol{\alpha}_U}(t), \mathbf{Y}_{\boldsymbol{\beta}_U}(t)) \leq 1 \;\; \forall t \in \{1, \ldots, U\},$$

*given $m = \Omega\left(\frac{U_h^2}{\epsilon^3}\right)$ samples, the empirical Importance Sampling-based Expected Alignment distance (I-SEA) $\hat{D}_{\text{I-SEA}}$ concentrates around its mean with probability $1-\delta$, for small constants $\delta$ and $\epsilon$, i.e.,*

$$\mathbf{P}\left[|\hat{D}_{\text{I-SEA}} - D_{\text{I-SEA}}| \leq \epsilon\right] \geq 1 - \delta.$$

**Time Complexity**. For time-series of average length $T$, and $U_h = O(T)$, I-SEA requires $\Omega(T^2)$ samples (say $c_1 T^2$) from Lem. 3, for some $c_1 > 0$. Let the distance computations for each path take $O(T)$ time. Then the overall time complexity is $O(T^3)$, which is similar to Che et al. (2017), and the constant number of KDE steps do not change the overall order. Note that the dependence on the length is a natural consequence of using concentration results used to establish sufficient conditions for the success of the estimation with high probability. As such, sampling based methods will encounter a similar dependence. Nevertheless in practice, we observe that the procedure succeeds with lower number of samples, the complexity can be controlled by restricting range of $U$, which still results in a valid distance metric.

## 2.2 Deep Metric Learning

I-SEA leverages neural network to transform the multi-variate time-series before utilizing the Importance Sampling-based distance computations (described in the previous section). To accomplish this, we train a neural network to learn an appropriate non-linear transformation for effectively embedding of the multi-variate time-series data. To this end, we adopt a triplet loss-based large margin approach (Weinberger and Saul 2009) for this exposition. Note that I-SEA can be trained with other deep metric learning loss functions; see Kaya and Ş. Bilge (2019). These loss functions rely on computing pair-wise distances to learn an embedding that captures the relationship between data samples effectively (Weinberger and Saul 2009). Since triangle inequality holds for distance metrics, training via the metric learning loss learns an effective transformation, capturing the relative geometry (Cover and Hart 1967; Shalev-Shwartz, Singer, and Ng 2004; Cox and Cox 2008; Weinberger and Saul 2009).

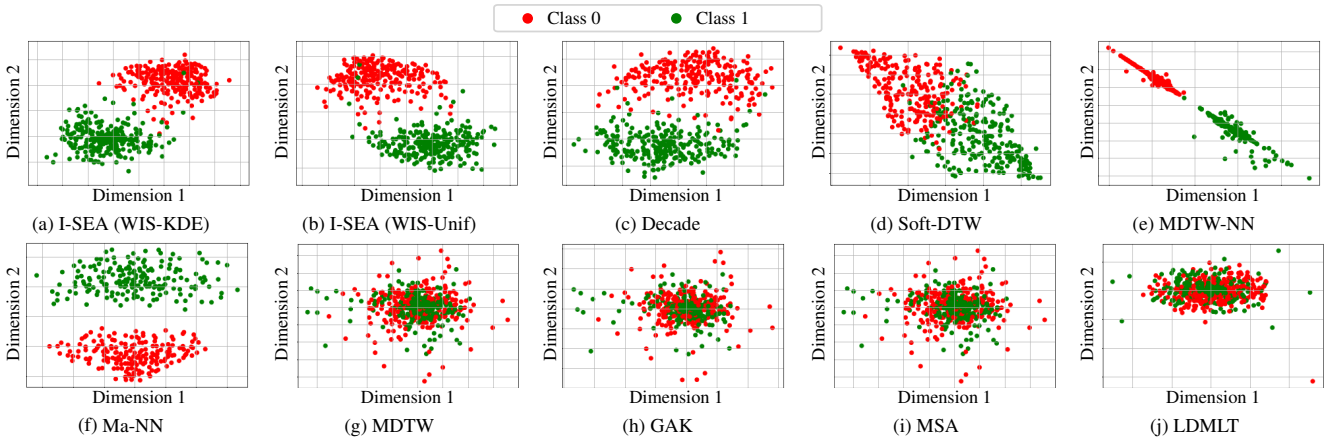
Specifically, the large margin approach aims to learn an embedding of the time-series data samples, by bringing a data sample $\mathbf{X}^{(i)}$ closer to its in-class *targets* (in the embedding space), while pushing away from the out-of-class *imposters* (using the triplet loss) (Weinberger and Saul 2009; Che et al. 2017). Formally, given $N$ samples $\{\mathbf{X}^{(i)}\}_{i=1}^{N}$ from $C$ classes, the targets $\mathcal{S}_i^+$, and the imposters $\mathcal{S}_i^-$ for the $i$-th data sample, we minimize the following objective by the choice of I-SEA's neural-network parameters in $D_{\text{I-SEA}}(\cdot)$,

$$\begin{aligned} \min_{D} \;\; & \textstyle\sum_{i \in [N], j \in \mathcal{S}_i^+} D^{(i,j)} \\ & + \lambda \textstyle\sum_{i \in [N], j \in \mathcal{S}_i^+, k \in \mathcal{S}_i^-} \max\{\delta + D^{(i,j)} - D^{(i,k)}, 0\}, \end{aligned} \tag{7}$$

where $D^{(i,j)} := D(\widetilde{\mathbf{X}}^{(i)}, \widetilde{\mathbf{X}}^{(j)})$ is the distance metric and $\widetilde{\mathbf{X}}$ denotes a non-linear transformation of $\mathbf{X}$ learned via a neural network $f_{NN}(\cdot)$, i.e., $\widetilde{\mathbf{X}} = f_{NN}(\mathbf{X})$. Here, $\lambda$ and $\delta$ control the separation between classes (*margin*).

# 3 Experimental Evaluation

We evaluate and compare the performance of I-SEA for metric learning tasks arising in two real-world healthcare time-series datasets which suffer from one of more of the following non-idealities such as missing data, irregular sampling, and a large feature set. We now describe the experiments; see Supp. C for raw data embeddings (Fig. C.1), details of the set-up, additional results, and baselines descriptions. The code is available at https://github.com/srambhatla/ISEA.




(a) I-SEA (WIS-KDE) (b) I-SEA (WIS-Unif) (c) Decade (d) Soft-DTW (e) MDTW-NN

(f) Ma-NN (g) MDTW (h) GAK (i) MSA (j) LDMLT

Figure 4: Embedding data in UCI-EEG (32) based on pair-wise distances/similarities using Multidimensional Scaling (MDS) (Cox and Cox 2008). Panels (a)-(f) show the neural network-based techniques, and panels (g)-(j) show the data-independent methods. Panel (j) shows the LDMLT-based embedding which uses data-dependent linear transformations only. (Best viewed in color).

### 3.1 Datasets

**The UCI EEG Dataset**. The UCI EEG dataset[3] consists of 600 traces collected from equal number of test subjects from alcoholics (class 0) and non-alcoholic (class 1) groups (binary classification) using 64 electrodes (sensors). Each trace is of 1 second duration sampled at 256Hz. For our experiments, we downsample these traces (by 8 and 4) resulting in data samples of length 32 "UCI-EEG (32)" and 64 "UCI-EEG (64)", respectively, with a feature-set of 64 variables.

**The PhysioNet Dataset**. The PhysioNet dataset (Goldberger et al. 2000) consists of in-hospital Intensive Care Unit (ICU) patient medical data recorded over the first 48 hours of their stay, and the eventual mortality outcome (0 or 1). We randomly sample 1108 traces to form a balanced dataset containing observations from 37 variables observed irregularly over the 48 hour period with missing entries (73%/sample), aggregating the observations for each hour[4]. The resulting samples are of length 48 with a feature-set of 37 variables.

### 3.2 Variants and Baselines

We evaluate the performance of I-SEA with popular task-dependent and independent baselines for time series analysis. We present two variants of I-SEA in the experiments to study the contribution of a) Weighted Importance Sampling (WIS) and b) Gaussian Kernel-based ($\kappa(\cdot)$) KDE for estimating $f(\cdot)$. Here, the first variant "I-SEA (WIS-Unif)" incorporates WIS only (i.e., $f(\cdot)$ is set to be Uniform distribution) and is used to analyze performance with respect to Che et al. (2017) that does not use WIS; The second variant "I-SEA (WIS-KDE)" incorporates both WIS and KDE. We employ Euclidean distance for $d(\cdot)$, and fix the hyper-parameters across all experiments; see Supp. C. We use data independent measure such as Multi-variate DTW (MDTW), Multiple Sequence Alignment (MSA) (Hogeweg and Hesper 1984), Global Alignment Kernel (GAK) (Cuturi 2011; Cuturi et al. 2007), and task-dependent measures such as Decade (Che et al. 2017), neural network-based Soft-DTW (Cuturi and Blondel 2017), neural network-based MDTW (MDTW-NN), Manhattan Neural Network (Ma-NN) (Mueller and Thyagarajan 2016) and LDMLT (LogDet divergence based Metric Learning with Triplets) (Mei et al. 2015). Of these, Decade (when Euclidean distances are used for local distances), Ma-NN, MSA, and GAK constitute a metric. For task-dependent baselines, we use large margin metric learning loss in Eq. (7). For neural network-based baselines, we use a two-layer feed-forward model with sigmoid activations to capture complex feature dependence with same input and output dimensions.

### 3.3 Evaluation Metrics

We use the K-Nearest Neighbor Mean Accuracy over the test sets corresponding to the 5 folds and its standard deviation to evaluate the embeddings learned using the techniques. Here, we evaluate the performance of the techniques for various values of $K$, i.e. $K \in \{1, 3, \ldots 19\}$. In addition, we evaluate the learned visualizations both qualitatively and quantitatively, via Multidimensional Scaling (MDS)-based 2-D projections (training) since MDS can handle metrics and non-metrics, and in terms of *triplet loss* – percent violation of triangle inequality over triplets (20k random triplets each for train and test set), respectively, using the learned distances.

### 3.4 Results

**K-Nearest Neighbor Performance**. We compare the performance of I-SEA with the baselines detailed in Supp. C.1 based on their K-Nearest Neighbor (K-NN) classification accuracy for metric learning tasks for the real-world datasets. Panels (a), (b), and (c) in Fig. 3 show the K-NN Misclassification ($1-$ `Accuracy`) performance of the baselines as compared to I-SEA for UCI-EEG (32 and 64), and the PhysioNet dataset respectively; see Supp. C.2 for detailed results. We also show the corresponding distribution of standard deviation across $K$s in the violin plots in panels (d), (e), and (f) of Fig. 3, respectively. We observe that across datasets,

[3]The UCI-EEG dataset is available at http://archive.ics.uci.edu/ml/datasets/EEG+Database.

[4]I-SEA can also handle irregularity in time-series by restricting the set of time-steps available while forming the alignment paths.

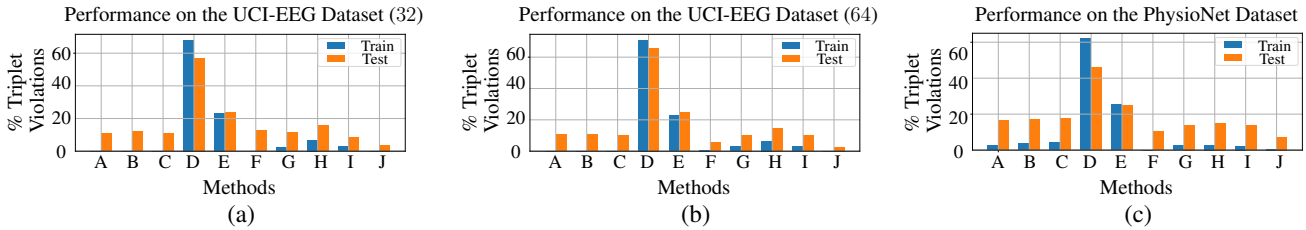



Figure 5: Percent Triplet Violations from Learned Embeddings. Panel (a), (b), and (c) show the triplet violations across different methods for UCI-EEG (32), UCI-EEG (64) and the PhysioNet dataset, respectively. Valid distance metrics score better.

I-SEA variants perform better across different choices of $K$, while also exhibiting the best variance properties, i.e., both low and consistent. This is because Importance Sampling at its core is a variance reduction method and leads to superior variance properties across datasets. This underscores the benefit of Importance Sampling for computing the Expected Alignment. These also point to learned embedding properties. Specifically, even as the number of neighbors grow, the K-NN performance for I-SEA is *stable*, indicating that the learned metric faithfully separates the in-class and out-of-class examples. This can also be observed to some extent in Decade and Ma-NN, highlighting the importance of a valid metric for time-series deep metric learning.

**Analyzing the Learned Embeddings**. In Fig. 4, we visualize the learned distances by I-SEA and the baselines using the 2-dimensional MDS embedding based on the pair-wise training set distances (over the last training fold)for the UCI-EEG (32); see Supp. C (Figs. C.2 and C.3) for other datasets. Complementary to these results, we show the percent triplet violations (violations of triangle inequality by data triplets) by each method over train and test sets for these (for the same fold) in Fig. 5. The embeddings are reveal how distance computations are used by metric learning. Specifically, we notice that all deep metric learning techniques – which use the large margin approach shown in Eq. (7) – exhibit some kind of clustering structure, while the rest of the baselines do not, highlighting the role of deep learning for metric learning.

Next, we observe that the MDTW-NN baseline for both datasets shows a clustering which seems to lie on a line. This is because although deep representations do help with separation, the pair-wise similarities cannot capture relationships between data samples. This can also be observed from Fig. 5, where we analyze triplet violations over 20k randomly chosen triplets. Generally, metrics (I-SEA, Decade, Ma-NN, GAK, MSA) perform better than similarity measures (Soft-DTW, MDTW-NN, MDTW), and among the neural network methods, valid metrics (I-SEA, Decade, Ma-NN) outperform the non-metrics (Soft-DTW, MDTW-NN), underscoring importance of distance metrics for reliable time-series embedding.

Overall, our results demonstrate the importance of a) a valid distance metric, b) Importance Sampling-based method to tackle high-dimensional distributions, and c) deep metric learning to capture complex feature dependence for learning effective time-series embeddings.

## 4 Discussion

**Summary**. Data-driven task-dependent metrics are critical for learning effective time-series embeddings. These learning procedures rely on computing pair-wise similarities to effectively encode the geometrical relationships in the given data. However, unlike distance metrics, similarity measures do not reliably represent the relationships between data points from pair-wise measurements. In this work, we develop a data-driven *metric* for multi-variate time-series data – I-SEA – which estimates the distance between data samples using Expected Alignment. A key contribution here is to enable accurate Expected Alignment computation by developing a way to implicitly estimate the expectation w.r.t. a (high-dimension) distribution over all alignment paths using Non-parametric Density Estimation and Importance Sampling. We establish the theoretical properties of the proposed metric, and demonstrate its superior performance in terms of variance reduction, accuracy, and quality of embeddings on real-world data. A key observation is that I-SEA shows low variance, while conventional Importance Sampling estimators are known to be unstable and result in high variance if the distribution over desired region has a small support (Precup 2000).

**Limitations and Future Work**. Expected Alignment for estimating the distance requires additional sampling and distance computations, which add to the computational overhead in practice at the training stage. Nevertheless, we demonstrate that the learned embeddings lead to better performance, while capturing the geometry of the data and maintaining the order of time and sample complexity.

Although I-SEA can handle irregular sampling in time-series, a detailed exploration along with strategies to address missing data challenge remains an open problem while also considering statistical metrics (Fawaz et al. 2019). Further, estimating high-dimensional distribution $p(\cdot)$ by leveraging recent work on sampling from rare distributions (with significantly more data) also provides exciting avenues for future explorations (O'Kelly et al. 2018).

**Conclusions**. Learning meaningful representations from time-series data is challenging due to the difficulty in constructing a valid distance metric, and the high-dimensionality of the underlying distribution. In this work, we present a way to tackle the high-dimensionality via Importance Sampling for effectively leveraging the inherent structure for various machine learning tasks. Our flexible framework can serve as a general recipe for future explorations, and for learning distribution-aware embeddings for multi-variate time-series.

## Acknowledgments

This work is supported in part by NSF Research Grant CCF-1837131.

## References

Baranowski, P.; Krzyszczak, J.; Slawinski, C.; Hoffmann, H.; Kozyra, J.; Nieróbca, A.; Siwek, K.; and Gluza, A. 2015. Multifractal analysis of meteorological time series to assess climate impacts. *Climate Research*, 65: 39–52.

Che, Z.; He, X.; Xu, K.; and Liu, Y. 2017. DECADE: a deep metric learning model for multivariate time series. In *KDD workshop on mining and learning from time series*.

Chorowski, J.; Bahdanau, D.; Serdyuk, D.; Cho, K.; and Bengio, Y. 2015. Attention-based models for speech recognition. *Advances in Neural Information Processing Systems*, 2015: 577–585.

Cover, T.; and Hart, P. 1967. Nearest neighbor pattern classification. *IEEE transactions on information theory*, 13(1): 21–27.

Cox, M. A.; and Cox, T. F. 2008. Multidimensional scaling. In *Handbook of data visualization*, 315–347. Springer.

Cuturi, M. 2011. Fast global alignment kernels. In *Proceedings of the International conference on machine learning (ICML)*, 929–936.

Cuturi, M.; and Blondel, M. 2017. Soft-dtw: a differentiable loss function for time-series. In *International Conference on Machine Learning*, 894–903. PMLR.

Cuturi, M.; Vert, J. P.; Birkenes, O.; and Matsui, T. 2007. A kernel for time series based on global alignments. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 2, II–413. IEEE.

Fawaz, H. I.; Forestier, G.; Weber, J.; Idoumghar, L.; and Muller, P. A. 2019. Deep learning for time series classification: a review. *Data mining and knowledge discovery*, 33(4): 917–963.

Glynn, P. W.; and Iglehart, D. L. 1989. Importance sampling for stochastic simulations. *Management science*, 35(11): 1367–1392.

Goldberger, A. L.; Amaral, L. A. N.; Glass, L.; Hausdorff, J. M.; Ivanov, P. C.; Mark, R. G.; Mietus, J. E.; Moody, G. B.; Peng, C. K.; and Stanley, H. E. 2000. PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. *circulation*, 101(23): e215–e220.

Hayashi, A.; Mizuhara, Y.; and Suematsu, N. 2005. Embedding time series data for classification. In *International workshop on machine learning and data mining in pattern recognition*, 356–365. Springer.

Hoffer, E.; and Ailon, N. 2015. Deep metric learning using triplet network. In *International workshop on similarity-based pattern recognition*, 84–92. Springer.

Hogeweg, P.; and Hesper, B. 1984. The alignment of sets of sequences and the construction of phyletic trees: an integrated method. *Journal of molecular evolution*, 20(2): 175–186.

Huang, G.; Quo, C.; Kusner, M. J.; Sun, Y.; Weinberger, K. Q.; and Sha, F. 2016. Supervised word mover's distance. In *Advances in Neural Information Processing Systems*, 4869–4877.

Jiang, S.; Jr, J. F.; and Gonzalez, M. C. 2012. Discovering urban spatial-temporal structure from human activity patterns. In *Proceedings of the ACM SIGKDD international workshop on urban computing*, 95–102.

Karamchandani, A.; Bjerager, P.; and Cornell, C. 1989. Adaptive importance sampling. In *Structural Safety and Reliability*, 855–862. ASCE.

Kaya, M.; and Ş. Bilge, H. 2019. Deep metric learning: A survey. *Symmetry*, 11(9): 1066.

Lajugie, R.; Garreau, D.; Bach, F. R.; and Arlot, S. 2014. Metric Learning for Temporal Sequence Alignment. In *Advances in Neural Information Processing Systems*.

Lhermitte, S.; Verbesselt, J.; Verstraeten, W. W.; and Coppin, P. 2011. A comparison of time series similarity measures for classification and change detection of ecosystem dynamics. *Remote sensing of environment*, 115(12): 3129–3152.

Ma, Q.; Zheng, J.; Li, S.; and Cottrell, G. W. 2019. Learning representations for time series clustering. *Advances in neural information processing systems*, 32: 3781–3791.

McFee, B.; Barrington, L.; and Lanckriet, G. 2012. Learning content similarity for music recommendation. *IEEE transactions on audio, speech, and language processing*, 20(8): 2207–2218.

Mei, J.; Liu, M.; Wang, Y. F.; and Gao, H. 2015. Learning a mahalanobis distance-based dynamic time warping measure for multivariate time series classification. *IEEE transactions on Cybernetics*, 46(6): 1363–1374.

Mueller, J.; and Thyagarajan, A. 2016. Siamese recurrent architectures for learning sentence similarity. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30.

Müller, M. 2007. Dynamic time warping. *Information retrieval for music and motion*, 69–84.

Murray, D. B. 1993. Forecasting a chaotic time series using an improved metric for embedding space. *Physica D: Nonlinear Phenomena*, 68(3-4): 318–325.

Myers, C.; Rabiner, L.; and Rosenberg, A. 1980. Performance tradeoffs in dynamic time warping algorithms for isolated word recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 28(6): 623–635.

O'Kelly, M.; Sinha, A.; Namkoong, H.; Tedrake, R.; and Duchi, J. C. 2018. Scalable End-to-End Autonomous Vehicle Testing via Rare-event Simulation. *Advances in Neural Information Processing Systems*, 31: 9827–9838.

Oord, A. V. D.; Dieleman, S.; and Schrauwen, B. 2013. Deep content-based music recommendation. In *Advances in Neural Information Processing Systems*, volume 26.

Precup, D. 2000. Eligibility traces for off-policy policy evaluation. *Computer Science Department Faculty Publication Series*, 80.

Qiu, J.; Wang, X.; Fua, P.; and Tao, D. 2019. Matching seqlets: An unsupervised approach for locality preserving

sequence matching. *IEEE transactions on pattern analysis and machine intelligence*.

Rubinstein, R. Y.; and Kroese, D. P. 2016. *Simulation and the Monte Carlo method*, volume 10. John Wiley & Sons.

Saigo, H.; Vert, J. P.; and Akutsu, T. 2006. Optimizing amino acid substitution matrices with a local alignment kernel. *BMC bioinformatics*, 7(1): 1–12.

Sakoe, H.; and Chiba, S. 1978. Dynamic programming algorithm optimization for spoken word recognition. *IEEE transactions on acoustics, speech, and signal processing*, 26(1): 43–49.

Salakhutdinov, R.; and Hinton, G. 2007. Learning a nonlinear embedding by preserving class neighbourhood structure. In *Artificial Intelligence and Statistics*, 412–419. PMLR.

Schütze, H.; Manning, C. D.; and Raghavan, P. 2008. *Introduction to information retrieval*, volume 39. Cambridge University Press Cambridge.

Shalev-Shwartz, S.; Singer, Y.; and Ng, A. Y. 2004. Online and batch learning of pseudo-metrics. In *Proceedings of the International Conference on Machine learning*, 94.

Shanmugam, D. 2018. *A tale of two time series methods: representation learning for improved distance and risk metrics*. Ph.D. thesis, Massachusetts Institute of Technology.

Shapiro, A. 2003. Monte Carlo sampling methods. *Handbooks in operations research and management science*, 10: 353–425.

Silverman, B. W. 2018. *Density estimation for statistics and data analysis*. Routledge.

Su, B.; and Hua, G. 2018. Order-preserving optimal transport for distances between sequences. *IEEE transactions on pattern analysis and machine intelligence*, 41(12): 2961–2974.

Su, B.; and Wu, Y. 2019. Learning distance for sequences by learning a ground metric. In *International Conference on Machine Learning*, 6015–6025. PMLR.

Tran, D.; and Sorokin, A. 2008. Human activity recognition with metric learning. In *European conference on computer vision*, 548–561. Springer.

Weinberger, K. Q.; and Saul, L. K. 2009. Distance metric learning for large margin nearest neighbor classification. *Journal of machine learning research*, 10(2).

Xing, E. P.; Ng, A. Y.; Jordan, M. I.; and Russell, S. 2002. Distance metric learning with application to clustering with side-information. In *Advances in Neural Information Processing Systems*, volume 15, 505–512.

Xiong, H.; and Chen, X. W. 2006. Kernel-based distance metric learning for microarray data classification. *BMC bioinformatics*, 7(1): 1–11.

Yang, K.; and Shahabi, C. 2004. A PCA-based similarity measure for multivariate time series. In *Proceedings of the ACM international workshop on Multimedia databases*, 65–74.

Yassine, A.; Singh, S.; and Alamri, A. 2017. Mining human activity patterns from smart home big data for health care applications. *IEEE Access*, 5: 13131–13141.

Zhang, J.; Marszałek, M.; Lazebnik, S.; and Schmid, C. 2007. Local features and kernels for classification of texture and object categories: A comprehensive study. *International journal of computer vision*, 73(2): 213–238.

Zhang, P. 1996. Nonparametric importance sampling. *Journal of the American Statistical Association*, 91(435): 1245–1253.