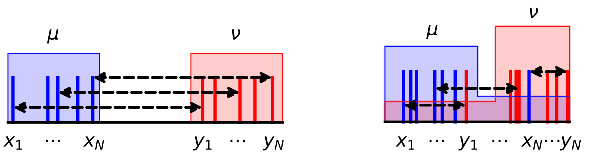
# On the Information Processing of One-Dimensional Wasserstein Distances with Finite Samples

**Cheongjae Jang**[1], **Jonghyun Won**[1], **Soyeon Jun**[2,3],
**Chun Kee Chung**[3], **Keehyoung Joo**[4], **Yung-Kyun Noh**[1,4]
[1]Hanyang University, Korea
[2]Icahn School of Medicine at Mount Sinai, USA
[3]Seoul National University, Korea
[4]Korea Institute for Advanced Study, Korea

## Abstract

Leveraging the Wasserstein distance—a summation of sample-wise transport distances in data space—is advantageous in many applications for measuring support differences between two underlying density functions. However, when supports significantly overlap while densities exhibit substantial pointwise differences, it remains unclear whether and how this transport information can accurately identify these differences, particularly their analytic characterization in finite-sample settings. We address this issue by conducting an analysis of the information processing capabilities of the one-dimensional Wasserstein distance with finite samples. By utilizing the Poisson process and isolating the rate factor, we demonstrate the capability of capturing the pointwise density difference with Wasserstein distances and how this information harmonizes with support differences. The analyzed properties are confirmed using neural spike train decoding and amino acid contact frequency data. The results reveal that the one-dimensional Wasserstein distance highlights meaningful density differences related to both rate and support.

**Extended version** — https://arxiv.org/abs/2511.12881



(a) Different support case (b) Overlapping support case

Figure 1: Sample transport between empirical distributions derived from the underlying one-dimensional distributions $\mu$ and $\nu$. Blue and red spikes represent samples drawn from $\mu$ and $\nu$, respectively, with sample transport distances illustrated by dotted arrows. In (a), the prominent difference is in the support of $\mu$ and $\nu$, while in (b), the two densities are significantly different despite having the same support. In (b), it is desirable for the measure derived from the sample transport between empirical distributions to represent the pointwise density difference in the underlying distributions.

## 1 Introduction

The Wasserstein distance, which leverages transport distances between samples, has become a popular tool across various domains for capturing differences between probability density functions (Rubner, Tomasi, and Guibas 2000; Haker et al. 2004; Kusner et al. 2015; Schiebinger et al. 2019; Sihn and Kim 2019). In particular, this metric is known to address support differences due to its reliance on sample transport distances. However, it remains unclear whether and how this transport information can appropriately identify pointwise density differences, particularly in terms of their analytic characterization in the finite-sample setting. This ambiguity persists even in one-dimensional spaces, where the presence of an analytic form for optimal transport simplifies the analysis of Wasserstein distance properties (Villani et al. 2009; Santambrogio 2015; Peyré, Cuturi et al. 2019; Bobkov and Ledoux 2019) (see Figure 1).

In this paper, we explore the information processing capabilities of the one-dimensional Wasserstein distance between empirical measures derived from sequences of randomly generated finite points, where rates are proportional to the underlying density. By employing Poisson processes that depend exclusively on the rate parameter, we derive that the expected transport distance between samples adequately identifies rate differences which can be equivalently understood as measuring the pointwise density difference. Furthermore, introducing a shift to the support of one empirical distribution, disregarding any other influencing factors, shows that the one-dimensional Wasserstein distance seamlessly accommodates both positional and rate differences.

Traditionally, the mechanism of encoding rate information and position information have been considered mutually exclusive, particularly in brain research. This dichotomy resembles the long-lasting debate on information processing in the brain: temporal coding versus rate coding. Given the necessity of using density difference information, KL-divergence can appropriately quantify rate differences while is overly sensitive to support differences. Conversely, Wasserstein distance has been suggested as a pragmatically useful alternative that avoids such sensitivities (Arjovsky, Chintala, and Bottou 2017; Ozair et al. 2019). However, questions remain about whether Wasserstein distances can reliably treat rate differences, similar to how KL-divergence does, and most importantly, if the distance is minimized

---

when samples come from equivalent densities.

Our examination of the one-dimensional Wasserstein distance reaffirms its practical value for applications requiring the quantification or optimization of differences in rate and support. Our discussion extends to the sliced Wasserstein distance—a measure derived as the expectation of the one-dimensional Wasserstein distance applied to projections of multidimensional distributions along arbitrary directions (Bonneel et al. 2015). These insights could thus be particularly valuable given the recent surge in applications of sliced Wasserstein distances (Kolouri et al. 2018; Deshpande, Zhang, and Schwing 2018; Lee et al. 2019).

In the experiments, we substantiate our theoretical findings with case studies involving neural spike train decoding and amino acid contact frequency data, showing that sample transport distances and the resulting one-dimensional Wasserstein distance can effectively capture density differences related to both rate and support, leading to enhanced classification performance and data representation. These findings highlight the potential utility of the Wasserstein distance-based analysis across diverse fields, particularly in neuroscience and molecular biology.

The paper is organized as follows. Section 2 provides background on the one-dimensional Wasserstein distance. Section 3 presents the derivations of the expected Wasserstein distance using two Poisson processes. Section 4 reports the experimental results that verify our findings and show their behavior beyond the Poisson setting. Throughout the paper, references to the appendix refer to the appendix of the extended version (Jang et al. 2025).

## 2 Background

The $p$-Wasserstein distance between two probability density functions $\mu$ and $\nu$ over $\mathbb{R}^d$ is defined as

$$W_p(\mu,\nu) = \left( \inf_{\gamma\in\Gamma} \int_{\mathbb{R}^d}\int_{\mathbb{R}^d} \|x-y\|^p d\gamma(x,y) \right)^{1/p}, \quad (1)$$

where $\Gamma$ denotes the set of all joint distributions on $\mathbb{R}^d \times \mathbb{R}^d$ that have respective marginals $\mu$ and $\nu$.

In the one-dimensional case ($d = 1$), this admits a closed-form expression using cumulative distribution functions (CDFs) $P$ and $Q$ of $\mu$ and $\nu$, respectively (Dall'Aglio 1956; Peyré, Cuturi et al. 2019). Specializing to $p = 1$ gives

$$W_1(\mu,\nu) = \int_{\mathbb{R}} |P(u) - Q(u)| du. \quad (2)$$

This distance is the primary focus of our study and is denoted as the Wasserstein distance or $W(\mu,\nu)$ throughout.

Let $\hat{\mu}_N$ and $\hat{\nu}_N$ denote empirical distributions constructed from i.i.d. samples $x_1,\ldots,x_N \sim \mu$ and $y_1,\ldots,y_N \sim \nu$, respectively: $\hat{\mu}_N = \frac{1}{N}\sum_{i=1}^N \delta_{x_i}$ and $\hat{\nu}_N = \frac{1}{N}\sum_{i=1}^N \delta_{y_i}$, where $\delta_x$ is a Dirac measure at $x \in \mathbb{R}$. Assuming ordered samples ($x_i < x_{i+1}$ and $y_i < y_{i+1}$ for $i = 1,\ldots,N-1$), the Wasserstein distance between $\hat{\mu}_N$ and $\hat{\nu}_N$ is the average transport distance between ordered samples (see Appendix A.1 for details):

$$W(\hat{\mu}_N,\hat{\nu}_N) = \frac{1}{N}\sum_{k=1}^N |x_k - y_k|.$$

Under mild conditions, $\mathbb{E}[W(\hat{\mu}_N,\mu)] \to 0$ as $N \to \infty$ (Boissard and Le Gouic 2014; Fournier and Guillin 2015; Weed and Bach 2019), and hence, $\mathbb{E}\left[W(\hat{\mu}_N,\hat{\nu}_N)\right] \to W(\mu,\nu)$ (Sommerfeld et al. 2019).

In scenarios involving infinite samples, there exist cases in which the Wasserstein distance can adeptly reveal differences in pointwise densities (or rates). For instance, consider two uniform distributions $\mu = \mathcal{U}[0,1/\lambda_1]$ and $\nu = \mathcal{U}[0,1/\lambda_2]$ with constant probability densities $\lambda_1$ and $\lambda_2$ for $\lambda_1 \leq \lambda_2$. As $N \to \infty$, the Wasserstein distance between the corresponding empirical distributions $\hat{\mu}_N$ and $\hat{\nu}_N$ can be derived analytically according to (2) as

$$\lim_{N\to\infty} \mathbb{E}\left[W(\hat{\mu}_N,\hat{\nu}_N)\right] = W(\mu,\nu) = \frac{1}{2}\left(\frac{1}{\lambda_1} - \frac{1}{\lambda_2}\right).$$

However, this analytic solution cannot be straightforwardly extended to the finite sample case. The finite sample expectations are instead equal to the sum of the expected sample distances, given as:

$$\mathbb{E}[W(\hat{\mu}_N,\hat{\nu}_N)] = \frac{1}{N}\sum_{k=1}^N \mathbb{E}[|x_k - y_k|]. \quad (3)$$

Although the asymptotic behavior implied by (2) offers some insight, it remains implicit. The finite-sample expression in (3), *per se*, does not explicitly clarify whether and how the expected sample distances $\mathbb{E}[|x_k - y_k|]$ or the empirical Wasserstein distance effectively capture rate differences in finite-sample cases. This paper aims to characterize these aspects both analytically and empirically.

## 3 Rate and Support Difference Encoding in Wasserstein Distance with Finite Samples

To directly examine how sample-level transport distances encode pointwise density and support differences in finite-sample settings, we utilize Poisson processes as a tractable analytic framework. These processes are commonly employed to model the occurrence of random events over time, such as neural spike events (Heeger et al. 2000; Kass and Ventura 2001). The rate parameter $\lambda > 0$, the defining characteristic of these processes, directly controls the frequency of events (higher $\lambda$ yields denser spikes). By focusing on this process, we can effectively isolate the impact of rate and support differences while keeping other variables controlled.

We analyze the rate difference encoding in Section 3.1 and support difference encoding in Section 3.2 for constant rate parameters, followed by a discussion of the time-varying rate parameter case in Section 3.3.

### 3.1 Rate Difference Encoding

Consider two sequences of event (or 'spike') times, $\{x_i\}_{i=1}^N$ and $\{y_i\}_{i=1}^N$, generated from Poisson processes with constant rates $\lambda_1$ and $\lambda_2$, where $\lambda_1 < \lambda_2$. For simplicity and clarity, we begin our analysis with constant rates and identical sample sizes ($N$), though the derivations can be extended to cases with different sample sizes and time-varying rates.

In a Poisson process with a constant rate $\lambda$, the $k$-th event time follows an Erlang distribution with shape $k$ and rate $\lambda$, having density $p(z;k,\lambda) = \frac{1}{(k-1)!}\lambda^k z^{k-1} e^{-\lambda z}$. Thus, for our sequences, $x_k \sim p(x_k;k,\lambda_1)$ and $y_k \sim p(y_k;k,\lambda_2)$.

Our goal is to analyze the expected Wasserstein distance between the empirical distributions of $\{x_i\}_{i=1}^N$ and $\{y_i\}_{i=1}^N$ as defined in (3). To do this, we first determine the expected distance between the $k$-th spikes $x_k$ and $y_k$:

$$\mathbb{E}[|x_k - y_k|] = \int_0^\infty\int_0^\infty |x_k - y_k| p(x_k;k,\lambda_1)p(y_k;k,\lambda_2)dx_k dy_k.$$

The following proposition, derived via a non-trivial computation of this expectation, elucidates how the expected distance between event time pairs $(x_k, y_k)$ reflects the difference in their rates, $\lambda_1$ and $\lambda_2$:

**Proposition 3.1.** *For the $k$-th spikes $x_k$ and $y_k$ obtained from two Poisson processes of rates $\lambda_1$ and $\lambda_2$, respectively, the expection of the distance between $x_k$ and $y_k$ is*

$$\mathbb{E}[|x_k - y_k|] = \frac{\lambda_1 + \lambda_2}{2\lambda_1\lambda_2}\mathbb{E}_{i\sim P(i|2k,p)}\left[|i - (2k-i)|\right], \quad (4)$$

*where $p = \lambda_1/(\lambda_1 + \lambda_2)$ and $P(i|2k,p) = \binom{2k}{i}p^i(1-p)^{2k-i}$ is the binomial distribution with the parameters $2k$ and $p$. The minimum of $\mathbb{E}[|x_k - y_k|]$ is achieved when $\lambda_1 = \lambda_2$, under the constraint of constant harmonic mean between rates, i.e., $\frac{2\lambda_1\lambda_2}{\lambda_1+\lambda_2} = C$ for a $C > 0$.*

***Proof.*** *The proof is provided in Appendix A.2.* ☐

Note that (4) is expressed solely in terms of the rates $\lambda_1$ and $\lambda_2$, exhibiting explicit symmetry between them. Further analysis of the expected distance in Proposition 3.1 reveals that the expected Wasserstein distance in (3) is minimized with $\lambda_1 = \lambda_2$, given that $\frac{\lambda_1+\lambda_2}{2\lambda_1\lambda_2}$ is fixed so that $\frac{1}{2}(\mathbb{E}[x_k] + \mathbb{E}[y_k])$ is constant for each $k$ (see Figure 2). This finding underscores that Wasserstein distances computed from finite samples reliably and effectively capture rate differences. A numerical validation is provided in Appendix B.

**An infinite-sample case discussion.** The above discussion becomes increasingly apparent with larger $N$. As $k \to \infty$, the expected distance between the $k$-th spikes, $x_k$ and $y_k$, normalized by $k$, is given by the following proposition:

**Proposition 3.2.** *Let $x_k$ and $y_k$ be the $k$-th spikes from two Poisson processes with rates $\lambda_1$ and $\lambda_2$, respectively, assuming without loss of generality that $\lambda_1 < \lambda_2$, and define $s_k = |x_k - y_k|/k$. Then,*

$$\lim_{k\to\infty} \mathbb{E}\left[s_k\right] = \frac{1}{\lambda_1} - \frac{1}{\lambda_2} \quad \textit{and} \quad \lim_{k\to\infty} \mathrm{Var}\left[s_k\right] = 0. \quad (5)$$

***Proof.*** *The proof is provided in Appendix A.3.* ☐

Based on Proposition 3.2, when $\lambda_1 < \lambda_2$, $\mathbb{E}[W(\hat{\mu}_N, \hat{\nu}_N)]$ for large $N$ can be approximated, up to the leading order, as

$$\mathbb{E}[W(\hat{\mu}_N, \hat{\nu}_N)] \approx \frac{1}{N}\sum_{k=1}^{N}\left(\frac{k}{\lambda_1} - \frac{k}{\lambda_2}\right) = \frac{N+1}{2}\left(\frac{1}{\lambda_1} - \frac{1}{\lambda_2}\right). \quad (6)$$

The dependence of (6) on the inverse rate difference demonstrates how the Wasserstein distance directly captures rate differences.

Additionally, the expression in (6) can be interpreted through its connection to uniform distributions. In a Poisson process with $N$ events occurring by a fixed time $t$, the joint distribution of the event times, $p(x_1, \ldots, x_N)$, is uniform over $0 < x_1 < \ldots < x_N < t$ (Gallager 2013). As $N \to \infty$, the leading-order expression in (6) corresponds to the Wasserstein distance between two uniform distributions $\mu = \mathcal{U}\left[0, (N+1)/\lambda_1\right]$ and $\nu = \mathcal{U}\left[0, (N+1)/\lambda_2\right]$.

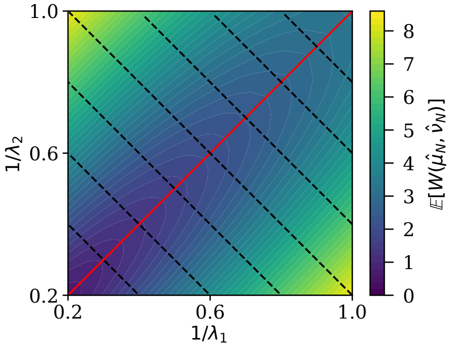


Figure 2: $\mathbb{E}[W(\hat{\mu}_N, \hat{\nu}_N)]$ for $\lambda_1, \lambda_2 \in [1, 5]$ and $N = 20$. When the harmonic mean of the rates is constant (black dashed lines), $\mathbb{E}[W(\hat{\mu}_N, \hat{\nu}_N)]$ shows its minimum where the rates are equal ($\lambda_1 = \lambda_2$, red solid line).

### 3.2 Support Difference Encoding

We now explore how differences in support influence the expected Wasserstein distance. This allows us to understand how rate and support difference information are harmonized, which is crucial for a comprehensive understanding of the properties of the Wasserstein distance. This also aligns with recent findings emphasizing the importance of time encoding over rate encoding in neural information processing, where absolute distances between spikes play a critical role (VanRullen, Guyonneau, and Thorpe 2005; Gütig 2014).

To analyze the effects of support differences, we consider the shift of support by $\Delta t \geq 0$, to the spike timings in one of the Poisson processes. We calculate the expected distance between $x_k + \Delta t$ and $y_k$ using a similar method described in Section 3.1. A comprehensive derivation is in Appendix A.4.

For clarity and foundational insight, here we focus on the expected distance between the first spikes of each sequence:

$$\mathbb{E}[|x_1 + \Delta t - y_1|] = e^{-\lambda_2 \Delta t}\left(\frac{\lambda_1 + \lambda_2}{\lambda_1\lambda_2}\mathbb{E}_{i\sim P(i|2,p)}[|i-1|]\right) + \Delta t + (1 - e^{-\lambda_2\Delta t})\left(\frac{1}{\lambda_1} - \frac{1}{\lambda_2}\right). \quad (7)$$

Compared to the rate difference encoding in (4) for $k = 1$, (7) introduces additional components: $\mathbb{E}[|x_1 - y_1|]$ weighted by $e^{-\lambda_2 \Delta t}$, the support shift $\Delta t$, and a term involving the inverse rate difference. When $\Delta t = 0$, the formula recovers the pure rate-based expectation in (4) with $k = 1$. As $\Delta t \to \infty$, the expression simplifies to $\Delta t + \frac{1}{\lambda_1} - \frac{1}{\lambda_2}$, where the shift dominates. A numerical validation of the above result is provided in Appendix B.

This analysis reveals how the Wasserstein distance can encode both rate and support differences and how these factors jointly influence the distance. The balance between these two types of information—rate difference and support difference—is moderated by the magnitude of the shift $\Delta t$.

### 3.3 Discussion on Time-Varying Rate Cases

Analyzing the case of time-varying rates provides valuable insight into how sample distances can reflect rate and support differences for arbitrary probability distributions. For the case of nonhomogeneous Poisson processes with a time-varying rate parameter $\mu(t)$, the distribution of the $k$-th spike arrival time is obtained as follows (Gallager 2013):

$$p(x_k; \mu(\cdot)) = \frac{\mu(x_k)}{(k-1)!} \exp(-m(x_k)) m(x_k)^{k-1}, \qquad (8)$$

where $m(x) = \int_0^x \mu(t)dt$.

To compute the expectation of $|x_k - y_l|$, where $y_l$ is the $l$-th spike arrival time for nonhomogeneous Poisson processes with a time-varying rate parameter $\nu(t)$, we employ a substitution integral: $x_k \mapsto u = m(x_k)$ and $y_l \mapsto v = n(y_l) = \int_0^{y_l} \nu(t)dt$. This leads to the following expression:

$$\begin{aligned} &\int_0^\infty \int_0^\infty |x_k - y_l| p(x_k; \mu(\cdot)) p(y_l; \nu(\cdot)) dx_k dy_l \\ &= \int_0^\infty \int_0^\infty |m^{-1}(u) - n^{-1}(v)| \frac{1}{(k-1)!} \exp(-u) u^{k-1} \qquad (9) \\ &\cdot \frac{1}{(l-1)!} \exp(-v) v^{l-1} du dv. \end{aligned}$$

This formulation allows our derivations to extend to more general cases where the double Laplace transform of $|m^{-1}(u) - n^{-1}(v)| u^{k-1} v^{l-1}$ is analytically tractable. In such scenarios, we expect differences in rates and supports to be explicitly captured in the sample transport distances and the resulting Wasserstein distances. We use this idea to construct an illustrative example that integrates sliced Wasserstein distances into our framework (see Appendix C).

## 4 Experiments

To validate our analysis of finite-sample Wasserstein distances, confirm that these analytic insights remain effective beyond the Poisson setting, and assess their practical usefulness in downstream tasks like classification and representation learning, we conduct experiments on synthetic (Section 4.1) and real-world data, including neural spike trains (Sections 4.2 and 4.3) and amino acid contacts (Section 4.4).

### 4.1 A Synthetic Data Experiment

We present a simple example illustrating how sample distances and the Wasserstein distance capture rate and support differences, effectively integrating both. For comparison, we also consider alternative measures that capture support or rate differences in distinct ways, such as the Hausdorff distance and the Jensen-Shannon (JS) divergence.

To create a setting with controllable rate and support differences, we use empirical measures from samples generated by two Poisson processes with varying rates $\mu(t)$ and $\nu(t)$, as shown in Figure 3(a) and detailed in Appendix D.

As depicted in Figure 3(b), the Hausdorff distance fails to capture rate differences, while the JS-divergence is overly sensitive with respect to $\Delta t$ when $|\Delta t| < 1$ and then saturates when $|\Delta t| \geq 1$. In contrast, the Wasserstein distance and the sample distance, such as $|x_{50} - y_{50}|$, effectively capture and harmonize both rate and support differences.

| Feature | $\log(r_1)$ | $\log(r_2)$ | $\lvert\Delta t\rvert$ |
|---|---|---|---|
| Directed Hausdorff | 43.7± 0.5 | 43.9± 0.3 | 70.4± 0.3 |
| Bin-wise JS divergence | 64.0± 0.4 | 68.4± 0.3 | 70.3± 0.1 |
| Sample transport cost | **81.5** ± 0.1 | **81.9**± 0.2 | **98.9**± 0.0 |

Table 1: $R^2$ scores for estimating rate and support differences in the synthetic example using various features.

| Method | Retina-All | Retina14 | Retina23 |
|---|---|---|---|
| FCN | 0.945 ± 7e-04 | 0.962 ± 7e-04 | 0.925 ± 3e-04 |
| FCN + SD1 | **0.951** ± 4e-04 | **0.971** ± 9e-04 | **0.931** ± 6e-04 |
| FCN + SD2 | 0.945 ± 1e-03 | **0.968** ± 1e-03 | **0.935** ± 4e-04 |
| IT | 0.937 ± 8e-04 | 0.955 ± 2e-03 | 0.889 ± 2e-03 |
| IT + SD1 | **0.951** ± 6e-04 | 0.960 ± 2e-03 | **0.912** ± 3e-03 |
| IT + SD2 | **0.950** ± 1e-03 | **0.966** ± 2e-03 | **0.913** ± 1e-03 |
| ResNet | 0.937 ± 7e-04 | 0.967 ± 2e-03 | 0.898 ± 8e-04 |
| ResNet + SD1 | **0.943** ± 1e-03 | 0.965 ± 8e-04 | **0.912** ± 1e-03 |
| ResNet + SD2 | **0.948** ± 5e-04 | 0.967 ± 2e-03 | **0.913** ± 2e-03 |
| XT | 0.944 ± 6e-04 | 0.970 ± 1e-03 | 0.930 ± 8e-04 |
| XT + SD1 | 0.947 ± 2e-03 | **0.979** ± 6e-04 | 0.932 ± 1e-03 |
| XT + SD2 | **0.950** ± 6e-04 | **0.978** ± 9e-04 | 0.932 ± 1e-03 |

Table 2: Classification test AUC for retinal stimulus types.

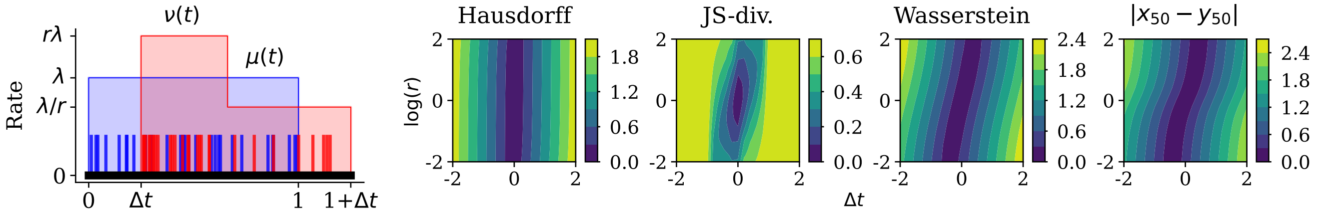
To quantitatively assess how well sample transport distance captures rate and support differences, we conduct an experiment to predict these differences using transport costs as features. For sequences generated from Poisson processes with a constant rate $\mu(t)$ and a piecewise constant rate $\nu(t)$ (defined by two rate ratios, $r_1$ and $r_2$, and a support shift $\Delta t$ relative to $\mu(t)$), we extract ten-dimensional features representing transport costs computed over corresponding decile partitions of the empirical measures. These features are used to estimate $\log(r_1)$, $\log(r_2)$, and $|\Delta t|$ of $\nu(t)$ by training a three-layer fully connected neural network.

For comparison, we also perform the same task using features derived from (i) the bin-wise divergence values for JS-divergence between the probability mass functions of the samples (computed with ten equal-sized bins), and (ii) two directed Hausdorff distances between the samples. Further details are in Appendix D.

Table 1 presents the $R^2$ values for each feature set. The results show that sample transport cost features achieve higher $R^2$ scores than other distance-related features, demonstrating their effectiveness in capturing both rate and support differences. This supports our theoretical analysis that sample transport distance effectively encodes these differences.

### 4.2 Retinal Ganglion Cell Stimulus Classification

Here, we evaluate the usefulness of information from sample transport distances for stimulus-type classification using retinal ganglion cell spike train data. Previous work (Lazarevich et al. 2023) showed that using inter-spike interval (ISI) values alone achieves high classification performance for individual spike train chunks. We show that complementing ISI values with sample transport distances



(a) The time-varying rates and samples

(b) Values between empirical measures with varying $\Delta t$ and $r$

Figure 3: A one-dimensional example to compare information processing of the Hausdorff distance, Jensen-Shannon divergence, and Wasserstein distance. In (a), blue and red spikes represent the samples generated by Poisson processes with time-varying rates $\mu(t)$ and $\nu(t)$, respectively. In (b), the values are averaged over 1,000 trials.

between their empirical measures—which capture distributional differences—can further improve this performance.

**Dataset.** We use spike time data from multi-electrode array recordings of salamander retinal ganglion cells under four stimulus types (Prentice et al. 2016). The classification tasks include a multiclass problem (Retina-All) to classify all types and binary problems (Retina14, Retina23) to distinguish between specific class pairs. Following the preprocessing pipeline in (Lazarevich et al. 2023), we construct the dataset $\mathcal{D} = \{(X_i, y_i)\}_{i=1}^{T}$, where $X_i = (x_{i,1}, \ldots, x_{i,200}) \in \mathbb{R}^{200}$ represents an ISI temporal vector extracted from single-neuron spike train windows and $y_i \in \{1, 2, 3, 4\}$ is the corresponding label. The dataset is split into $T_{train}$ training and $T_{test}$ test sets.

**Methods.** To extract features that capture distributional differences in ISI values, we compute the sample transport cost between the empirical measure of each input sequence $X_i = \{x_{i,k}\}_{k=1}^{200}$ and that of the training set $Y = \{x_{i,k}\}_{i=1}^{T_{train}}{}_{,k=1}^{200}$. By evaluating transport costs over each $1/200$ probability mass interval, we obtain a 200-dimensional feature, called SD1. Alternatively, transport costs can be computed between $X_i$ and the class-wise aggregated empirical distributions, yielding another feature set, SD2.

We train four types of 1D CNN models, such as FCN and ResNet (Wang, Yan, and Oates 2017), InceptionTime (IT; Ismail Fawaz et al. 2020), and XceptionTime (XT; Rahimian et al. 2019), using ISI data and SD1 or SD2. These inputs are passed through separate CNNs, and the outputs are concatenated for final classification. SGD is used to minimize the cross-entropy loss. See Appendix E for details.

**Results.** Table 2 reports the test AUC for each classification task, with p-values below 0.05 in bold. Incorporating SD1 or SD2 features alongside ISI data shows statistically significant improvements in performance over the baseline trained using only ISI data. This demonstrates that effectively leveraging distribution differences—e.g., rate and support differences—via sample transport distances can enhance stimulus type classification performance.

### 4.3 An Analysis of Human Neural Spike Trains

We examine the information processing capabilities of the Wasserstein distance using human neural spike train data. By segmenting long spike trains with a sliding window, we explore temporal shifts and rate differences across segments. Analyzing embeddings that preserve pairwise distances between spike trains, we show that the Wasserstein distance effectively encodes these variations, offering meaningful insights into the dynamics of individual spike trains.

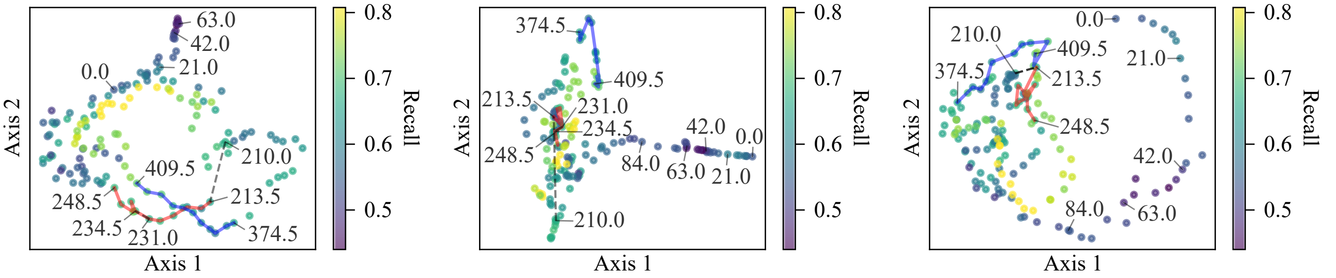
**Dataset.** The neural spike trains used in this study were recorded from the hippocampus of a human subject during the retrieval phase of a word memory task (Jun et al. 2020; Jun, Kim, and Chung 2023). A 667-sec spike time series was captured simultaneously from four hippocampal microelectrodes and segmented into a series of 140-sec windows with a 3.5-sec sliding interval, yielding a dataset $\mathcal{D} = \{X_i\}_{i=1}^{T}$ with $T = 151$. Each $X_i = (X_i^{(1)}, X_i^{(2)}, X_i^{(3)}, X_i^{(4)})$ represents a four-channel spike train, where $X_i^{(k)} = \{x_j^{(i,k)}\}_{j=1}^{l_{i,k}}$ is the spike train recorded from the $k$-th electrode in window $i$ (see Appendix F.1 for details).

**Embedding algorithm.** We employ the Isomap algorithm (Tenenbaum, Silva, and Langford 2000) to obtain two-dimensional embeddings that aim to preserve pairwise Wasserstein distances between spike trains in $\mathcal{D}$. For each four-channel spike train $X_i$, we construct empirical distributions $\hat{\mu}_i^{(k)}$ for each channel $X_i^{(k)}$. The composite Wasserstein distance between spike trains is then calculated as $W^*(X_i, X_j) = \sqrt{\sum_{k=1}^{4} W^2(\hat{\mu}_i^{(k)}, \hat{\mu}_j^{(k)})}$.

For comparison, we also obtain Isomap embeddings using other commonly used dissimilarity measures in neural spike train analysis, including the spike count difference $d(X_i, X_j) = \sqrt{\sum_{k=1}^{4} (|X_i^{(k)}| - |X_j^{(k)}|)^2}$ and the Victor-Purpura (VP) distance (Victor and Purpura 1997). Additional embedding results using a kernel feature space distance from (Park et al. 2013) are provided in Appendix F.2.

**Rate difference and temporal shift encoding in Wasserstein embeddings.** Figure 4 shows the Isomap embeddings generated using the Wasserstein distance, spike count difference, and VP distance. In each embedding, point color indicates the recall of old words during the memory task for the corresponding time window.

When examining embeddings obtained from the spike count difference measure and the VP distance, we found that the primary variation only distinguishes the initial win-



(a) Wasserstein distance (b) Spike count difference (c) Victor-Purpura distance

Figure 4: Three Isomap embeddings of human neural spike trains. We present the embedding obtained using the Wasserstein distance in (a), that from the spike count difference in (b), and that from the Victor-Purpura (VP) distance in (c).

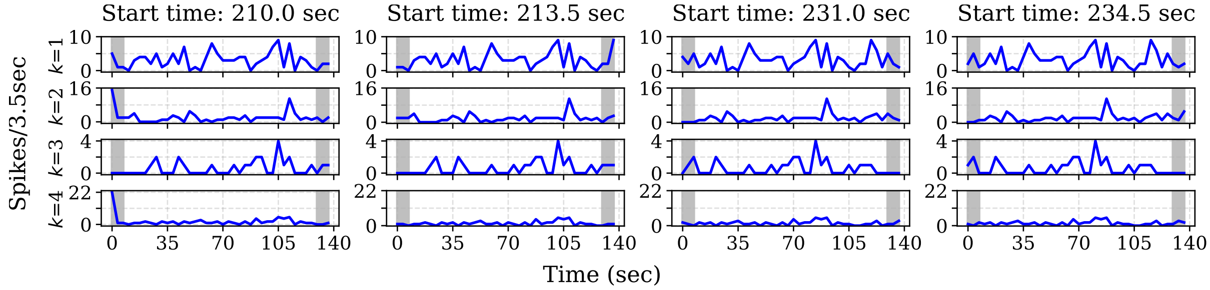


Figure 5: Spike count histograms for windows starting at 210.0 seconds, 213.5 seconds, 231.0 seconds, and 234.5 seconds. Shaded areas in each histogram indicate the beginning and end regions of each window.

dows from the rest, failing to capture meaningful patterns in later windows. In contrast, the Wasserstein distance embedding goes beyond this limitation. It shows a smoother trajectory (note the red lines representing identical time intervals across subfigures in Figure 4) and more consistent variations in recall values.

A closer examination reveals instances of sharp jumps between successive points, the one marked by the black dotted line in Figure 4(a). The two spike count histograms on the left in Figure 5, corresponding to this jump, show significant neural activity at the beginning or end of the window. This creates a large rate difference between adjacent windows. The Wasserstein distance captures this by representing it as a substantial embedding difference, enabling the detection of significant modes in the signal. Conversely, with moderate activity and no extreme spikes at either end (as shown in the two histograms on the right of Figure 5), temporal shift becomes the primary information, resulting in a continuous trajectory similar to the red and blue lines in Figure 4(a). Notably, the VP distance does not effectively capture these distinctions, as Figure 4(c) shows.

The information processing capabilities of the Wasserstein distance hold strong potential for neuroscience applications. By leveraging suitable embedding algorithms, it enables similar neural activity patterns, even those occurring far apart in time, to be represented closely in the embedding space. An example is the alignment of trajectories such as the red and blue lines in Figure 4(a), where the subject's recall performance is comparable at the intersection point. While these embeddings do not, by themselves, offer definitive conclusions, the analysis suggests that the Wasserstein distance effectively captures both similarities and dissimilarities in neural activity, offering valuable insights for neural signal interpretation. Importantly, these insights are qualitatively distinct from—and complementary to—those derived from conventional measures such as spike count difference and the Victor–Purpura distance.

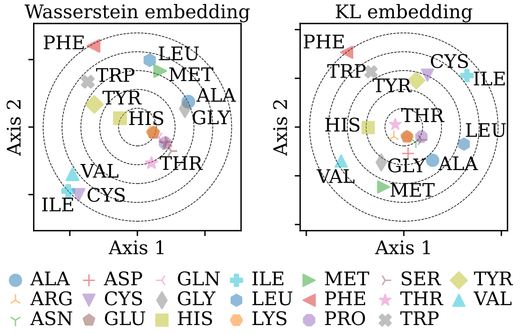
## 4.4 An Analysis of Amino Acid Contacts

We analyze amino acid contact data to demonstrate the effectiveness of the Wasserstein distance in capturing rate differences. In natural proteins, there are twenty types of amino acids, and the rate at which one amino acid appears near another depends on their spatial distance. This rate, known as the contact frequency, is distinct for each amino acid pair.

In this experiment, we use contact frequency data to generate embeddings for the twenty amino acids. By capturing rate differences via the Wasserstein distance, these embeddings reveal specific properties of amino acids—offering insights that, in some cases, go beyond what is captured by KL divergence in a molecular biology context.

Throughout we use the terms "amino acid" and "(protein) residue" interchangeably and denote each amino acid by its standard three-letter code.

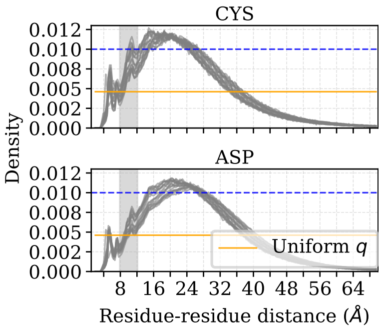
**Dataset.** Amino acid contact frequencies are represented as a set of pairwise distributions $\{p_{ij} \mid 1 \leq i, j \leq 20, i \neq j\}$, where each $p_{ij}$ denotes the normalized histogram of spatial distances between the $i$-th and $j$-th amino acids (ordered alphabetically). Each histogram is defined over 217



(a) Two-dimensional minimum distortion embeddings

| Amino Acid |
|---|
| CYS |
| PHE, ILE |
| VAL |
| LEU, MET, TRP |
| HIS |
| TYR |
| ALA |
| GLY |
| THR |
| ... |

(b) A hydrophobicity table



(c) $p_{ij}$ of the amino acids CYS and ASP

Figure 6: Results from the minimum distortion embeddings of amino acids using pairwise dissimilarities $d_{ij}$. In (a), the left shows the two-dimensional embedding obtained from Wasserstein distances, and the right shows that from KL-divergences. In (b), we present a hydrophobicity rankings of the top twelve hydrophobic amino acids as reported by (Rose et al. 1985). In (c), shaded areas indicate regions of significant rate differences in $p_{ij}$, with gray curves showing $p_{ij}$ for all $j \neq i$.

uniformly spaced bins with identical support across all pairs, see Figure 6(c) for examples. These distributions are derived from 12,508 proteins sourced from the Protein Data Bank (PDB) (Berman et al. 2000).

**Embedding algorithm.** We construct amino acid embeddings based on their pairwise distance distributions $p_{ij}$. To define pairwise residue dissimilarity, we use $d_{ij} = \gamma_W \cdot (W(p_{ij}, q) - \beta_W)$, where $q$ is a uniform reference distribution with the same binning as $p_{ij}$, and $\gamma_W, \beta_W > 0$ are scaling factors. This dissimilarity reflects the rate difference between $p_{ij}$ and $q$, and it can effectively capture the relative profile of each pairwise distribution $p_{ij}$ compared to others.

We apply $d_{ij}$ to the Riemannian geometric manifold learning algorithm of (Jang, Noh, and Park 2021) to obtain embeddings to preserve geometric structures (e.g., distances and angles) induced from the dissimilarity. For comparison, we also obtain embeddings using KL-based dissimilarities defined as $d_{ij} = \gamma_{KL} \cdot (D_{KL}(p_{ij}||q) - \beta_{KL})$, with scaling factors $\gamma_{KL}, \beta_{KL} > 0$. See Appendix G for more details.

**Rate differences induced by long-range contacts.** Figure 6(a) shows two-dimensional embeddings of amino acids based on the Wasserstein distance and KL-divergence. In both embeddings, hydrophobic residues such as CYS, ILE, and PHE (as indicated in the hydrophobicity table in Figure 6(b)), appear near the periphery, while less hydrophobic residues are positioned closer to the center. This separation reflects the rate differences between $p_{ij}$ and the reference distribution $q$, suggesting that the contact frequencies of hydrophobic residues deviate more substantially from $q$.

Further examination of the distributions $p_{ij}$ for CYS in Figure 6(c) (and similarly for ILE and PHE in Figure 12 of Appendix G) reveals that these larger rate differences, particularly compared to residues in the second row of the figures, are concentrated in the range of 8-12Å. From a protein structural perspective, high density in this range indicates a propensity for long-range contacts, a characteristic commonly observed for medium-to-high hydrophobic residues such as CYS, ILE, and PHE (Rose et al. 1985; Gromiha and Selvaraj 1999, 2004). Their peripheral positioning in both embeddings (Figure 6(a)) effectively highlights these specific properties of amino acids, demonstrating that both Wasserstein and KL-based embeddings capture biologically meaningful features such as long-range contacts.

While both embeddings capture this trend, they differ in how well they align with known hydrophobicity rankings. For example, CYS, the most hydrophobic amino acid by (Rose et al. 1985), appears at the outermost edge in the Wasserstein embedding, but closer to the center in the KL-based one. Overall, the norms of the Wasserstein embeddings show a stronger correlation with the hydrophobicity rankings reported in (Rose et al. 1985): Kendall's tau coefficients for radial ordering among the top 10 most hydrophobic residues are 0.722 for Wasserstein vs. 0.582 for KL, and for all residues, 0.807 vs. 0.731, respectively. These results highlight the Wasserstein distance's strength in capturing rate-based structural patterns in amino acid contact data.

## 5 Conclusion

In this paper, we have investigated the information processing capabilities of the one-dimensional Wasserstein distance with finite samples. While its ability to capture support differences in density functions is well understood, its mechanism for encoding pointwise density differences through sample transport distances—particularly in finite sample settings—has remained unclear and lacked analytic characterization. We address this gap by demonstrating rate encoding in one-dimensional distances between empirical measures derived from Poisson processes. Our analysis of expected sample distances provides intuitive interpretations and novel insights into how transport distances within samples encode rate differences and integrate them with support variation. Furthermore, our analytic and empirical findings highlight the complementary nature of the Wasserstein distance relative to other measures of distributional difference, such as KL-divergence. We validated these insights through experiments on neural spike train decoding and amino acid contact analysis, demonstrating the broad applicability of Wasserstein-based approaches across diverse domains.

# Acknowledgments

This work was supported by NRF/MSIT (No. RS-2024-00421203), which supported C. K. Chung and Y.-K. Noh. This work was also supported by IITP/MSIT (No. RS-2023-00220628), which supported K. Joo and Y.-K. Noh. C. Jang was partly supported by NRF/ME (No. RS-2023-00249714). C. K. Chung was partly supported by NRF/MSIT (NRF-2021M3E5D2A01019093). Y.-K. Noh was partly supported by IITP/MSIT (IITP-2021-0-02068, RS-2020-II201373). We thank the KIAS Center for Advanced Computation for providing computing resources.

# References

Arjovsky, M.; Chintala, S.; and Bottou, L. 2017. Wasserstein generative adversarial networks. In *International conference on machine learning*, 214–223. PMLR.

Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; and Bourne, P. E. 2000. The protein data bank. *Nucleic acids research*, 28(1): 235–242.

Bobkov, S.; and Ledoux, M. 2019. *One-dimensional empirical measures, order statistics, and Kantorovich transport distances*, volume 261. American Mathematical Society.

Boissard, E.; and Le Gouic, T. 2014. On the mean speed of convergence of empirical and occupation measures in Wasserstein distance. In *Annales de l'IHP Probabilités et statistiques*, volume 50, 539–563.

Bonneel, N.; Rabin, J.; Peyré, G.; and Pfister, H. 2015. Sliced and radon wasserstein barycenters of measures. *Journal of Mathematical Imaging and Vision*, 51: 22–45.

Dall'Aglio, G. 1956. Sugli estremi dei momenti delle funzioni di ripartizione doppia. *Annali della Scuola Normale Superiore di Pisa-Classe di Scienze*, 10(1-2): 35–74.

Deshpande, I.; Zhang, Z.; and Schwing, A. G. 2018. Generative modeling using the sliced wasserstein distance. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3483–3491.

Fournier, N.; and Guillin, A. 2015. On the rate of convergence in Wasserstein distance of the empirical measure. *Probability theory and related fields*, 162(3): 707–738.

Gallager, R. G. 2013. *Stochastic processes: theory for applications*. Cambridge University Press.

Gromiha, M. M.; and Selvaraj, S. 1999. Importance of long-range interactions in protein folding. *Biophysical chemistry*, 77(1): 49–68.

Gromiha, M. M.; and Selvaraj, S. 2004. Inter-residue interactions in protein folding and stability. *Progress in biophysics and molecular biology*, 86(2): 235–277.

Gütig, R. 2014. To spike, or when to spike? *Current opinion in neurobiology*, 25: 134–139.

Haker, S.; Zhu, L.; Tannenbaum, A.; and Angenent, S. 2004. Optimal mass transport for registration and warping. *International Journal of computer vision*, 60: 225–240.

Heeger, D.; et al. 2000. Poisson model of spike generation. *Handout, University of Standford*, 5(1-13): 76.

Ismail Fawaz, H.; Lucas, B.; Forestier, G.; Pelletier, C.; Schmidt, D. F.; Weber, J.; Webb, G. I.; Idoumghar, L.; Muller, P.-A.; and Petitjean, F. 2020. Inceptiontime: Finding alexnet for time series classification. *Data Mining and Knowledge Discovery*, 34(6): 1936–1962.

Jang, C.; Noh, Y.-K.; and Park, F. C. 2021. A Riemannian geometric framework for manifold learning of non-Euclidean data. *Advances in Data Analysis and Classification*, 15(3): 673–699.

Jang, C.; Won, J.; Jun, S.; Chung, C. K.; Joo, K.; and Noh, Y.-K. 2025. On the Information Processing of One-Dimensional Wasserstein Distances with Finite Samples. *arXiv preprint arXiv:2511.12881*.

Jun, S.; Kim, J. S.; and Chung, C. K. 2023. Hippocampal neuronal activity preceding stimulus predicts later memory success. *Eneuro*, 10(2).

Jun, S.; Lee, S. A.; Kim, J. S.; Jeong, W.; and Chung, C. K. 2020. Task-dependent effects of intracranial hippocampal stimulation on human memory and hippocampal theta power. *Brain stimulation*, 13(3): 603–613.

Kass, R. E.; and Ventura, V. 2001. A spike-train probability model. *Neural computation*, 13(8): 1713–1720.

Kolouri, S.; Pope, P. E.; Martin, C. E.; and Rohde, G. K. 2018. Sliced-wasserstein autoencoder: An embarrassingly simple generative model. *arXiv preprint arXiv:1804.01947*.

Kusner, M.; Sun, Y.; Kolkin, N.; and Weinberger, K. 2015. From word embeddings to document distances. In *International conference on machine learning*, 957–966. PMLR.

Lazarevich, I.; Prokin, I.; Gutkin, B.; and Kazantsev, V. 2023. Spikebench: an open benchmark for spike train time-series classification. *PLOS Computational Biology*, 19(1): e1010792.

Lee, C.-Y.; Batra, T.; Baig, M. H.; and Ulbricht, D. 2019. Sliced wasserstein discrepancy for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10285–10295.

Ozair, S.; Lynch, C.; Bengio, Y.; Van den Oord, A.; Levine, S.; and Sermanet, P. 2019. Wasserstein dependency measure for representation learning. *Advances in Neural Information Processing Systems*, 32.

Park, I. M.; Seth, S.; Paiva, A. R.; Li, L.; and Principe, J. C. 2013. Kernel methods on spike train space for neuroscience: a tutorial. *IEEE Signal Processing Magazine*, 30(4): 149–160.

Peyré, G.; Cuturi, M.; et al. 2019. Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 11(5-6): 355–607.

Prentice, J. S.; Marre, O.; Ioffe, M. L.; Loback, A. R.; Tkačik, G.; and Berry, M. J. 2016. Error-robust modes of the retinal population code. *PLoS computational biology*, 12(11): e1005148.

Rahimian, E.; Zabihi, S.; Atashzar, S. F.; Asif, A.; and Mohammadi, A. 2019. Xceptiontime: A novel deep architecture based on depthwise separable convolutions for hand gesture classification. *arXiv preprint arXiv:1911.03803*.

Rose, G. D.; Geselowitz, A. R.; Lesser, G. J.; Lee, R. H.; and Zehfus, M. H. 1985. Hydrophobicity of amino acid residues in globular proteins. *Science*, 229(4716): 834–838.

Rubner, Y.; Tomasi, C.; and Guibas, L. J. 2000. The earth mover's distance as a metric for image retrieval. *International journal of computer vision*, 40(2): 99.

Santambrogio, F. 2015. Optimal transport for applied mathematicians. *Birkäuser, NY*, 55(58-63): 94.

Schiebinger, G.; Shu, J.; Tabaka, M.; Cleary, B.; Subramanian, V.; Solomon, A.; Gould, J.; Liu, S.; Lin, S.; Berube, P.; et al. 2019. Optimal-transport analysis of single-cell gene expression identifies developmental trajectories in reprogramming. *Cell*, 176(4): 928–943.

Sihn, D.; and Kim, S.-P. 2019. A spike train distance robust to firing rate changes based on the earth mover's distance. *Frontiers in Computational Neuroscience*, 13: 82.

Sommerfeld, M.; Schrieber, J.; Zemel, Y.; and Munk, A. 2019. Optimal transport: Fast probabilistic approximation with exact solvers. *Journal of Machine Learning Research*, 20(105): 1–23.

Tenenbaum, J. B.; Silva, V. d.; and Langford, J. C. 2000. A global geometric framework for nonlinear dimensionality reduction. *science*, 290(5500): 2319–2323.

VanRullen, R.; Guyonneau, R.; and Thorpe, S. J. 2005. Spike times make sense. *Trends in neurosciences*, 28(1): 1–4.

Victor, J. D.; and Purpura, K. P. 1997. Metric-space analysis of spike trains: theory, algorithms and application. *Network: computation in neural systems*, 8(2): 127–164.

Villani, C.; et al. 2009. *Optimal transport: old and new*, volume 338. Springer.

Wang, Z.; Yan, W.; and Oates, T. 2017. Time series classification from scratch with deep neural networks: A strong baseline. In *2017 International joint conference on neural networks (IJCNN)*, 1578–1585. IEEE.

Weed, J.; and Bach, F. 2019. Sharp asymptotic and finite-sample rates of convergence of empirical measures in Wasserstein distance. *Bernoulli*, 25(4A): 2620–2648.