A Transportation L^p Distance for Signal Analysis

Matthew Thorpe[1], Serim Park[1], Soheil Kolouri[3], Gustavo K. Rohde[2], and Dejan Slepčev[1]

[1]Carnegie Mellon University, Pittsburgh, PA 15213, USA
[2]University of Virginia, Charlottesville, VA 22908, USA
[3]HRL Laboratories, Malibu, CA 90265, USA

September 2016

Abstract

Transport based distances, such as the Wasserstein distance and earth mover's distance, have been shown to be an effective tool in signal and image analysis. The success of transport based distances is in part due to their Lagrangian nature which allows it to capture the important variations in many signal classes. However these distances require the signal to be nonnegative and normalized. Furthermore, the signals are considered as measures and compared by redistributing (transporting) them, which does not directly take into account the signal intensity. Here we study a transport-based distance, called the TL^p distance, that combines Lagrangian and intensity modelling and is directly applicable to general, non-positive and multi-channelled signals. The framework allows the application of existing numerical methods. We give an overview of the basic properties of this distance and applications to classification, with multi-channelled, non-positive one and two-dimensional signals, and color transfer.

1 Introduction

Enabled by advances in numerical implementation [3,8,38], and their Lagrangian nature, transportation based distances for signal analysis are becoming increasingly popular in a large range of applications. Recent applications include astronomy [5,11,12], biomedical sciences [2,18–20,56,60,61,64,65], colour transfer [6,10,36,44,45], computer vision and graphics [4,31,43,47,48,53,54], imaging [26,29,46], information theory [57], machine learning [1,7,13,24,27,35,55], operational research [49] and signal processing [39,42].

The success of transport based distances is due to the large number of applications that consider signals that are Lagrangian in nature (spatial rearrangements, i.e. transport, are a key factor when considering image differences). Many signals contain similar features for which transport based distances will outperform distances that only consider differences in intensity, such as L^p. Optimal transport (OT) distances, for example the earth mover's distance or Wasserstein distance, are examples of transport distances. However these distances do not directly account for signal intensity. The L^p distance is the other extreme, this distance is based on intensity and does not take into account Lagrangian properties.

In this paper we develop the TL^p distance introduced in [14] which combines both Lagrangian and intensity based modelling. Our aim is to show that by including both transport and intensity within the distance we can better represent the similarities between classes of data in many problems. For example, if a distance can naturally differentiate between classes, that is the within class distance is small compared to the between class separation, then the classification problem is made easier. This requires designing distances that can faithfully represent the structure within a given data set.

Optimal transport distances interpret signals as either probability measures or as densities of probability measures. This places restrictions on the type of signals one can consider. Probability measures must be non-negative, integrate to unity and be single-channelled. In order to apply OT to a wider class of signals one has to use ad-hoc methods to transform the signal into a probability measure. This can often dampen the features, for example renormalization may reduce the intensity range of a signal.

The TL^p distance does not need the signal to be a probability measure and therefore the above restrictions do not apply. Rather, the TL^p distance models the intensity directly. The framework is sufficiently general as to include signals on either a discrete or continuous domains that can be negative, multi-channelled and integrate to an arbitrary value.

Another property of OT, due to the lack of intensity modelling, is its insensitivity to high frequency perturbations. This is due to transport being on the order of the wavelength of the perturbation. By modelling the intensity directly, and therefore accounting for amplitude, the TL^p distance does not suffer this property.

The aim of this paper is to develop the TL^p framework and demonstrate its applicability in a range of applications. We consider classification problems on data sets where we show that the TL^p better represents the underlying geometry, i.e. achieves a better between class to within class distance, than popular alternative distances.

We also consider the colour transfer problem in a context where spatial information, as well as intensity, is important. To apply standardised tests in applications such as medical imaging it is often necessary to normalise colour variation [23,32,52].

One solution is to match the means and variance of each colour channel (in some colourspace e.g. RGB or LAB). However, by transferring the colour of one image onto the other it is possible to recolour an image with *exactly* the same colour profile.

A popular method is to use the OT distance on the histogram of images [6, 10, 36, 44, 45]. This allows one to take into account the intensity of pixels but includes no spatial information. The TL^p distance is able to include both spatial and intensity information.

Our methodology, therefore, has more in common with registration methods that aim to find a transformation that maximizes the similarity between two images where our measure of similarity includes both spatial and intensity information. One should compare our approach to [20] where the authors develop a numerical method for the Monge formulation of OT with the addition of an intensity term for image warping and registration. However, unlike the method presented in [20], the formulation presented here defines a metric.

Paper Overview. The outline for this paper is the following. In the next section we review OT and give a formal definition of the TL^p distance followed by examples to illustrate its features and to compare with the OT and L^p distances. In Section 3 we give a more general definition and explain some of its key properties. In Section 4 we include applications of the TL^p distances. We first consider classification on synthetic one and two dimensional, non-positive signals with no assumption on total mass and to real-world multivariate signals and two-dimensional images. A further application to the colour transfer problem is then given. Conclusions are given in Section 5.

2 Formal Definitions and Examples

2.1 Review of Optimal Transport and the TL^p Distance

We begin by reviewing optimal transport in first the Kantorovich formulation and then the Monge formulation.

The Kantorovich Formulation of Optimal Transport. For measures μ and ν on $\Omega \subset \mathbb{R}^d$ with the same mass and a continuous cost function $c : \Omega \times \Omega \to [0, \infty)$ the Kantorovich formulation of OT is given by

$$\mathrm{OT}(\mu, \nu) = \min_{\pi} \int_{\Omega \times \Omega} c(x, y) \, \mathrm{d}\pi(x, y) \tag{1}$$

where the minimum is taken over probability measures π on $\Omega \times \Omega$ such that the first marginal is μ and the second marginal is ν, i.e. $\pi(A \times \Omega) = \mu(A)$ and $\pi(\Omega \times B) = \nu(B)$ for all open sets A and B. We denote the set of such π by $\Pi(\mu, \nu)$. We call measures $\pi \in \Pi(\mu, \nu)$ transport plans since $\pi(A \times B)$ is the amount of mass in A that is transferred to B. Minimizers π^* of $\mathrm{OT}(\mu, \nu)$, which we call optimal plans, exist when c is lower semi-continuous [62].

A common choice is $c(x, y) = |x - y|_p^p = \sum_{i=1}^d |x_i - y_i|^p$ in which case we define $d_{\mathrm{OT}}(\mu, \nu) = \sqrt[p]{\mathrm{OT}(\mu, \nu)}$. When $p = 2$ this is known as the Wasserstein distance and when $p = 1$ the earth mover's distance. With an abuse of notation we will sometimes write $d_{\mathrm{OT}}(f, g)$ when μ and ν have densities f and g respectively.

When μ has a continuous density then the support of any optimal plan π^* is contained on the graph of a function T^*. In particular this implies $\pi^*(A, B) = \mu(\{x : x \in A, T^*(x) \in B\})$ and furthermore that the optimal plan defines a mapping between μ and ν, see for example Figure 1a. This leads us to the Monge formulation of OT.

The Monge Formulation of Optimal Transport. An appealing property of optimal transport distances are their formulation in a Lagrangian setting. One can rewrite the optimal transport problem in the Monge formulation as

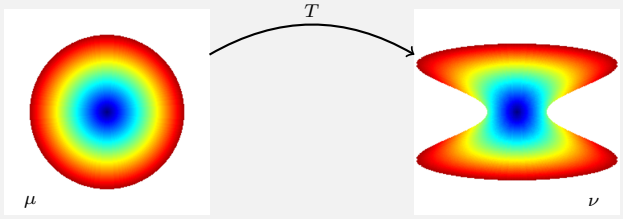
$$\mathrm{OT}_M(\mu, \nu) = \inf_{T} \int_{\Omega \times \Omega} c(x, T(x)) \, \mathrm{d}\mu(x) \tag{2}$$

where the infimum is taken over transport maps $T : \Omega \to \Omega$ that rearrange μ into ν, i.e. $\nu = T_{\#}\mu$ where we define the push-forward of μ onto the range of T by $T_{\#}\mu(A) = \mu(T^{-1}(A))$, see Figure 1b. This is now a non-convex optimization problem with nonlinear constraints. However when, for example, μ and ν have densities, then optimal transport maps T^* exist and give a natural interpolation between two measures. In particular when $c(x, y) = |x - y|^p$ the map $T_t(x) = (1 - t)x + tT^*(x)$ describes the path of particle x and furthermore the measure of μ pushed forward by T_t is the geodesic (shortest path) between μ and ν. This property has had many uses in transport based morphometry applications such as biomedical [2, 40, 59, 63], super-resolution [26] and has much in common with large deformation diffeomorphism techniques in shape analysis [16, 21].

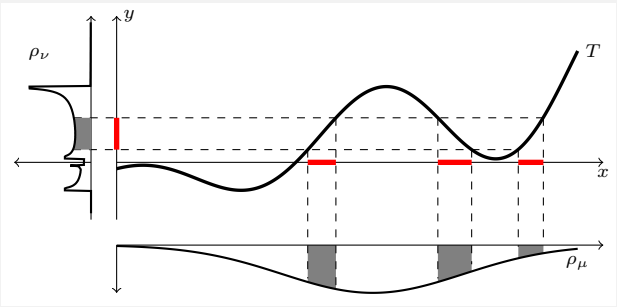
Optimal Transport in Signal and Image Processing. To further motivate our development of the TL^p distance we point out some features of optimal transport important to signal and image processing. We refer to [25] and references therein for more details and a review of the subject.

Key to the success of OT is the ability to provide generative models which accurately represent various families of data distributions. The success and appeal of OT owes to (1) ability to capture well the signal variations due to spatial rearrangements (shifts, translations, transport), (2) the OT distances are theoretically well understood and have appealing features (for example Wasserstein distance has a Riemannian structure and geodesics can be characterized), (3) efficiency and accuracy of numerical methods, (4) simplicity compared to other Lagrangian methods such as large deformation diffeomorphic metric mapping.

The Monge formulation of OT defines a mapping between images which has been used in, for example, *image registration* [17–20, 37, 61, 65] where one wishes to find a common geometric reference frame between two or more images. In addition to the properties listed above the success of OT is due to the fact that (5) the Monge problem is symmetric (i.e. if T is the optimal map from the first image to the second, then T^{-1} is the optimal map from the second image to the first) and (6)



(a) The Monge formulation of OT looks for maps that rearrange μ into ν. The optimal transport map between the uniform measure on the circle and the uniform measure on a bean (with the same mass) is given above. The mapping is represented by the colouring.



(b) The measure $T_{\#}\mu$ is represented with support on the y-axis and the measure μ on the x-axis in the above graph. The pushforward measure $T_{\#}\mu(A)$ is, by definition, equal to $\mu(T^{-1}(A))$. When T is differentiable, one-to-one and $\mathrm{d}\mu(x) = \rho_\mu(x)\ \mathrm{d}x$, $\mathrm{d}\nu(y) = \rho_\nu(y)\ \mathrm{d}y$ then $\nu = T_{\#}\mu$ is equivalent to $\det(DT(x))\rho_\nu(T(x)) = \rho_\mu(x)$.

Figure 1: On the left an example optimal transport map for OT, on the right an illustration of the pushforward measure.

OT provides a landmark-free and parameter-free registration scheme.

We now introduce the TL^p distance in the simplest setting.

The Transportation L^p Distance. In this paper we use the TL^p distance (introduced in more generality in the next section), for functions $f, g : \Omega \to \mathbb{R}^m$ defined by

$$d^p_{TL^p}(f,g) = \min_{\pi} \int_{\Omega\times\Omega} |x-y|_p^p + |f(x)-g(y)|_p^p \ \mathrm{d}\pi(x,y)$$

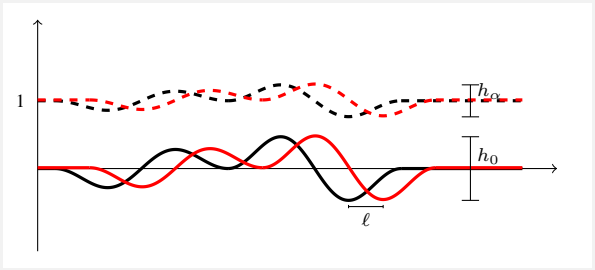
where the minimum is taken over all probability measures π on $\Omega \times \Omega$ such that both the marginals are the Lebesgue measure $\mathcal{L}$ on Ω, i.e. $\pi \in \Pi(\mathcal{L}, \mathcal{L})$. This can be understood in two ways. The first is as an optimal transport distance of the Lebesgue measure with itself and cost $c(x,y) = |x-y|_p^p + |f(x)-g(y)|_p^p$. This observation allows one to apply existing numerical methods for OT where the effective dimension is d (recall that $\Omega \subseteq \mathbb{R}^d$). The second is as an OT distance between the Lebesgue measure raised onto the graphs of f and g. That is, given $f, g : \Omega \to \mathbb{R}$ then we define the measures $\tilde{\mu}, \tilde{\nu}$ on the graphs of f and g by $\tilde{\mu}(A \times B) = \mathcal{L}(\{x : x \in A, f(x) \in B\})$ and $\tilde{\nu}(A \times B) = \mathcal{L}(\{y : y \in A, g(y) \in B\})$ for any open sets $A \subseteq \Omega, B \subseteq \mathbb{R}^m$. The TL^p distance between f and g is the OT distance between $\tilde{\mu}$ and $\tilde{\nu}$. Transport in TL^p is of the form $(x, f(x)) \mapsto (y, g(y))$ and therefore has two components. We refer to horizontal transport as the transport $x \mapsto y$ in Ω, and vertical transport as the transport $f(x) \mapsto g(y)$. In the next section we discuss the behaviour of TL^p through three examples.

2.2 Examples Illustrating the Behaviour of TL^p

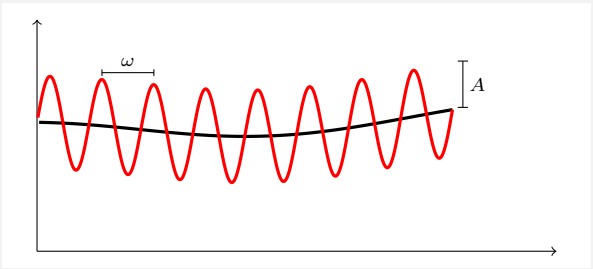
No mass renormalization. Unlike for OT, in the TL^p distance there is no need to assume that f and g are non-negative or that they have the same mass. If a signal is negative then a typical (ad-hoc) fix in OT is to add a constant to make the signal non-negative before computing the distance. How to choose this constant is often unclear unless a lower bound is known a-priori. Furthermore this may damage sensitivity to translations as the defining features of the signal become compressed. For example, considering the functions in Figure 2a, let $g = f(\cdot - \ell)$ be the translation of f. OT will lose sensitivity when comparing $\hat{f} = \frac{f+\alpha}{\int (f+\alpha)}$ and $\hat{g} = \frac{g+\alpha}{\int (g+\alpha)}$. In particular $d_{\mathrm{OT}}(\hat{f}, \hat{g})$ scales with the height of the renormalised function, which is of the order of $\frac{1}{\alpha}$, and the size of the shift: $d_{\mathrm{OT}}(\hat{f}, \hat{g}) \propto \frac{h_0 \ell}{\alpha}$ where h_0 is the height of f. To ensure positivity one must choose α large but this also implies a small OT distance. Note also that both L^p and TL^p are invariant under adding a constant whereas OT is not.

Sensitivity to High Frequency Perturbations. The TL^p distance inherits sensitivity to high frequency perturbations from the L^p norm. For example, let $g = f + A\xi$ where ξ is high frequency perturbation with amplitude A and wavelength ω. Then the distance moved by each particle in the Monge formulation of OT is on the order of the wavelength ω of ξ, which is small, and independent of the amplitude A. On the other hand both the TL^p distance and the L^p distance are independent of the wavelength but scale linearly with amplitude, see Figure 2b. In particular OT is insensitive to high frequency noise regardless of how large the amplitude whereas both TL^p and L^p scale linearly with the amplitude.

Ability of TL^p to Track Translations. Another desirable property of both TL^p and OT are their ability to keep track of translations for further than L^p. Let $f = A\chi_{[0,1]}$ be the indicator function of the set $[0,1]$ on $\mathbb{R}$ scaled by $A > 1$ and $g(x) = f(x - \ell)$ the translation of f by ℓ. Once $\ell > 1$ then L^p can no longer tell how far apart two humps are. On the other hand OT can track the hump indefinitely. In this example the TL^p distance couples the graphs of f and g in one of three ways, see Figure 3. The first is when the transport is horizontal only in the graph (Figure 3 top left). In the second (top right) there is a mixture of horizontal and vertical transport. And in the third



(a) f (solid, black) and g (solid, red) are translations of each other, the shifted and renormalised signals are $\hat{f} = \frac{f+\alpha}{\int (f+\alpha)}$ (dashed, black) and $\hat{g} = \frac{g+\alpha}{\int (g+\alpha)}$ (dashed, red). $d_{\mathrm{OT}}(\hat{f}, \hat{g}) = O\left(\frac{h_0 \ell}{\alpha}\right)$ whilst TL^p requires no renormalization.



(b) f (black) is a low frequency signal and g is a high frequency signal (red). $d_{\mathrm{OT}}(f, g) = O(\omega) \ll 1$ whilst $d_{TL^p}(f, g) = O(A)$.

Figure 2: A Comparison of TL^p with OT.

there is only vertical transport (bottom left), in which case the TL^p distance coincides with the L^p distance. One can calculate the range of the TL^p distance which is on the order of A.

3 Definitions and Basic Properties of TL^p

In the previous section we defined the TL^p distance for signals defined with respect to the Lebesgue measure. In this section we generalise to signals defined on a general class of measures. In particular we treat a signal as a pair (f, μ) where $f \in L^p(\mu; \mathbb{R}^m)$ for a measure $\mu \in \mathcal{P}_p(\Omega)$ (the set of probability measures with finite p^{th} moment) and a function $f : \Omega \to \mathbb{R}^m$. The general framework allows us to treat signal and discrete signals within the same framework as well as allowing one to design the underlying measure in order to emphasise certain parts of the signal. We are also able to compare signals with different discretisations. However, unless otherwise stated, $\mu = \nu$ is the Lebesgue measure. In addition there is no assumption on the dimension m of the codomain. This allows us to consider multi-channelled signals.

The TL^p_λ distance for pairs $(f, \mu) \in TL^p$ where

$$TL^p := \{(f, \mu) \,:\, f \in L^p(\mu), \mu \in \mathcal{P}_p(\Omega)\}$$

is defined by

$$d^p_{TL^p_\lambda}((f, \mu), (g, \nu)) = \min_{\pi \in \Pi(\mu,\nu)} \int_{\Omega \times \Omega} c_\lambda(x, y; f, g) \, \mathrm{d}\pi \quad (3)$$

$$c_\lambda(x, y; f, g) = \frac{1}{\lambda}|x - y|^p_p + |f(x) - g(y)|^p_p \quad (4)$$

and $\Pi(\mu, \nu)$ is the space of measures on $\Omega \times \Omega$ such that the first marginal is μ and the second marginal is ν. Note that if $f = g$ is constant then we recover the OT distance between the measures μ and ν. In the special cases, when $\mu = \nu = \mathcal{L}$ are the Lebesgue measure, we write $d_{TL^p_\lambda}(f, g) := d_{TL^p_\lambda}((f, \mathcal{L}), (g, \mathcal{L}))$ and, when $\lambda = 1$, $d_{TL^p}(f, g) := d_{TL^p_1}(f, g)$. The result of [14, Proposition 3.3] implies that $d_{TL^p_\lambda}$ is a metric on TL^p.

Proposition 3.1. *[14] For any $p \in [1, \infty]$, $(TL^p, d_{TL^p_\lambda})$ is a metric space.*

When $\mu = \nu = \mathcal{L}$ is the Lebesgue measure then an admissible plan is the identity plan: $\pi(A \times B) = \mathcal{L}(A \cap B)$. This implies that the TL^p_λ distance is bounded above by the L^p distance (for any λ).

In fact the parameter λ controls how close the distance is to an L^p distance. As $\lambda \to 0$ then the cost of transport: $\frac{1}{\lambda}\int_{\Omega\times\Omega} |x - y|^p_p \, \mathrm{d}\pi(x, y)$, is very expensive which favours transport plans that are approximately the identity mapping. Hence $d_{TL^p_0}(f, g) := \lim_{\lambda \to 0} d_{TL^p_\lambda}(f, g) = \|f - g\|_{L^p}$. The following result, and the remainder of the results in this section, can be found in [58].

Proposition 3.2. *[58] Let $f, g \in L^p$ (with respect to the Lebesgue measure). The TL^p_λ distance is decreasing as a function of λ and*

$$\lim_{\lambda \to 0} d_{TL^p_\lambda}(f, g) = \|f - g\|_{L^p}.$$

Moreover, if either the derivative of f or g is bounded then

$$d^p_{TL^p_\lambda}(f, g) \geq \begin{cases} \epsilon^{p-1}(\lambda)\|f - g\|^p_{L^p} & \text{if } p > 1 \\ \|f - g\|^p_{L^p} & \text{if } p = 1 \text{ and } \lambda < \frac{1}{\kappa} \end{cases}$$

where $\epsilon(\lambda) = \frac{1}{1 + (\lambda\kappa)^{\frac{1}{p-1}}}$ and $\kappa = \min\{\|Df\|^p_{L^\infty}, \|Dg\|^p_{L^\infty}\}$.

The above proposition implies that, when $p = 1$, if $\frac{1}{\lambda}$ is chosen larger than the length scale given by the derivative then the TL^1_λ distance is exactly the L^1 distance.

Recall that we can consider the TL^p_λ distance as an OT distance on the graphs of f and g. When there exists a map realising the minimum in $d_{TL^p_\lambda}$ then we can understand the transport as a map $(x, f(x)) \mapsto (y, g(y))$. We refer to the transport $x \mapsto y$ in the domain Ω as horizontal transport and transport $f(x) \mapsto g(y)$ in the codomain of f and g as vertical transport. We see that horizontal transport is favoured as $\lambda \to \infty$. For example, if we consider $f(x) = \chi_{[0,1]}$ and $g(x) = \chi_{[1,2]}$ defined on the

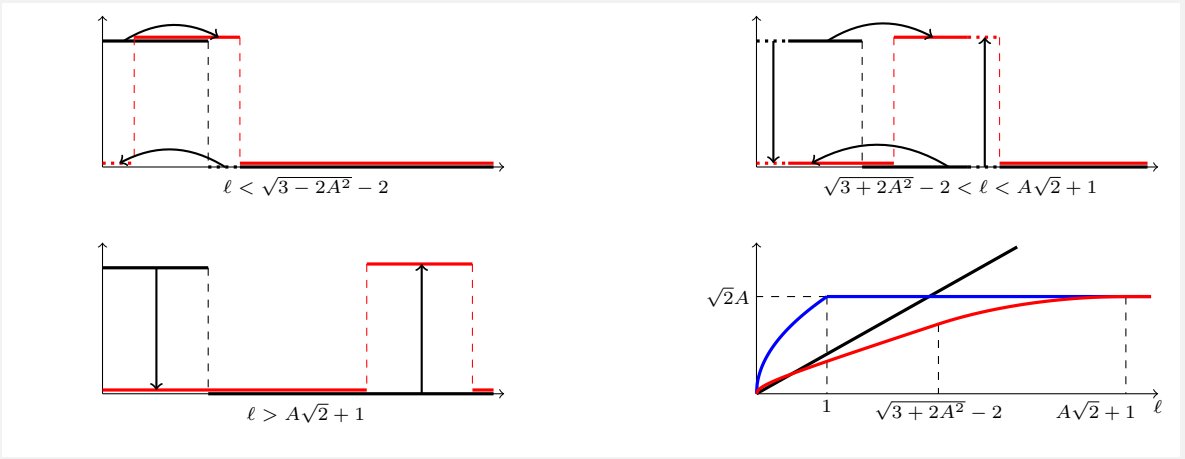


Figure 3: TL^2 transport between $f(x) = A\chi_{[0,1]}$ (black) and $g(x) = f(x-\ell)$ (red) and the TL^2 distance (red), L^2 distance (blue) and OT (black) between f and g (bottom right).

interval $[0,2]$ then the mapping $T(x) = x+1$ if $x \in [0,1]$ and $T(x) = x-1$ otherwise has cost

$$d^p_{TL^p_\lambda}(f,g) \leq \int_0^2 \frac{|x-T(x)|^p}{\lambda} + |f(x)-g(T(x))|^p \, \mathrm{d}x = \frac{2}{\lambda} \to 0 \quad \text{as } \lambda \to \infty.$$

In this example $d_{TL^p_\infty}(f,g) := \lim_{\lambda\to\infty} d_{TL^p_\lambda}(f,g) = 0$. The TL^p_∞ distance is an OT distance between the measures $f_\#\mu$ and $g_\#\nu$.

Proposition 3.3. *[58] Let $\Omega \subseteq \mathbb{R}^d$, $f, g : \Omega \to \mathbb{R}^m$ measurable functions and $\mu, \nu \in \mathcal{P}_p(\Omega)$ where $p \geq 1$, then*

$$\lim_{\lambda\to\infty} d_{TL^p_\lambda}((f,\mu),(g,\nu)) = d_{\mathrm{OT}}(f_\#\mu, g_\#\nu)$$

where d_{OT} is the OT distance (on $\mathcal{P}(\mathbb{R}^m)$) with cost $c(x,y) = |x-y|^p_p$.

As the example before the proposition showed, $d_{TL^p_\infty}(f,g)$ is not a metric, however is non-negative, symmetric and the triangle inequality holds.

We observe that when μ is a uniform measure (either in the discrete or continuous sense) the measure $f_\#\mu$ is the histogram of f. The OT distance between histograms is a popular tool in histogram specification. Minimizers to the Monge formulation of $d_{\mathrm{OT}}(f_\#\mu, g_\#\nu)$ define a mapping between the histograms $f_\#\mu$ and $g_\#\nu$ [36, 44, 45]. However this mapping contains no spatial information. If instead one uses minimizers to the Monge formulation of the TL^p_λ distance (5) ($\lambda < \infty$) then one can include spatial information in the histogram specification. We explore this further in Section 4.4 and apply the method to the colour transfer problem.

It is well known that there exists a minimizer (when c is lower semi-continuous) for OT. Since $d_{TL^p_\lambda}$ is closely related to an OT distance between measures in $\mathbb{R}^{d+m}$ (i.e. measures supported on graphs) then there exists a minimizer to TL^p_λ.

Proposition 3.4. *[58] Let $\Omega \subseteq \mathbb{R}^d$ be open and bounded, $f \in L^p(\mu)$, $g \in L^p(\nu)$ where $\mu, \nu \in \mathcal{P}(\Omega)$, $\lambda \in [0, +\infty]$ and $p \geq 1$. Under these conditions there exists an optimal plan $\pi \in \Pi(\mu,\nu)$ realising the minimum in $d_{TL^p_\lambda}((f,\mu),(g,\nu))$.*

As in the OT case it is natural to set the TL^p_λ problem in the Monge formulation (2)

$$d_{TL^p_\lambda}((f,\mu),(g,\nu)) = \inf_{T:T_\#\mu=\nu} \int_\Omega c_\lambda(x, T(x); f, g) \, \mathrm{d}\mu(x). \tag{5}$$

Minimizers to the above will not always exist. For example, consider when $f = g$ then the TL^p_λ distance is the OT distance between μ and ν. If one chooses $\mu = \frac{1}{3}\delta_{x_1} + \frac{1}{3}\delta_{x_2} + \frac{1}{3}\delta_{x_3}$ and $\nu = \frac{1}{2}\delta_{y_1} + \frac{1}{2}\delta_{y_2}$ where all of x_i, y_j are distinct then there are no maps $T : \{x_1, x_2, x_3\} \to \{y_1, y_2\}$ that pushforward μ to ν.

However, in terms of numerical implementation, an interesting and important case is when μ and ν are discrete measures (see also [62, pg 5, 14-15] for the following argument with the Monge OT problem). Let $\mu = \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$ and $\nu = \frac{1}{n}\sum_{i=1}^n \delta_{y_i}$ then $\pi = (\pi_{ij})_{i,j=1}^n \in \Pi(\mu,\nu)$ is a doubly stochastic matrix up to a factor of $\frac{1}{n}$, that is

$$\pi_{ij} \geq 0 \, \forall i,j, \ \sum_{i=1}^n \pi_{ij} = \frac{1}{n} \, \forall j \text{ and } \sum_{j=1}^n \pi_{ij} = \frac{1}{n} \, \forall i, \tag{6}$$

and the TL^p_λ distance can be written

$$d^p_{TL^p_\lambda}((f,\mu),(g,\nu)) = \min \sum_{i=1}^n \sum_{j=1}^n c_\lambda(x_i, y_j; f, g)\pi_{ij} \tag{7}$$

where the minimum is taken over π satisfying (6). It is known (by Choquet's Theorem) that the solution to this minimisation problem is an extremal point in the matrix set $\Pi(\mu,\nu)$. It is also known (by Birkhoff's Theorem) that extremal points in $\Pi(\mu,\nu)$ are permutation matrices. This implies that there exists an optimal plan π^* that can be written as $\pi^*_{ij} = \frac{1}{n}\delta_{j-\sigma(i)}$ for a permutation $\sigma : \{1,\dots,n\} \to \{1,\dots,n\}$. Hence there exists an optimal plan to the Monge formulation of TL^p_λ.

Proposition 3.5. *For any $f \in L^p(\mu)$ and $g \in L^p(\nu)$ where $\mu = \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$ and $\nu = \frac{1}{n}\sum_{j=1}^n \delta_{y_j}$ there exists a permutation $\sigma : \{1, 2, \ldots, n\} \to \{1, 2, \ldots, n\}$ such that*

$$d_{TL^p_\lambda}((f,\mu),(g,\nu)) = \frac{1}{n}\sum_{i=1}^{n} c_\lambda(x_i, x_{\sigma(i)}; f, g).$$

The above theorem implies that in the discrete case there exists optimal plans (which are matrices) which will be sparse. In particular, π^* is an $n \times n$ matrix with only n non-zero entries. This motivates the use of numerical methods that can take advantage of expected sparsity in the solution (e.g. iterative linear programming methods such as [38]).

4 TL^p in Multivariate Signal and Image Processing

Written in the form (3) the TL^p distance is an optimal transport problem between the measures μ and ν with the cost function c given by (4) and which depends upon f and g. Hence, to compute TL^p there are many algorithms for OT that we may apply, for example the multi-scale approaches of Schmitzer [51] and Oberman and Ruan [38], or the entropy regularized approaches of Cuturi [8] and Benamou, Carlier, Cuturi, Nenna and Peyre [3]. Our choice was the iterative linear programming method of Oberman and Ruan [38] for the multivariate signals which we find works well both in terms of accuracy and computation time. Our choice for the images was the entropy regularized solution due to Cuturi [3, 8]. Whilst this only produces an approximation of the TL^p distance we find it computationally efficient for 2D images. For convenience we include a review of the numerical methods in Appendix B.

With respect to choosing λ there are two approaches we could take. The first is to compute the TL^p distance for a range of λ and then use cross-validation. There are two disadvantages to this approach: we would still have to know the range of λ and computing the TL^p_λ distance for multiple choices of λ would considerably increase computation time. The second approach, and the one we use for each example in this section, is to estimate λ by comparing length scales and desired behaviour. In particular we choose λ so that both horizontal and vertical transport make a contribution. For the applications in this section we want to stay away from the asymptotic regimes $\lambda \approx 0$ and $\lambda \gg 1$. By balancing the vertical and horizontal length scale we can formally find an approximation of λ which in our results below works well.

We first consider two synthetic examples. Considering synthetic examples allows us to better demonstrate where TL^p will be successful. In particular synthetic examples can simplify the analysis and allow us to draw attention to features that may be obscured in real world applications.

The first synthetic example considers three classes where we can analytically compute the within class distances and between class separation. This allows us to compare how well we expect TL^p to perform in a classification problem.

The second synthetic example uses simulated 2D data from one-hump and two-hump functions. We test how well TL^p recovers the classes and compare with OT and L^p.

Our first real world application is to classifying multivariate times series and 2D images. We choose a multivariate time series data set where we expect transport based methods to be successful but it is not clear how one could apply OT distances (one would want to define a 'multi-valued measure'). Our chosen data set consists of sequences of sign language data (we define the data set in more detail shortly) which contains the position of both hands (parametrised by 22 variables) at each time. The TL^p distance can treat these signals as functions $f : [0, 1] \to \mathbb{R}^{22}$. We expect to see certain features in the signals however these may be shifted based on the speed of the speaker. The second data set contains 2D images that must be normalised in order to apply the OT distance, this distorts some of the features leading to a poor performance.

The second real world application is to histogram specification and colour transfer. Histogram specification or matching, where one defines a map T that matches one histogram with another, is widely used to define a colour transfer scheme. In particular let $f : \{x_i\}_{i=1}^N \to \mathbb{R}^3$ represent a colour image by mapping pixels x_i to a colour $f(x_i)$ (for example in RGB space), one defines a multidimensional histogram of colours on an image by $\varphi(c) = \frac{1}{N}\#\{x_i : f(x_i) = c\}$. For colour images the histogram φ is a measure on $\mathbb{R}^3$. For notational clarity we will call φ the colour histogram. One can equivalently define a histogram for grayscale images as a measure on $\mathbb{R}$.

Let φ and ψ be two colour histograms for images f and g respectively. The OT map T defines a rearrangement of φ onto ψ, that is $\psi = T_\#\varphi$. In colour transfer the map T is used to colour the image f using the palette of g by $\hat{f}(x) = g(T(x))$.

The histogram contains only intensity information and in particular there is no spatial dependence. Using the TL^p_λ-optimal map we define a spatially correlated histogram specification and explain how this can be applied to the colour transfer problem.

4.1 1D Class Separation for Synthetic Data

Objective. We compare the expected classification power of TL^p, L^p and OT with three classes of 1D signals that differ by position (translations), shape (1 hump versus 2 hump) and frequency (hump versus chirp).

Data Sets. We consider data from three classes defined in Figure 4. The first class contains single hump function and the second class contains two hump functions. The third class consists of functions with one hump and one chirp, defined to be a high frequency perturbation of a hump. The classes are chosen to test the performance of TL^2 with L^2 and OT with regards to identifying translations (where we expect L^2 to do poorly) with a class containing high frequency perturbations (where we expect OT to do poorly).

Methods. For a distance to have good performance in classification and clustering problems it should be able to separate classes. To be able to quantify this we use the ratio of 'between class separation' to 'class coverage radius' that we define now.

($\mathcal{C}_1$) **One hump functions:** of the form

$$f = \chi_{[\ell,\ell+\alpha]}$$

where $\ell \in [0, 1-\alpha]$.

($\mathcal{C}_2$) **Two hump functions:** of the form

$$f = \frac{1}{2}\left(\chi_{[\ell,\ell+\alpha]} + \chi_{[\ell+\beta+\alpha,\ell+\beta+2\alpha]}\right)$$

where $\ell \in [0, 1-\beta-2\alpha]$.

($\mathcal{C}_3$) **One hump, one chirp functions:** of the form

$$f = \sum_{i=0}^{\frac{\alpha}{\gamma}-1} \chi_{\left[\ell+i\gamma,\ell+\frac{(2i+1)\gamma}{2}\right]} + \frac{1}{2}\chi_{[\ell+\beta+\alpha,\ell+\beta+2\alpha]}$$

where $\ell \in [0, 1-\beta-2\alpha]$.

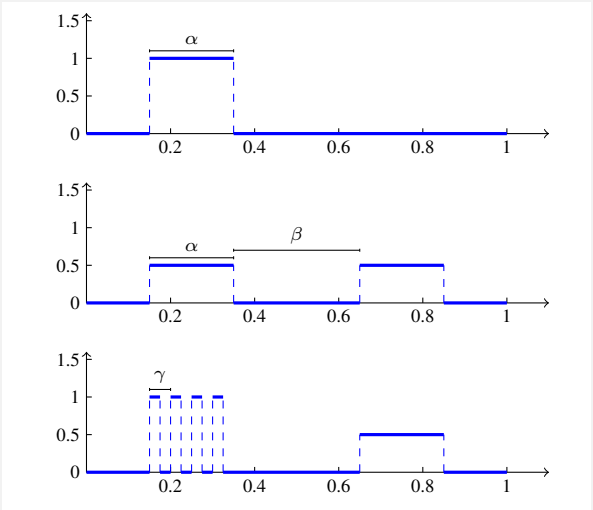


Figure 4: For fixed $\alpha, \beta, \gamma \in (0,1)$ where $\beta > \alpha \gg \gamma$ the definition of the classes $\mathcal{C}_i$.

Let $\mathcal{C}_i^N = \{f_j^i\}_{j=1}^N$ be a sample of N functions from class $\mathcal{C}_i$. For a given radius r we let $G_i(r)$ be the graph defined by connecting any two points in $\mathcal{C}_i^N$ with distance less than r. The distance will be defined using the TL^p_λ, L^2 and OT metrics. Let $R_{TL^p_\lambda}(\mathcal{C}_i^N)$ be the smallest r such that $G_i(r)$ is a connected graph using the TL^p_λ metric. Analogously we can define R_{L^p} and R_{OT}.

We define 'between class separation' as the Hausdorff distance between classes:

$$d_{H,\rho}(\mathcal{C}_i^N, \mathcal{C}_j^N) = \max\left\{\sup_{f\in\mathcal{C}_i^N}\inf_{g\in\mathcal{C}_j^N}\rho(f,g), \sup_{g\in\mathcal{C}_j^N}\inf_{f\in\mathcal{C}_i^N}\rho(f,g)\right\}$$

where we will consider ρ to be one of the TL^2, L^2 or OT metrics. Large values of $d_{H,\rho}(\mathcal{C}_i^N, \mathcal{C}_j^N)$ imply that the classes $\mathcal{C}_i^N$ and $\mathcal{C}_j^N$ are well separated.

When $R_\rho(\mathcal{C}_i^N) \le d_{H,\rho}(\mathcal{C}_i^N, \mathcal{C}_j^N)$ then we say that the class $\mathcal{C}_i^N$ is separable from class $\mathcal{C}_j^N$ since for any $f \in \mathcal{C}_i^N$ the nearest neighbour in $\left(\mathcal{C}_i^N \cup \mathcal{C}_j^N\right) \setminus \{f\}$ is also in class $\mathcal{C}_i^N$. We define the pairwise property

$$\kappa_{ij}(\rho; N) = \frac{\mathbb{E}d_{H,\rho}(\mathcal{C}_i^N, \mathcal{C}_j^N)}{\max\{\mathbb{E}R_\rho(\mathcal{C}_i^N), \mathbb{E}R_\rho(\mathcal{C}_j^N)\}}$$

where we take the expectation over sample classes $\mathcal{C}_i^N$. We will assume that the distribution over each class is uniform in the parameter ℓ. When $\kappa_{ij}(\rho; N) > 1$ then we expect classes $\mathcal{C}_i^N$ and $\mathcal{C}_j^N$ to be separable from each other.

As a performance metric we use the smallest value of N such that $\kappa_{ij}(\rho; N) \ge 1$. We let

$$N_{ij}^*(\rho) = \min\{N : \kappa_{ij}(\rho; N) \ge 1\}.$$

This measures how many data points we need in order to expect a good classification accuracy.

Results. We leave the calculation to the appendix but the conclusion is

$$N_{12}^*(TL^2) < N_{12}^*(\mathrm{OT}) < N_{12}^*(L^2)$$
$$N_{13}^*(TL^2) < N_{13}^*(\mathrm{OT}) < N_{13}^*(L^2)$$
$$N_{23}^*(TL^2) < N_{23}^*(\mathrm{OT}) < N_{23}^*(L^2).$$

In each case the TL^2 distance outperforms L^2 and OT.

In each class the L^2 distance has a larger value of R. This implies a larger data set is needed to accurately cover each class. This is due to the Lagrangian nature of signals within each class (translations) that is poorly represented by L^2. OT has the lowest (and therefore best) value of R in each class. Since each class is Lagrangian then the OT distance is very small between functions of the same class.

When considering between class separation the TL^2 and L^2 distances coincide and give a bigger (and better) between class distance than OT. Since the class $\mathcal{C}_3$ can be written as a high frequency perturbation of functions in the class $\mathcal{C}_2$ then the OT distance struggles to tell the difference between these classes. The distance $d_{H,\mathrm{OT}}(\mathcal{C}_2^N, \mathcal{C}_3^N)$ is comparatively small so that one needs more data points in order to fully resolve these classes. We see a similar effect when considering $d_{H,\mathrm{OT}}$ for the other classes.

4.2 2D Classification for Synthetic Data

Objective. We use simulated data to illustrate better separation of TL^p compared to L^p and OT distances for 2D data from two classes of 1-hump and 2-hump functions.

Data Sets. The data set consists of two dimensional images simulated from the following classes

$$\mathbb{P} = \left\{ p\lfloor_{[0,1]^2} : p(x) = \alpha\phi(x|\gamma,\sigma),\ \gamma \sim \text{unif}([0,1]^2),\ \alpha \sim \text{unif}([0.5,1]) \right\}$$

$$\mathbb{Q} = \left\{ q\lfloor_{[0,1]^2} : q(x) = \alpha\phi(x|\gamma_1,\sigma) - \alpha\phi(x|\gamma_2,\sigma),\ \gamma_1,\gamma_2 \overset{\text{iid}}{\sim} \text{unif}([0,1]^2),\ \alpha \sim \text{unif}([0.5,1]) \right\}$$

where $\phi(\cdot|\gamma,\sigma)$ is the multivariate normal pdf with mean $\gamma \in \mathbb{R}^2$ and co-variance $\sigma \in \mathbb{R}^{2\times 2}$. We choose $\sigma = 0.01 \times \text{Id}$ where Id is the 2×2 identity matrix. The first class, $\mathbb{P}$, are the set of multivariate Gaussians restricted to $[0,1]^2$ with mean uniformly sampled in $[0,1]^2$ and weighted by α uniformly sampled in $[0.5,1]$. The second class, $\mathbb{Q}$, are the set of weighted differences between two Gaussian pdf's restricted to $[0,1]^2$ with means γ_1,γ_2 sampled uniformly in $[0,1]^2$. Note that the second class contains non-positive functions. See Figure 5 for examples from each class.

We simulate 25 from each set and denote the resulting set of functions by $\mathcal{F} = \{f_i\}_{i=1}^N$ where $N = 50$.

Methods. Let $(\{f_i\}_{i=1}^N, D_{TL^2_\lambda})$ be a finite dimensional metric space where $D_{TL^2_\lambda}$ is the $N \times N$ matrix containing all pairwise distances is the TL^2_λ distance i.e. $D_{TL^2_\lambda}(i,j) = d_{TL^2_\lambda}(f_i,f_j)$. Similarly for $(\{f_i\}_{i=1}^N, D_{L^2})$ and $(\{f_i\}_{i=1}^N, D_{OT})$ where the optimal transport distance is defined by $d_{\text{OT}}(f,g) = \sqrt{\text{OT}(f,g)}$ and OT is given by (1) for $c(x,y) = |x-y|_2^2$.

To apply the optimal transport distance we need to renormalise so that signals are all non-negative and integrate to the same value. We do this by applying the nonlinear transform $\mathcal{N}(f) = \frac{f-\beta}{\int(f-\beta)}$ where $\beta = \min_{f\in\mathcal{F}} \min_{x\in[0,1]^2} f(x)$. Neither the L^2 nor TL^2_λ distances require normalisation.

We use non-metric multidimensional scaling (MDS) [28] to represent the graph in k dimensions. More precisely the aim is to approximate $(\{f_i\}_{i=1}^N, D_\cdot)$ by a metric space $(\{x_i\}_{i=1}^N, D_{|\cdot|_2})$ embedded in $\mathbb{R}^k$ ($D_{|\cdot|_2}$ is the matrix of pairwise distances using the Euclidean distance, i.e. $D_{|\cdot|_2}(i,j) = |x_i - x_j|_2$). This is done by minimising the stress S defined by

$$S_{TL^2_\lambda}(k) = \frac{\sum_{i,j=1}^N \left(|x_i - x_j|_2^2 - F(D_{TL^2_\lambda}(i,j)) \right)^2}{\sum_{i,j=1}^N |x_i - x_j|_2^2}$$

over $\{x_i\}_{i=1}^N \subset \mathbb{R}^k$ and monotonic transformations $F : [0,\infty) \to [0,\infty)$, with $S_{L^2_\lambda}$, S_{OT} defined analogously. The classical solution to finding the MDS projection (for Euclidean distances) is to use the k dominant eigenvectors of the matrix of squared distances, after double centring, as coordinates weighted by the square root of the eigenvalue. More precisely, define $D^{(2)} = -\frac{1}{2} J \left[|f_i - f_j|^2 \right]_{ij} J$ where $J = \text{Id} - \frac{1}{N}\mathbb{I}$ and $\mathbb{I}$ is the $N \times N$ matrix of ones. Let Λ_k be the matrix with the k largest eigenvalues of $D^{(2)}$ on the diagonal and E_k to be the corresponding matrix of eigenvectors. Then $X = E_k \Lambda_k^{\frac{1}{2}}$ is the MDS projection. Increasing the dimension of the projected space k leads to a better approximation. In Figure 5 we show the projection in L^2, TL^2 and the OT distance for $k = 2$ as well as the dependence of k on S for each choice of distance.

Results. Our results in Figure 5 show that TL^2 is the better distance for this problem. There is no separation in either L^2 or OT whereas TL^2 completely separates the data. It should not therefore be surprising that the 1NN classifier in TL^2 outperforms the other distances. In fact, using 5 fold cross validation (CV) we get 100% accuracy in TL^2, compared to 72% in L^2 and 86% in OT. In addition we see that the stress S_ρ is much smaller and converges quickly to zero for TL^2 which indicates that the TL^2 distance is, in this problem, more amenable to a low dimensional representation than either OT or L^2.

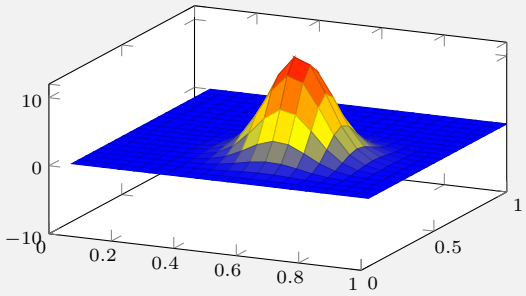
4.3 Classification with Real World Data Sets

Objective. We evaluate TL^2_λ as a distance to classify real world data sets where spatial and intensity information is expected to be important and compare with popular alternative distances. We choose one dataset which is of the type multivariate time series and a second data set consisting of images.

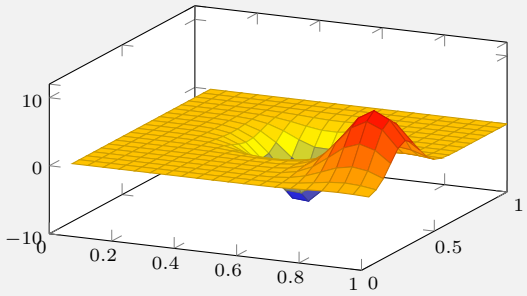
Data Sets. We use two data sets. The first is the *AUSLAN* [22, 30] data set which contains 95 classes (corresponding to different words) from a native AUSLAN speaker (Australian Sign Language) using 22 sensors on a CyberGlove (recording position of x, y, z axis, roll, yaw, pitch for left and right hand). Therefore signals are considered as functions from $\{t_1, t_2, \ldots t_N\}$ to $\mathbb{R}^{22}$. We used the following 25 (out of the 95) classes: alive, all, boy, building, buy, cold, come, computer, cost, crazy, danger, deaf, different, girl, glove, go, God, joke, juice, man, where, which, yes, you and zero. There are 27 signals in each class which give a total of 675 signals.

We make two pre-processing steps. The first is to truncate each signal so it is 46 frames in length. Empirically we find that the signal is constant after the 46^{th} frame and therefore there is no loss of information in truncating the signal. The second preprocessing step is to normalise each channel independently. This is because some channels are orders of magnitude greater than others and would otherwise dominate each choice of distance.

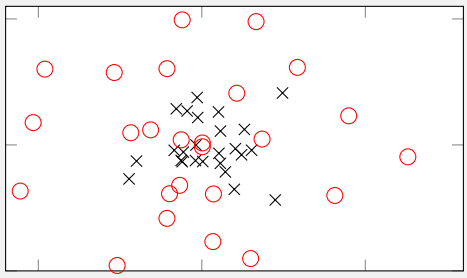
The second data set we use is a subset of the 28×28 Caltech Silhouettes database [33]. This data set was derived from the Caltech 101 data set [9], which consists of images from 101 categories, by finding and filling in the outline for the object of focus in each image. See Figure 6b for examples. The subset we uses consists of the following 11 images: anchor, barrel, crocodile head, dollar bill, emu, gramophone, pigeon, pyramid, rhino, rooster and stegosaurus. The number of images in each class varied from 42 to 59. There were 565 images in total.



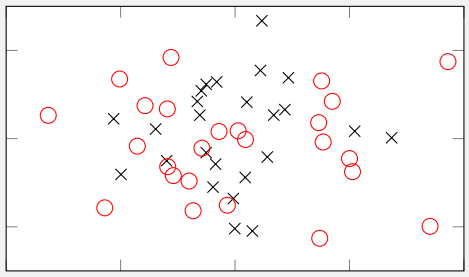
(a) Example signal from class $\mathbb{P}$.



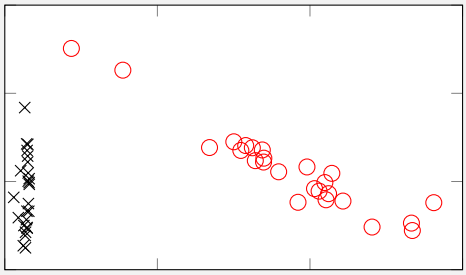
(b) Example signal from class $\mathbb{Q}$.



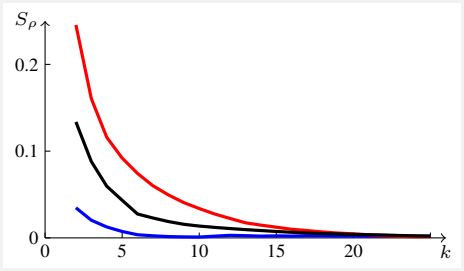
(c) MDS projection in L^2.



(d) MDS projection in OT.



(e) MDS projection in TL^2.



(f) Dependence of the stress S_ρ on the dimension of the projected space k for L^2 (red), OT (black) and TL^2 (blue).

Figure 5: Example signals and results for the synthetic data in Section 4.2.

Methods. For the multivariate time series we compare the performance of a 1NN classifier using the L^2 and TL^2_λ distances as well as the state-of-the-art method dynamic time warping [15]. There are three common variations of dynamic time warping. One can apply dynamic time warping directly to the signals f and g (denoted by DTW), to the derivative f' of the signals (denoted by DDTW) and to a weighted average of DTW and DDTW (denoted by WDTW). We define

$$d_{DDTW}(f,g) = d_{DTW}(f',g')$$
$$d_{WDTW}(f,g) = \alpha d_{DTW}(f,g) + (1-\alpha) d_{DDTW}(f,g).$$

The parameter α is chosen by 5-fold 2nd depth cross validation.

One can define the analogous distances for L^2 and TL^2 by

$$d_{DL^2}(f,g) = d_{L^2}(f',g')$$
$$d_{DTL^2_\lambda}(f,g) = d_{TL^2_\lambda}(f',g')$$
$$d_{WL^2}(f,g) = \alpha d_{L^2}(f,g) + (1-\alpha) d_{DL^2}(f,g)$$
$$d_{WTL^2_\lambda}(f,g) = \alpha d_{TL^2_\lambda}(f,g) + (1-\alpha) d_{DTL^2_\lambda}(f,g).$$

We do not have to choose the same value of λ in TL^2 and DTL^2 however considering that signals are normalised, we will use the same value. Note that DL^2, DTW, DDTW, WDTW and DTL^2_λ are *not* metrics.

We remark that an alternative method for including derivatives in the TL^p distance would be to extend the signal to include the derivative. We briefly assume that f is defined over a continuous domain. Let $f : \mathbb{R} \to \mathbb{R}$, and $\tilde{f} = \left(f, \frac{\mathrm{d}f}{\mathrm{d}x}\right)$ then we define

$$d_{TW^{1,p}_\lambda}(f,g) = d_{TL^p_\lambda}(\tilde{f},\tilde{g}).$$

We take our notation $TW^{k,p}_\lambda$ from the Sobolev space notation where $W^{k,p}$ is the Sobolev space with k weak derivatives integrable in L^p. There is no reason to limit this to one derivative, and we may define $\tilde{f} = \left(f, \frac{\mathrm{d}f}{\mathrm{d}x}, \ldots, \frac{\mathrm{d}^k f}{\mathrm{d}x^k}\right)$ and

$$d_{TW^{k,p}_\lambda}(f,g) = d_{TL^p_\lambda}(\tilde{f},\tilde{g}).$$

When the signals are discrete one should use a discrete approximation of the derivative. In order to be consistent with previous

extensions of dynamic time warping we do not develop this approach here.

Dynamic time warping is only defined on time series so we are not able to apply it to the Caltech Silhouettes database. Instead we use the optimal transport distance (with $p = 2$). To apply the optimal transport distance each image $f \in \mathbb{R}^2 \to \{0, 1\}$ is normalised by $\hat{f}(x) = \frac{f(x)}{\int_{[0,1]^2} f(y)\,\mathrm{d}y}$. There is no normalisation for either L^2 or TL^2. We find the 1NN classifier using TL^2, L^2 and OT distances.

We will use $\lambda = 1$ in AUSLAN and $\lambda = 0.1$ in Caltech Silhouettes for the TL^2 based distances. The underlying measure μ is chosen to be the uniform measure defined on $[0, 1]$ or $[0, 1]^2$.

Results. We considered two methods for comparing the performance of each distance. The first is the 1NN classification accuracy in each distance. We use the 1NN classification accuracy as a measure as to how well each distance captures the underlying geometry. A higher accuracy implies closest neighbours are more likely to belong to the same class.

The results are given in Table 1 where we report error rates using 5-fold cross-validation. In terms of the 1NN classifier for the AUSLAN data set we see that TL^2 is significantly better than L^2 and is a modest improvement over dynamic time warping. And for the Caltech Silhouettes dataset OT performs poorly with TL^2 the best performer.

In the same spirit as Section 4.1 we define the performance metric $\kappa_{ij}(\rho)$ as the ratio of distance between class i and class j and the maximum class coverage radius of class i and class j. For the distance between classes we use the Hausdorff distance (see Section 4.1) and for the class coverage radius we use the minimum radius r such that connecting any two data points in class i closer than r defines a connected graph. We plot the results in Figures 6c and 6d. The x axis represents pairs of classes where for visual clarity we have ordered the pairs so that the $\kappa(L^2)$ is increasing. A large value of κ_{ij} indicates that it is easier to identify class i from class j whereas a small value indicates that identifying the two classes is a difficult problem.

For AUSLAN we see that TL^2 has, for the majority of pairs of classes, a larger value of κ_{ij} than L^2 and DTW and therefore better represents the class structure. For the Caltech Silhouettes dataset L^2 has the worst separation even though it outperformed OT in the 1NN test. The TL^2 distance is much more consistent than OT, we can see that although between some classes OT is the best distance with other classes OT does extremely poorly (worse than L^2). On the other hand TL^2 is better than L^2 for every pair of classes.

4.4 Histogram Specification and Colour Transfer with TL^p

Histogram specification and colour transfer. Histogram specification concerns the problem of matching one histogram onto another. For a function f on a discrete domain X the histogram is given by $f_{\#}\mu$ where μ is the uniform discrete measure supported on N points. We do not make any assumption on the dimension of the codomain of f (so that f may be multivalued and the histogram may be multidimensional). This coincides with the definition given in the introduction to the section, that is

$$f_{\#}\mu(y) = \frac{1}{N}\#\left\{x \in X \,:\, f(x) = y\right\}.$$

Given two functions $f : X \to \mathbb{R}^m$ and $g : Y \to \mathbb{R}^m$, with histograms φ and ψ respectively, histogram specification is the problem of finding a map $T : X \to Y$ such that $\psi = T_{\#}\varphi$.

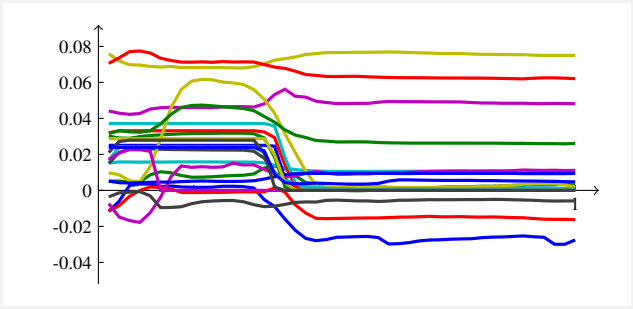
The colour transfer problem is the problem of colouring one image f with the palette of an exemplar image g. A common method used to solve this problem is to use histogram specification where T is the minimizer to Monge's optimal transport problem (2) between φ and ψ [6, 10, 36, 44, 45]. Let our colour space be denoted by $\mathcal{C}$ where for example if the colour space is 8 bit RGB then $\mathcal{C} = \{0, 1, \ldots, 255\}^3$. The colour histogram then defines a measure over $\mathcal{C}$. If we consider two such histograms φ and ψ corresponding to images $f : X \to \mathcal{C}$ and $g : Y \to \mathcal{C}$ respectively then a histogram specification is a map $T : \mathcal{C} \to \mathcal{C}$ that satisfies $\psi = T_{\#}\varphi$. The recoloured image $\hat{f} = g \circ T$ has the same colour histogram as g. The solution $\hat{f}$ is a recolouring of f using the palette of g.

If we consider grayscale images then $\mathcal{C} = [0, 1]$ and the optimal transport map (assuming it exists) is a monotonically increasing function. In particular this implies that if pixel x is lighter than pixel y (i.e. $f(x) > f(y)$) then in the recoloured image $\hat{f} = T \circ f$ pixel x is still lighter than pixel y. In this sense the OT solution preserves intensity ordering. But note that no spatial information is used to define T; only the difference in intensity between pixels is used and not the distance between pixels.

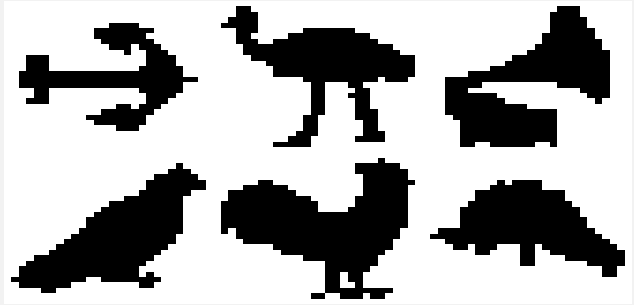
Spatially correlated histogram specification. Let φ and ψ be the histograms corresponding to images $f : X \to \mathbb{R}^m$ and $g : Y \to \mathbb{R}^m$ respectively. If we recall Proposition 3.3 then $\lim_{\lambda\to\infty} d_{TL^p_\lambda}((f, \mu), (g, \nu)) = d_{\mathrm{OT}}(f_{\#}\mu, g_{\#}\nu)$ (where μ and ν are the discrete uniform measures over the sets X and Y). For $\lambda < \infty$ the TL^p_λ distance includes spatial *and* intensity information. Hence the TL^p_λ distance provides a generalization of OT induced histogram specification.

Analogously to the OT induced histogram specification method we define the spatially correlated histogram specification to be histogram specification using the map $T : X \to Y$ which is a minimizer to Monge's formulation of the TL^p_λ distance (5). When the images are of the same size then, by Proposition 3.5 such a map exists. The recoloured image $\hat{f}$ of f is given by $\hat{f} = g \circ T$. Furthermore when the images are of the same size the map T is a rearrangement of the pixels in X and therefore the histograms are invariant under T. In particular the histogram of $\hat{f}$ is the same as the histogram of g.

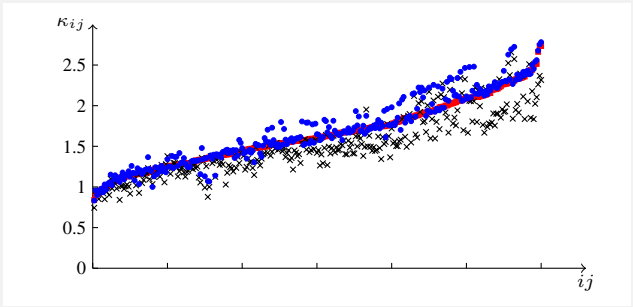
Although we propose the spatially correlated histogram specification as a method to incorporate spatial structure we now point out its value as a numerically efficient approximation to OT induced histogram specification for colour images. Motivated by Proposition 3.3 one expects that for large λ the TL^p_λ map is approximately the OT map between colour histograms. The OT problem is in the $\mathcal{C}$ space which, for colour images is 3 dimensional. However, the TL^p_λ problem is in the domain of the images, which is typically 2 dimensional. Hence one can



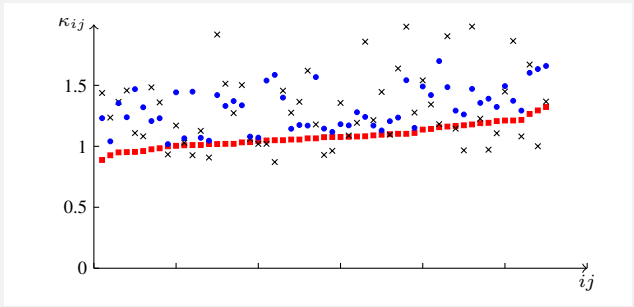
(a) Example signal from AUSLAN.



(b) Example signals from Caltech Silhouettes.



(c) A measure of the class identification of each pair of classes for L^2 (red squares), DTW (black crosses) and TL^2 (blue circles) for AUSLAN.



(d) A measure of the class identification of each pair of classes for L^2 (red squares), OT (black crosses) and TL^2 (blue circles) for Caltech Silhouettes.

Figure 6: Example signals and results for the data sets described in Section 4.3.

Dataset	L^2	DL^2	WL^2	DTW	DDTW	WDTW	TL^2_λ	DTL^2_λ	WTL^2_λ
AUSLAN	10.4	14.7	9.8	9.6	3.1	2.7	9.5	2.4	2.7

Dataset	L^2	OT	TL^2_λ
Caltech Silhouettes	12.8	19.6	11.0

Table 1: Error rates (%) for 1NN classification.

use TL^p_λ to approximate OT induced histogram specification in a lower dimensional space.

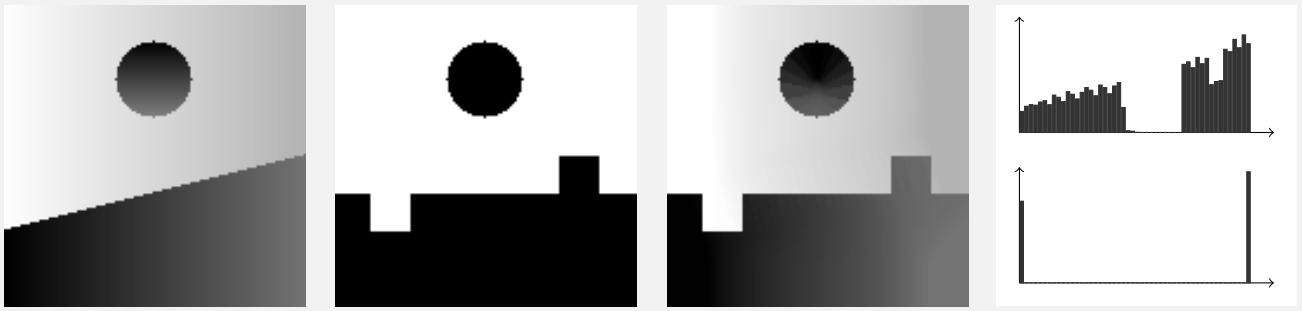
We briefly remark that histogram specification methods often include additional regularization terms. Such choices of regularization on the transport map include penalizing the gradients [10, 44, 45], sparsity [45] and average transport [41]. One could apply any of the above regularizations to spatially correlated histogram specification.

Examples. First, let us consider the grayscale images in Figure 7. The objective is to combine the shading of the first image with the geometry of the second image. We are motivated by the scenario where one wishes to combine information about a scene obtained by two different measurements: one where intensities (dynamical range) are well resolved, but the spatial resolution (geometry) is not well captured, and another where dynamical range is poorly captured, but the geometry is well resolved. We furthermore allow that the scenes captured may be somewhat different. The desire is to combine the images to obtain a single image with both good geometry and intensity. The solution we propose is to use spatially correlated histogram specification to re-shade the image with low quality intensity.

The result, as given in Figure 7, produces what we consider to be the desired output. The shading has been transferred and the geometry has not been lost. One is not able to apply histogram specification (induced by the OT map) due to the lack of existence of an optimal transport map from the histogram of the original image φ to the histogram exemplar image ψ. This is due to the histogram of the original image being a sum of two delta masses as in Figure 7d.

As a more challenging example we consider the colour images in Figure 8. The exemplar image contains a few trees with the northern lights in the background, whilst the other image has a few trees with a mostly clear sky in the background. The challenge is to recreate the northern lights in the second image.

As one would expect, in Figure 8f we see that the histogram specification induced by OT loses the spatial structure. Indeed, it is hard to recognise the northern lights. The spatially correlated histogram specification solution does a much better job at preserving the ordering locally. As λ increases it becomes



(a) Exemplar image. (b) Original image to be shaded. (c) The TL^p_λ solution. (d) The grayscale histograms of the exemplar image (top) and the original image (bottom).

Figure 7: Spatially correlated histogram specification of synthetic grayscale images.

cheaper to match pixels that are further apart and therefore, for large λ, the matching does not preserve the local structure in the exemplar image.

5 Conclusions

In this paper we have developed and applied a distance that directly accounts for the intensity of the signal within a Lagrangian framework. This differs from optimal transport that does not directly measure intensity and the L^p distance which measures intensity only. Through applications we have shown the potential of this distance in signal analysis.

The distance is widely applicable, unlike optimal transport the distance does not require treating signals as measures. Treating a signal as a measure implies the following constraints: non-negative signals, conservation of mass and single channelled signals. None of these assumptions are necessary for the TL^p_λ distance. Furthermore the framework is general enough to include discrete and continuous signals as well as allowing practitioners to emphasise features which in many cases should allow for a better representation of data sets, for example one could include derivatives.

Efficient existing methods, such as entropy regularized or multi-scale linear programming, for optimal transport are applicable to the TL^p_λ distance. In fact any numerical method for optimal transport that can cope with arbitrary cost functions is immediately available.

Via the representation as an optimal transport distance between measures supported on graphs we expect many other results for OT to carry through to TL^p. For example, one could extend the LOT method [63] for signal representation and analysis to the TL^p framework. This would allow pairwise distances of a data set to be computed with numerical cost that is linear in number of images. We leave the development for future work.

The applications we considered were classification and histogram specification in the context of colour transfer. For classification we chose data sets with a Lagrangian nature but were either multi-channelled (so that optimal transport distances are not available) or non-positive (in which case one has to rescale in order to apply optimal transport). We showed the TL^p_λ distance better represented the underlying geometry. For the colour transfer problem we defined a spatially correlated histogram specification method which produced more visually appealing results when combining the colour of one image with the geometry of another.

Although the main motivation was to develop a distance which better represents Lagrangian data sets we also note that the TL^p_λ distance provides a numerically efficient approximation for the optimal transport induced histogram specification method by, for 2-dimensional images colour images, reducing the effective dimension of the problem from three for optimal transport to two for TL^p_λ. We also observe that the effective dimension of multi-channelled time signals is one. In particular the effective dimension is independent of the number of channels.

The applications we have considered are for demonstration on the performance of TL^p_λ. A next step would be to consider a more detailed study of a specific problem. For example in the colour transfer application we could have considered regularization terms which would have improved the performance. It was not the aim to propose a state-of-the-art method for each application, indeed each application would constitute a paper within its own right.

Acknowledgements

Authors gratefully acknowledge funding from the NSF (CCF 1421502) and the NIH (GM090033, CA188938) in contributing to a portion of this work. DS also acknowledges funding by NSF (DMS-1516677). The authors are also grateful to the Center for Nonlinear Analysis of CMU for its support.

A Performance of TL^p_λ in Classification Problems with Simple and Oscillatory Signals

We compare the performance of TL^2_λ, L^2 and OT distances with respect to classification/clustering for the three classes $\{\mathcal{C}_i\}_{i=1,2,3}$ of signals defined in Figure 4. We test how each distance performs by finding the smallest number of data points

(a) Exemplar image.

(b) Original image to be coloured.

(c) TL^p_λ solution for $\lambda = 0.1$.

(d) TL^p_λ solution for $\lambda = 1$.

(e) TL^p_λ solution for $\lambda = 10$.

(f) OT colour transfer solution (no spatial information).

Figure 8: Spatially correlated histogram specification of real colour images.

such that the classes $\mathcal{C}_i^N = \{f_i\}_{i=1}^N \subset \mathcal{C}_i$ are separable. For sufficiently large N the approximation $d_{H,\rho}(\mathcal{C}_i^N, \mathcal{C}_j^N) \approx d_{H,\rho}(\mathcal{C}_i, \mathcal{C}_j)$ is used to simplify the computation. Similarly, as a proxy for $\mathbb{E}R_\rho(\mathcal{C}_i^N)$ we use $R_\rho(\hat{\mathcal{C}}^N)$ where

$$\hat{\mathcal{C}}_i^N = \left\{ f_\ell \,:\ell = \ell^i_{\min} + \frac{n-1}{N-1}\left(\ell^i_{\min} - \ell^i_{\max}\right), \right. \\ \left. n \in \{1, 2, \ldots, N\} \right\}$$

is the uniform sample from class $\mathcal{C}_i$ (recall that class $\mathcal{C}_i$ is parameterized by $\ell \in [\ell^i_{\min}, \ell^i_{\max}]$ and with an abuse of notation we use the subscript of f_ℓ to denote the dependence of ℓ).

It follows that the class separation distances and class coverage radius are approximated by

$$\begin{aligned}
d^2_{H,L^2}(\mathcal{C}_1^N, \mathcal{C}_2^N) &\approx \frac{\alpha}{2} & R^2_{L^2}(\mathcal{C}_1^N) &\approx \frac{2}{N} \\
d^2_{H,L^2}(\mathcal{C}_1^N, \mathcal{C}_3^N) &\approx \frac{3\alpha}{4} & R^2_{L^2}(\mathcal{C}_2^N) &\approx \frac{1}{N} \\
d^2_{H,L^2}(\mathcal{C}_2^N, \mathcal{C}_3^N) &\approx \frac{\alpha}{4} & R^2_{L^2}(\mathcal{C}_3^N) &\approx \frac{2\alpha}{N\gamma} \\
d^2_{H,\mathrm{OT}}(\mathcal{C}_1^N, \mathcal{C}_2^N) &\approx \frac{\beta^2\alpha}{4} & R^2_{\mathrm{OT}}(\mathcal{C}_1^N) &\approx \frac{\alpha}{N^2} \\
d^2_{H,\mathrm{OT}}(\mathcal{C}_1^N, \mathcal{C}_3^N) &\approx \frac{\beta^2\alpha}{4} & R^2_{\mathrm{OT}}(\mathcal{C}_2^N) &\approx \frac{\alpha}{N^2} \\
d^2_{H,\mathrm{OT}}(\mathcal{C}_2^N, \mathcal{C}_3^N) &\approx \frac{\alpha\gamma^2}{8} & R^2_{\mathrm{OT}}(\mathcal{C}_3^N) &\approx \frac{\alpha}{N^2} \\
d^2_{H,TL^2_\lambda}(\mathcal{C}_1^N, \mathcal{C}_2^N) &\approx \frac{\alpha}{2} & R^2_{TL^2_\lambda}(\mathcal{C}_1^N) &\approx \frac{\alpha^2}{N} \\
d^2_{H,TL^2_\lambda}(\mathcal{C}_1^N, \mathcal{C}_3^N) &\approx \frac{3\alpha}{4} & R^2_{TL^2_\lambda}(\mathcal{C}_2^N) &\approx \frac{4\alpha^2}{N}
\end{aligned}$$

$$d^2_{H,TL^2_\lambda}(\mathcal{C}^N_2, \mathcal{C}^N_3) \approx \frac{\alpha}{4} \qquad R^2_{TL^2_\lambda}(\mathcal{C}^N_3) \approx \frac{\alpha^2}{N}.$$

We have

$$\begin{aligned}
\kappa^2_{12}(L^2; N) &\approx \frac{\alpha N}{4}, & \kappa^2_{13}(L^2; N) &\approx \frac{3\gamma N}{8},\\
\kappa^2_{12}(\mathrm{OT}; N) &\approx \frac{\beta^2 N}{4}, & \kappa^2_{13}(\mathrm{OT}; N) &\approx \frac{\beta^2 N^2}{4},\\
\kappa^2_{12}(TL^2_\lambda; N) &\approx \frac{N}{8\alpha}, & \kappa^2_{13}(TL^2_\lambda; N) &\approx \frac{3N}{4\alpha},\\
\kappa^2_{23}(L^2; N) &\approx \frac{\gamma N}{8},\\
\kappa^2_{23}(\mathrm{OT}; N) &\approx \frac{\gamma^2 N^2}{8},\\
\kappa^2_{23}(TL^2_\lambda; N) &\approx \frac{N}{16\alpha}.
\end{aligned}$$

Finally we can compute N^*,

$$\begin{aligned}
N^*_{12}(L^2) &\approx \frac{4}{\alpha}, & N^*_{13}(L^2) &\approx \frac{8}{3\gamma}, & N^*_{23}(L^2) &\approx \frac{8}{\gamma}\\
N^*_{12}(\mathrm{OT}) &\approx \frac{2}{\beta}, & N^*_{13}(\mathrm{OT}) &\approx \frac{2}{\beta}, & N^*_{23}(\mathrm{OT}) &\approx \frac{\sqrt{8}}{\gamma}\\
N^*_{12}(TL^2) &\approx \frac{\alpha}{8}, & N^*_{13}(TL^2) &\approx \frac{4\alpha}{3}, & N^*_{23}(TL^2) &\approx 16\alpha
\end{aligned}$$

which for $\beta > \frac{\alpha}{2}$, $\beta > \frac{3\gamma}{4}$ and $\gamma < \frac{\sqrt{2}\alpha}{8}$ implies the ordering given Section 4.1.

B Numerical Methods

In principle any numerical method for OT capable of dealing with an arbitrary cost function can be adapted to compute TL^p_λ. Here we describe two numerical methods we used in Section 4.

B.1 Iterative Linear Programming

Here we describe the iterative linear programming method of Oberman and Ruan [38] which we abbreviate OR. Although this method is not guaranteed to find the minimum in (3) we find it works well in practice and is easier to implement than, for example, methods due to Schmitzer [50] that provably minimize (3) but require a more advanced refinement procedure. See also [34] and references therein for a multiscale descent approach.

The linear programming problem restricted to a subset $\mathcal{M} \subseteq \Omega^2_h$ is

$$\begin{aligned}
&\text{minimize:} \sum_{(i,j)\in\mathcal{M}} c_\lambda(x_i, x_j; f_h, g_h)\pi_{ij} \text{ over } \pi\\
&\text{subject to} \sum_{i\,:\,(i,j)\in\mathcal{M}} \pi_{ij} = q_j, \quad \sum_{j\,:\,(i,j)\in\mathcal{M}} \pi_{ij} = p_i
\end{aligned} \tag{LP$_h$}$$

where c_λ is given by (4). When $\mathcal{M} = \Omega^2_h$ then the TL^p_λ distance between (f_h, μ_h) and (g_h, ν_h) is the minimum to the above linear programme. Furthermore if π_h is the minimizer in the TL^p_λ distance then it is also the solution to the linear programme in (LP$_h$) for any $\mathcal{M}$ containing the support of π_h. That is if one already knows (or can reasonably estimate) the set of nodes $\mathcal{M}$ for which the optimal plan is non-zero then one need only consider the linear programme on $\mathcal{M}$. This is advantageous when $\mathcal{M}$ is a much smaller set. Motivated by Proposition 3.5 we expect to be able to write the optimal plan as a map. This implies whilst π_h has n^2 unknowns we only expect n of them to be non-zero.

The method proposed by OR is given in Algorithm 1. An initial discretisation scale h_0 is given and an estimate π_{h_0} found for the linear programme (LP$_h$) with $\mathcal{M} = \Omega^2_{h_0}$. One then iteratively finds $\mathcal{M}_r \subseteq \Omega^2_{h_r}$, where $h_r = \frac{h_{r-1}}{2}$, to be the set of nodes defined by the following refinement procedure. Find the set of nodes for which $\pi_{h_{r-1}}$ is non-zero, add the neighbouring nodes and then project onto the refined grid $\Omega^2_{h_r}$. The optimal plan π_{h_r} on $\Omega^2_{h_r}$ is then estimated by solving the linear programme (LP$_h$) with $\mathcal{M} = \mathcal{M}_r$.

The grid Ω_{h_r} will scale as $(2^{rd}h_0^{-1})^2$. If the linear programme is run N times then at the r^{th} step the linear programme has on the order of $2^{rd}h_0^{-1}$ variables. In particular on the last (and most expensive) step the number of variables is $O(2^{Nd}h_0^{-1})$. This compares to size $(2^{Nd}h_0^{-1})^2$ if the linear programme was run on the final grid without this refinement procedure.

Algorithm 1 An Iterative Linear Programming Approach [38]

Input: functions $f, g \in L^p(\Omega)$, measures $\mu, \nu \in \mathcal{P}(\Omega)$ and parameters h_0, N.

Set $r = 0$.
repeat
  Define $\mathcal{S}_r = \Omega^2_{h_r}$ where Ω_{h_r} is the square grid lattice with distances between neighbouring points h_r and discretise functions f, g and measures μ, ν on Ω_h.
  if $r = 0$ **then**
    Solve (LP$_h$) on $\mathcal{S}_0$ and call the output π_{h_0}.
  else
    Find the set of nodes on $\mathcal{S}_{r-1}$ for which $\pi_{h_{r-1}}$ is non-zero and call the set $\mathcal{K}_{r-1}$.
    To $\mathcal{K}_{r-1}$ add all neighbouring nodes and call this set $\mathcal{N}_{r-1}$.
    Define $\mathcal{M}_r$ to be the set of nodes on $\mathcal{S}_r$ that are children of nodes in $\mathcal{N}_{r-1}$.
    Solve (LP$_h$) restricted to $\mathcal{M}_r$ and call the optimal plan π_{h_r}.
  end if
  Set $h_{r+1} = \frac{h_r}{2}$ and $r \mapsto r + 1$.
until $r = N$

Output: The optimal plan $\pi_{h_{N-1}}$ for (LP$_h$).

B.2 Entropic Regularisation

Cuturi, in the context of computing optimal transport, proposed regularizing the minimization in (3) with entropy [8]. This was further developed by Benamou, Carlier, Cuturi, Nenna and Peyré [3], abbreviate to BCCNP, which is the method we de-

scribe here. Instead of considering the distance TL^p_λ we consider

$$S_\epsilon = \inf_{\pi \in \Pi(\mu,\nu)} \left\{ \sum_{i=1}^n \sum_{j=1}^n c_\lambda(x_i, x_j; f, g)\pi_{ij} - \epsilon H(\pi) \right\}$$

where $H(\pi) = -\sum_{i=1}^n \sum_{j=1}^n \pi_{ij} \log \pi_{ij}$ is the entropy. In the OT case the distance S_ϵ is also known as the Sinkhorn distance. It is a short calculation to show

$$S_\epsilon = \epsilon \inf_{\pi \in \Pi(\mu,\nu)} \{\mathrm{KL}(\pi|\mathcal{K})\}$$

where $\mathcal{K}_{ij} = \exp\left(-\frac{c_\lambda(x_i,x_j;f,g)}{\epsilon}\right)$ (the exponential is taken pointwise) and KL is the Kullback-Leibler divergence defined by

$$\mathrm{KL}(\pi|\mathcal{K}) = \sum_{i=1}^n \sum_{j=1}^n \pi_{ij} \log\left(\frac{\pi_{ij}}{\mathcal{K}_{ij}}\right).$$

It can be shown that the optimal choice of π for S_ϵ can be written in the form $\pi^* = \mathrm{diag}(u)\mathcal{K}\mathrm{diag}(v)$ where $u, v \in \mathbb{R}^n$ are limits, as $r \to \infty$, of the sequence

$$v^{(0)} = \mathbb{I}, \quad u^{(r)} = \frac{\underline{p}}{\mathcal{K}v^{(r)}}, \quad v^{(r+1)} = \frac{\underline{q}}{\mathcal{K}^\top u^{(r)}}$$

and $\underline{p} = (p_1, \ldots, p_n)$, $\underline{q} = (q_1, \ldots, q_n)$ (multiplication is the usual matrix-vector multiplication, division is pointwise and $\top$ denotes the matrix transpose). The algorithm given in 2 is a special case of iterative Bregman projections.

The stopping condition proposed in [8] is to let $\pi^{(r)} = \mathrm{diag}(u^{(r)})\mathcal{K}\mathrm{diag}(v^{(r)})$ then stop when

$$\left| \frac{\sum_{i,j=1}^n \mathcal{K}_{ij}\pi^{(r)}_{ij} - \epsilon H(\pi^{(r)})}{\sum_{i,j=1}^n \mathcal{K}_{ij}\pi^{(r-1)}_{ij} - \epsilon H(\pi^{(r-1)})} - 1 \right| < 10^{-4}.$$

Note that although as $\epsilon \to 0$ we will recover the unregularised TL^p_λ distance we also suffer numerical instability as $\mathcal{K} \to 0$ exponentially in ϵ.

Algorithm 2 An Entropy Regularised Approach [3, 8]

Input: discrete functions $f = (f_1, \ldots, f_n), g = (g_1, \ldots, g_n)$, discrete measures $\mu = \sum_{i=1}^n p_i \delta_{x_i}, \nu = \sum_{j=1}^n q_j \delta_{x_j}$, the parameter ϵ and a stopping condition.

1: Set $r = 0$, $\mathcal{K} = \left(\exp\left(-\frac{c(x_i,x_j;f,g)}{\epsilon}\right)\right)_{ij}$ and $v^{(0)} = \mathbb{I} \in \mathbb{R}^n$.
2: **repeat**
3: Let $r \mapsto r + 1$,

$$v^{(r)} = \frac{\underline{q}}{\mathcal{K}^\top u^{(r-1)}} \quad \text{and} \quad u^{(r)} = \frac{\underline{p}}{\mathcal{K}v^{(r)}}$$

where $\underline{p} = (p_1, \ldots, p_n)$, $\underline{q} = (q_1, \ldots, q_n)$.
4: **until** Stopping condition has been reached
5: Set $\pi = \mathrm{diag}(u^{(r)})\mathcal{K}\mathrm{diag}(v^{(r)})$.

Output: An estimate π on the optimal plan for S_ϵ where the accuracy is determined by the stopping condition.

References

[1] F. Åström, S. Petra, B. Schmitzer, and C. Schnörr. Image labeling by assignment. *arXiv:1603.05285*, 2016.

[2] S. Basu, S. Kolouri, and G. K. Rohde. Detecting and visualizing cell phenotype differences from microscopy images using transport-based morphometry. *Proceedings of the National Academy of Sciences*, 111(9):3448–3453, 2014.

[3] J.-D. Benamou, G. Carlier, M. Cuturi, L. Nenna, and G. Peyré. Iterative bregman projections for regularized transportation problems. *SIAM Journal on Scientific Computing*, 37(2):A1111–A1138, 2015.

[4] N. Bonneel, J. Rabin, G. Peyré, and H. Pfister. Sliced and Radon Wasserstein barycenters of measures. *Journal of Mathematical Imaging and Vision*, 51(1):22–45, 2015.

[5] Y. Brenier, U. Frisch, M. Hénon, G. Loeper, S. Matarrese, R. Mohayaee, and A. Sobolevski. Reconstruction of the early universe as a convex optimization problem. *Monthly Notices of the Royal Astronomical Society*, 346(2):501–524, 2003.

[6] L. Chizat, G. Peyré, B. Schmitzer, and F.-X. Vialard. Scaling algorithms for unbalanced optimal transport problems. *arXiv:1607.05816*, 2016.

[7] N. Courty, R. Flamary, and D. Tuia. *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2014, Nancy, France, September 15-19, 2014. Proceedings, Part I*, chapter Domain Adaptation with Regularized Optimal Transport, pages 274–289. Springer Berlin Heidelberg, 2014.

[8] M. Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems*, pages 2292–2300, 2013.

[9] L. Fei-Fei, R. Fergus, and P. Perona. Learning generative visual models from few training examples: an incremental Bayesian approach tested on 101 object categories. IEEE, CVPR 2004, Workshop on Generative-Model Based Vision, 2004.

[10] S. Ferradans, N. Papadakis, J. Rabin, G. Peyré, and J.-F. Aujol. *Scale Space and Variational Methods in Computer Vision: 4th International Conference, SSVM 2013, Schloss Seggau, Leibnitz, Austria, June 2-6, 2013. Proceedings*, chapter Regularized Discrete Optimal Transport, pages 428–439. Springer Berlin Heidelberg, 2013.

[11] U. Frisch, S. Matarrese, R. Mohayaee, and A. Sobolevski. A reconstruction of the initial conditions of the universe by optimal mass transportation. *Nature*, 417(6886):260–262, 2002.

[12] U. Frisch and A. Sobolevskii. Application of optimal transportation theory to the reconstruction of the early universe. *Journal of Mathematical Sciences (New York)*, 133(1):303–309, 2004.

[13] C. Frogner, C. Zhang, H. Mobahi, M. Araya-Polo, and T. A. Poggio. Learning with a Wasserstein loss. In *Advances in Neural Information Processing Systems (NIPS) 28*, 2015.

[14] N. García Trillos and D. Slepčev. Continuum limit of Total Variation on point clouds. *Archive for Rational Mechanics and Analysis*, pages 1–49, 2015.

[15] T. Górecki and M. Łuczak. Multivariate time series classification with parametric derivative dynamic time warping. *Expert Systems with Applications*, 42(5):2305–2312, 2015.

[16] U. Grenander and M. I. Miller. Computational anatomy: An emerging discipline. *Q. Appl. Math.*, LVI(4):617–694, 1998.

[17] E. Haber, T. Rehman, and A. Tannenbaum. An efficient numerical method for the solution of the L_2 optimal mass transfer problem. *SIAM Journal on Scientific Computing*, 32(1):197–211, 2010.

[18] S. Haker and A. Tannenbaum. On the Monge-Kantorovich problem and image warping. *IMA Volumes in Mathematics and its Applications*, 133:65–86, 2003.

[19] S. Haker, A. Tannenbaum, and R. Kikinis. *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2001: 4th International Conference Utrecht, The Netherlands, October 14–17, 2001 Proceedings*, chapter Mass Preserving Mappings and Image Registration, pages 120–127. Springer Berlin Heidelberg, 2001.

[20] S. Haker, L. Zhu, A. Tannenbaum, and S. Angenent. Optimal mass transport for registration and warping. *International Journal of Computer Vision*, 60(3):225–240, 2004.

[21] S. C. Joshi and M. I. Miller. Landmark matching via large deformation diffeomorphisms. *IEEE Transactions on Image Processing*, 9(8):1357–1370, 2000.

[22] M. W. Kadous. *Temporal classification: Extending the classification paradigm to multivariate time series*. PhD thesis, The University of New South Wales, 2002.

[23] A. M. Khan, N. Rajpoot, D. Treanor, and D. Magee. A nonlinear mapping approach to stain normalization in digital histopathology images using image-specific color deconvolution. *IEEE Transactions on Biomedical Engineering*, 61(6):1729–1738, 2014.

[24] S. Kolouri, S. Park, and G. K. Rohde. The Radon cumulative distribution transform and its application to image classification. *Image Processing, IEEE Transactions on*, 25(2):920–934, 2016.

[25] S. Kolouri, S. Park, M. Thorpe, D. Slepčev, and G. K. Rohde. Transport-based analysis, modeling, and learning from signal and data distributions. *In Preparation*, 2016.

[26] S. Kolouri and G. K. Rohde. Transport-based single frame super resolution of very low resolution face images. In *Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on*, pages 4876–4884, 2015.

[27] S. Kolouri, A. B. Tosun, J. A. Ozolek, and G. K. Rohde. A continuous linear optimal transport approach for pattern analysis in image datasets. *Pattern Recognition*, 51:453–462, 2016.

[28] J B Kruskal. Nonmetric multidimensional scaling: A numerical method. *Psychometrika*, 29(2):115–129, 1964.

[29] J. Lellmann, D. A. Lorenz, C. Schönlieb, and T. Valkonen. Imaging with Kantorovich–Rubinstein discrepancy. *SIAM Journal on Imaging Sciences*, 7(4):2833–2859, 2014.

[30] M. Lichman. UCI machine learning repository, 2013.

[31] Y. Lipman and I. Daubechies. Conformal Wasserstein distances: Comparing surfaces in polynomial time. *Advances in Mathematics*, 227(3):1047–1077, 2011.

[32] D. Magee, D. Treanor, D. Crellin, M. Shires, K. Smith, K. Mohee, and P. Quirke. Colour normalisation in digital histopathology images. In *Proc Optical Tissue Image analysis in Microscopy, Histopathology and Endoscopy (MICCAI Workshop)*, volume 100, 2009.

[33] B. M. Marlin, K. Swersky, B. Chen, and N. de Freitas. Inductive principles for restricted boltzmann machine learning. In *AISTATS*, pages 509–516, 2010.

[34] Q. Mérigot. A multiscale approach to optimal transport. *Computer Graphics Forum*, 30(5):1583–1592, 2011.

[35] G. Montavon, K.-R. Müller, and M. Cuturi. Wasserstein training of Boltzmann machines. *arXiv:1507.01972*, 2015.

[36] J. Morovic and P.-L. Sun. Accurate 3d image colour histogram transformation. *Pattern Recognition Letters*, 24(11):1725–1735, 2003.

[37] O. Museyko, M. Stiglmayr, K. Klamroth, and G. Leugering. On the application of the Monge–Kantorovich problem to image registration. *SIAM Journal on Imaging Sciences*, 2(4):1068–1097, 2009.

[38] A. M. Oberman and Y. Ruan. An efficient linear programming method for optimal transportation. *arXiv:1509.03668*, 2015.

[39] L. Oudre, J. Jakubowicz, P. Bianchi, and C. Simon. Classification of periodic activities using the Wasserstein distance. *IEEE Transactions on Biomedical Engineering*, 59(6):1610–1619, 2012.

[40] J. A. Ozolek, A. B. Tosun, W. Wang, C. Chen, S. Kolouri, S. Basu, H. Huang, and G. K. Rohde. Accurate diagnosis of thyroid follicular lesions from nuclear morphology using supervised learning. *Medical Image Analysis*, 18(5):772–780, 2014.

[41] N. Papadakis, A. Bugeau, and V. Caselles. Image editing with spatiograms transfer. *IEEE Transactions on Image Processing*, 21(5):2513–2522, 2012.

[42] S. Park, S. Kolouri, S. Kundu, and G. Rohde. The cumulative distribution transform and linear pattern classification. *arXiv:1507.05936*, 2015.

[43] O. Pele and M. Werman. Fast and robust Earth Mover's Distances. In *2009 IEEE 12th International Conference on Computer Vision*, pages 460–467, 2009.

[44] J. Rabin, S. Ferradans, and N. Papadakis. Adaptive color transfer with relaxed optimal transport. In *2014 IEEE International Conference on Image Processing (ICIP)*, pages 4852–4856, 2014.

[45] J. Rabin and N. Papadakis. *Geometric Science of Information: Second International Conference, GSI 2015, Palaiseau, France, October 28-30, 2015, Proceedings*, chapter Non-convex Relaxation of Optimal Transport for Color Transfer Between Images, pages 87–95. Springer International Publishing, 2015.

[46] J. Rabin and G. Peyré. Wasserstein regularization of imaging problem. In *Image Processing (ICIP), 2011 18th IEEE International Conference on*, pages 1521–1544, 2011.

[47] J. Rabin, G. Peyré, and L. D. Cohen. *Computer Vision – ECCV 2010: 11th European Conference on Computer Vision, Heraklion, Crete, Greece, September 5-11, 2010, Proceedings, Part V*, chapter Geodesic Shape Retrieval via Optimal Mass Transport, pages 771–784. Springer Berlin Heidelberg, 2010.

[48] Y. Rubner, C. Tomasi, and L. J. Guibas. The Earth Mover's Distance as a metric for image retrieval. *International Journal of Computer Vision*, 40(2):99–121, 2000.

[49] E. J. Russell. Letters to the editor-extension of Dantzig's algorithm to finding an initial near-optimal basis for the transportation problem. *Operations Research*, 17(1):187–191, 1969.

[50] B. Schmitzer. *Scale Space and Variational Methods in Computer Vision: 5th International Conference, SSVM 2015, Lège-Cap Ferret, France, May 31 - June 4, 2015, Proceedings*, chapter A sparse algorithm for dense optimal transport, pages 629–641. Springer International Publishing, 2015.

[51] B. Schmitzer. A sparse multi-scale algorithm for dense optimal transport. *arXiv:1510.05466*, 2016.

[52] R. T. Shinohara, E. M. Sweeney, J. Goldsmith, N. Shiee, F. J. Mateen, P. A. Calabresi, S. Jarso, D. L. Pham, D. S. Reich, and C. M. Crainiceanu. Statistical normalization techniques for magnetic resonance imaging. *NeuroImage: Clinical*, 2014.

[53] J. Solomon, F. de Goes, G. Peyré, M. Cuturi, A. Butscher, A. Nguyen, T. Du, and L. Guibas. Convolutional Wasserstein distances: Efficient optimal transportation on geometric domains. *ACM Trans. Graph.*, 34(4):66:1–66:11, 2015.

[54] J. Solomon, R. Rustamov, L. Guibas, and A. Butscher. Earth mover's distances on discrete surfaces. *ACM Transactions on Graphics (TOG)*, 33(4):67, 2104.

[55] J. Solomon, R. Rustamov, G. Leonidas, and A. Butscher. Wasserstein propagation for semi-supervised learning. In T. Jebara and E. P. Xing, editors, *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pages 306–214. JMLR Workshop and Conference Proceedings, 2014.

[56] Z. Su, W. Zeng, Y. Wang, Z.-L. Lu, and X. Gu. Shape classification using Wasserstein distance for brain morphometry analysis. In *Information Processing in Medical Imaging*, volume 24, pages 411–423, 2015.

[57] E. Tannenbaum, T. Georgiou, and A. Tannenbaum. Signals and control aspects of optimal mass transport and the Boltzmann entropy. In *49th IEEE Conference on Decision and Control (CDC)*, pages 1885–1890, 2010.

[58] M. Thorpe and D. Slepčev. Transportation L^p distances: Properties and extensions. *In Preparation*, 2016.

[59] A. B. Tosun, O. Yergiyev, S. Kolouri, J. F. Silverman, and G. K. Rohde. Novel computer-aided diagnosis of mesothelioma using nuclear structure of mesothelial cells in effusion cytology specimens. In *Proc. SPIE*, volume 9041, pages 90410Z–90410Z–6, 2014.

[60] A. B. Tosun, O. Yergiyev, S. Kolouri, J. F. Silverman, and G. K. Rohde. Detection of malignant mesothelioma using nuclear structure of mesothelial cells in effusion cytology specimens. *Cytometry Part A*, 87(4):326–333, 2015.

[61] T. ur Rehman, E. Haber, G. Pryor, J. Melonakos, and A. Tannenbaum. 3D nonrigid registration via optimal mass transport on the gpu. *Medical image analysis*, 13(6):931–940, 2009.

[62] C. Villani. *Topics in Optimal Transportation*. Graduate Studies in Mathematics. American Mathematical Society, 2003.

[63] W. Wang, D. Slepčev, S. Basu, J. A. Ozolek, and G. K. Rohde. A linear optimal transportation framework for quantifying and visualizing variations in sets of images. *International Journal of Computer Vision*, 101(2):254–269, 2012.

[64] L. Zhu, S. Haker, and A. Tannenbaum. Flattening maps for the visualization of multibranched vessels. *Medical Imaging, IEEE Transactions on*, 24(2):191–198, 2005.

[65] L. Zhu, Y. Yang, S. Haker, and A. Tannenbaum. An image morphing technique based on optimal mass preserving mapping. *Image Processing, IEEE Transactions on*, 16(6):1481–1495, 2007.