2019

Entropic Regularisation of Robust Optimal Transport

Rozenn Dahyot
Trinity College Dublin, Ireland

Hana Alghamdi
Trinity College Dublin Ireland

Mairead Grogan
Trinity College Dublin, Ireland

Recommended Citation

Dahyot, R., Alghamdi, H. & Grogan, M. (2019). Entropic regularisation of robust optimal transport. *IMVIP 2019: Irish Machine Vision & Image Processing*, Technological University Dublin, Dublin, Ireland, August 28-30. doi:10.21427/w611-mb37

Entropic Regularisation of Robust Optimal Transport

Rozenn Dahyot, Hana Alghamdi and Mairead Grogan

School of Computer Science and Statistics
Trinity College Dublin, Ireland
Rozenn.Dahyot@tcd.ie, alghamdh@tcd.ie, groganma@tcd.ie

Abstract

Grogan et al. [11, 12] have recently proposed a solution to colour transfer by minimising the Euclidean distance $\mathscr{L}_2$ between two probability density functions capturing the colour distributions of two images (palette and target). It was shown to be very competitive to alternative solutions based on Optimal Transport for colour transfer. We show that in fact Grogan et al's formulation can also be understood as a new robust Optimal Transport based framework with entropy regularisation over marginals.

Keywords: M-estimation, $\mathscr{L}_2E$ estimator, Optimal Transport, Colour Transfer

1 Introduction

Optimal transport (OT) [16] has been successfully used as a way for defining cost functions for optimisation when performing colour transfer [17] (cf. Fig. 2) and more recently in machine learning [4, 16]. The optimal transport cost (e.g Wasserstein distance) itself is also used as a similarity metric for retrieval [18]. For colour transfer, Grogan et al. [11, 12] have recently proposed an alternative approach for designing the cost function based on the $\mathscr{L}_2$ divergence (see section 2) allowing user interaction (see Fig. 1[1]). This $\mathscr{L}_2$ based cost function is a weighted sum of multiple terms (terms $\mathscr{T}_{0,1,2,3}$ in Eq. 1) able to take into account correspondences between images (via term $\mathscr{T}_3$ in Eq. 1) when these are available, as well as the unsupervised scenario when no correspondence is available (via term $\mathscr{T}_2$ in Eq. 1). In addition, $\mathscr{L}_2$ includes entropies (terms $\mathscr{T}_0$ and $\mathscr{T}_1$ in Eq. 1). To further constrain the cost function when estimating the colour transformation ϕ, additional penalties can be added to prevent colours exceeding a certain range or forcing the estimated solution ϕ to be smooth (resp. terms $\mathscr{T}_4$ and $\mathscr{T}_5$ in Eq. 1). The estimate $\hat{\phi}$ is computed as

Figure 1: Colour correspondences with colour palettes allows user interactions when recolouring a reference image (top left) [12].

$$\hat{\phi} = \arg\min_{\phi} \mathscr{C}(\phi)$$

[1] Images extracted from video `https://youtu.be/FfrdyKMBVRc` (demo for [12]) have been used for designing Fig. 1.

with:

$$\mathscr{C}(\phi) = \left| \begin{array}{ll} +\frac{1}{n^2}\sum_{j_1=1}^{n}\sum_{j_2=1}^{n}\mathscr{N}(0; y^{(j_1)} - y^{(j_2)}, 2h^2\mathrm{I}) & (\mathscr{T}_0) \\ +\frac{1}{\tilde{n}^2}\sum_{i_1=1}^{\tilde{n}}\sum_{i_2=1}^{\tilde{n}}\mathscr{N}(0; \tilde{y}^{(i_1)} - \tilde{y}^{(i_2)}, 2h^2\mathrm{I}) & (\mathscr{T}_1) \\ -\frac{2}{n\tilde{n}}\sum_{i=1}^{\tilde{n}}\sum_{j=1}^{n}\mathscr{N}(0; \tilde{y}^{(i)} - y^{(j)}, 2h^2\mathrm{I}) & (\mathscr{T}_2) \\ -\lambda_1\frac{1}{\tilde{\tilde{n}}}\sum_{k=1}^{\tilde{\tilde{n}}}\mathscr{N}(0; \tilde{y}^{(k)} - y^{(k)}, 2h^2\mathrm{I}) & (\mathscr{T}_3) \\ +\lambda_2\,\mathscr{P}(\tilde{y}) & (\mathscr{T}_4) \\ +\lambda_3\,\mathscr{P}(\phi) & (\mathscr{T}_5) \end{array} \right. \quad (1)$$

with $\mathscr{N}(z; a, \Sigma)$ indicating a normal distribution for random vector z with expectation a and covariance matrix Σ. I is the identity matrix, h is a user defined bandwidth and $\lambda_{1,2,3}$ are weights. This paper aims at proposing an OT formulation for the terms $\mathscr{T}_2$ and $\mathscr{T}_3$ (see Sec. 3) as an alternative to $\mathscr{L}_2$ (presented in Sec. 2). In particular we show that these terms corresponds to robust Wasserstein distances where the bandwidth h (Eq. 1) enables the seamless control of the level of robustness in a similar fashion as the scale parameter controlling M-estimators [13]. This reformulation allows the following contributions: first, to extend OT in supervised and semi-supervised scenarios, and second to propose a robust Wasserstein cost (Sec. 3). We start first by explaining in more detail the notations used and the $\mathscr{L}_2$ cost function.

2 $\mathscr{L}_2$ divergence

We consider that the following are available:

- a dataset $\mathscr{S} = \left\{y^{(j)}\right\}_{j=1,\cdots,n}$: the term $\mathscr{T}_0$ (Eq. 1) uses the samples from this dataset.
- a dataset $\widetilde{\mathscr{S}} = \left\{\tilde{y}^{(i)} = \phi(x^{(i)})\right\}_{i=1,\cdots,\tilde{n}}$ computed using a transfer (or mapping) function ϕ on data points $\{x^{(i)}\}_{i=1,\cdots,\tilde{n}}$. The term $\mathscr{T}_1$ (Eq. 1) uses the samples from this dataset.
- a dataset of correspondences $\widetilde{\widetilde{\mathscr{S}}} = \left\{(y^{(k)}, \tilde{y}^{(k)} = \phi(x^{(k)}))\right\}_{k=1,\cdots,\tilde{\tilde{n}}}$: the term $\mathscr{T}_3$ (Eq. 1) uses the samples from this dataset.

All data points have the same dimension (i.e. $\dim(y^{(l_1)}) = \dim(\tilde{y}^{(l_2)})$) for any samples taken from $\mathscr{S}$, $\widetilde{\mathscr{S}}$ or $\widetilde{\widetilde{\mathscr{S}}}$. Figure 2 shows an illustration of our datasets in the context of colour transfer[2]. In this $\mathscr{L}_2$ framework [11], only one random vector (r.v.) y is defined. Using $\mathscr{S}$ and $\widetilde{\mathscr{S}}$, two probability density functions noted $\mu(y)$ and $\tilde{\mu}(y|\phi)$ respectively are computed for r.v. y as kernel density estimates with a Normal kernel (or Gaussian Mixture Models):

$$\mu(y) = \frac{1}{n}\sum_{y^{(j)}\in\mathscr{S}}\mathscr{N}(y; y^{(j)}, h^2)$$

and

$$\tilde{\mu}(y|\phi) = \frac{1}{\tilde{n}}\sum_{y^{(i)}\in\widetilde{\mathscr{S}}}\mathscr{N}(y; \tilde{y}^{(i)}, h^2)$$

The unknown mapping function ϕ transforms the samples in $\widetilde{\mathscr{S}}$ that act as the means of the normal kernels in the mixture $\tilde{\mu}(y|\phi)$. Hence, $\tilde{\mu}$ can be warped onto μ by finding the appropriate function ϕ. The best choice for

[2]Images from the video posted at `https://twitter.com/gabrielpeyre/status/797605863295053826` have been used for designing Fig. 2.

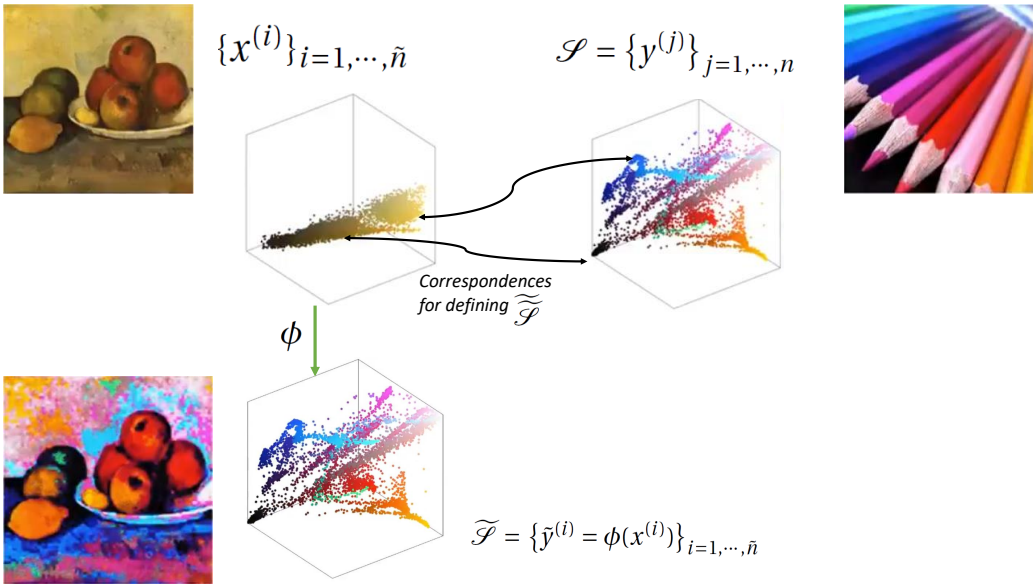


Figure 2: Colour transfer aims at changing the colour feel of a target image (top left) by using the colour content of an other source (top right image) to produce a recoloured image (bottom left). Data correspond to triplets of values (e.g. see the coloured coded 3D point clouds in the RGB cube associated with each image): the mapping or transfer function ϕ is estimated so that the point cloud $\widetilde{\mathscr{S}}$ overlaps well with $\mathscr{S}$.

function ϕ can be chosen as minimising the Euclidean $\mathscr{L}_2$ distance between μ and $\tilde{\mu}$ defined as [14]:

$$\mathscr{L}_2(\mu,\tilde{\mu}) = \|\mu-\tilde{\mu}\|^2 = \int (\mu(y)-\tilde{\mu}(y|\phi))^2\, dy = \underbrace{\|\mu\|^2}_{\mathscr{T}_0} \underbrace{-2\langle\mu|\tilde{\mu}\rangle}_{\mathscr{T}_2} + \underbrace{\|\tilde{\mu}\|^2}_{\mathscr{T}_1} \tag{2}$$

from which terms $\mathscr{T}_{0,1,2}$ in the cost function $\mathscr{C}(\phi)$ originate (Eq. 1). Such a formulation of $\mathscr{L}_2$ has been used for colour transfer [11] and shape registration [14, 1]. The connection between $\mathscr{L}_2$ with robust M-estimators has also been shown [3, 19, 14].

Removing $\mathscr{T}_0$ from the cost function $\mathscr{C}$. $\mathscr{T}_0$ does not depends on ϕ and can be discarded, shortening $\mathscr{L}_2$ into $\mathscr{L}_2E$ [19] for estimating ϕ. Both $\mathscr{T}_0$ and $\mathscr{T}_1$ correspond to entropies since $-\log\left(\|\mu\|^2\right)$ and $-\log\left(\|\tilde{\mu}\|^2\right)$ are the quadratic Renyi entropies of μ and $\tilde{\mu}$ respectively [11].

Using correspondences. The term $\mathscr{T}_3$ to account for correspondences in $\widetilde{\widetilde{\mathscr{S}}}$, is explained intuitively with notation $-2\langle\mu|\tilde{\mu}\rangle$ by Grogan et al [12], where this time μ and $\tilde{\mu}$ are likewise kernel density estimates (with Normal kernel) using only observations in the dataset of correspondences $\widetilde{\widetilde{\mathscr{S}}}$:

$$\tilde{\mu}(y) = \frac{1}{\tilde{\tilde{n}}} \sum_{y^{(k_1)}\in\tilde{\tilde{\mathscr{S}}}} \mathscr{N}(y; y^{(k_1)}, h^2\mathrm{I}) \quad \text{and} \quad \tilde{\mu}(y) = \frac{1}{\tilde{\tilde{n}}} \sum_{y^{(k_2)}\in\tilde{\tilde{\mathscr{S}}}} \mathscr{N}(y; \tilde{y}^{(k_2)}, \tilde{h}^2\mathrm{I})$$

and the scalar product $\langle\mu|\tilde{\mu}\rangle$ then corresponds to:

$$\langle\mu|\tilde{\mu}\rangle = \frac{1}{\tilde{\tilde{n}}^2} \sum_{y^{(k_1)}\in\tilde{\tilde{\mathscr{S}}}} \sum_{y^{(k_2)}\in\tilde{\tilde{\mathscr{S}}}} \mathscr{N}(y^{(k_1)}; \tilde{y}^{(k_2)}, (h^2+\tilde{h}^2)\mathrm{I}) \tag{3}$$

Hence the notation $\langle\mu|\tilde{\mu}\rangle$ is not mathematically correct to explain $\mathscr{T}_3$ (i.e note the single sum for $\mathscr{T}_3$ in Eq. 1 versus the double sum appearing in Eq. 3). So even if the intuition for $\mathscr{T}_3$ is sound and proves to be efficient in practice against the state of the art techniques for colour transfer [12, 11], its origin cannot be explained mathematically with $\mathscr{L}_2$ and we provide next a better explanation for $\mathscr{T}_3$ based on Optimal Transport.

3 Optimal Transport

We propose to reformulate both $\mathscr{T}_2$ and $\mathscr{T}_3$ from an OT perspective. OT aims at choosing ϕ with the minimum transport (displacement) cost between two random vectors noted y and $\tilde{y}$. The OT cost function is expressed here with the Wasserstein distance [16] as follow:

$$\mathscr{W}(\mu,\tilde{\mu}) = \min_{\gamma} \left\{ \int\int c(y,\tilde{y})\ \gamma(y,\tilde{y})\ dy\ d\tilde{y} = \langle c|\gamma\rangle \right\} \tag{4}$$

where c is a cost often chosen as $c(y,\tilde{y}) = \|y-\tilde{y}\|^2$ (quadratic Wasserstein distance), and γ is the joint probability density function of y and $\tilde{y}$ having μ and $\tilde{\mu}$ for marginals respectively i.e. $\int \gamma(y,\tilde{y})\ dy = \tilde{\mu}(\tilde{y})$ and $\int \gamma(y,\tilde{y})\ d\tilde{y} = \mu(y)$. We first present our choices for these distributions (Sec. 3.1) and then propose a new robust cost (in Sec. 3.2). An alternative OT based explanation for terms $\mathscr{T}_2$ and $\mathscr{T}_3$ then emerges (Sec. 3.3).

3.1 Models for γ_ϕ, μ and $\tilde{\mu}_\phi$

Kernel density estimates with Normal kernels are used as joint density functions γ_ϕ and using the datasets available, three estimates of $\gamma_\phi \in \{\gamma_u, \gamma_s, \gamma_{s+u}\}$ can be proposed:

- using independent sets $\mathscr{S}$ and $\widetilde{\mathscr{S}}$ (unsupervised scenario i.e without correspondences):

$$\gamma_u(y,\tilde{y}|\phi) = \left(\frac{1}{n} \sum_{y^{(j)}\in\mathscr{S}} \mathcal{N}(y; y^{(j)}, h^2\mathrm{I}) \right) \times \left(\frac{1}{\tilde{n}} \sum_{\tilde{y}^{(i)}\in\widetilde{\mathscr{S}}} \mathcal{N}(\tilde{y}; \tilde{y}^{(i)}, \tilde{h}^2\mathrm{I}) \right) \tag{5}$$

  with the marginals

$$\mu_u(y) = \frac{1}{n} \sum_{y^{(j)}\in\mathscr{S}} \mathcal{N}(y; y^{(j)}, h^2\mathrm{I})$$

  and

$$\tilde{\mu}_u(\tilde{y}|\phi) = \frac{1}{\tilde{n}} \sum_{\tilde{y}^{(i)}\in\widetilde{\mathscr{S}}} \mathcal{N}(\tilde{y}; \tilde{y}^{(i)}, \tilde{h}^2\mathrm{I})$$

- using the set of correspondences $\widetilde{\widetilde{\mathscr{S}}}$ (supervised):

$$\gamma_s(y,\tilde{y}|\phi) = \frac{1}{\tilde{\tilde{n}}} \sum_{(y^{(k)},\tilde{y}^{(k)})\in\widetilde{\widetilde{\mathscr{S}}}} \mathcal{N}(y; y^{(k)}, h^2\mathrm{I})\ \mathcal{N}(\tilde{y}; \tilde{y}^{(k)}, \tilde{h}^2\mathrm{I}) \tag{6}$$

  providing the marginals

$$\mu_s(y) = \frac{1}{\tilde{\tilde{n}}} \sum_{y^{(k)}\in\widetilde{\widetilde{\mathscr{S}}}} \mathcal{N}(y; y^{(k)}, h^2\mathrm{I})$$

  and:

$$\tilde{\mu}_s(\tilde{y}|\phi) = \frac{1}{\tilde{\tilde{n}}} \sum_{\tilde{y}^{(k)}\in\widetilde{\widetilde{\mathscr{S}}}} \mathcal{N}(\tilde{y}; \tilde{y}^{(k)}, \tilde{h}^2\mathrm{I})$$

- Using all datasets, the following mixture can be considered (semi-supervised):

$$\gamma_{s+u}(y,\tilde{y}|\phi) = (1-\lambda)\ \gamma_u(y,\tilde{y}|\phi) + \lambda\ \gamma_s(y,\tilde{y}|\phi) \tag{7}$$

  where $0 \le \lambda \le 1$ is a parameter controlling the importance between the estimates γ_u and γ_s. In this case, the marginals are:

$$\mu_{s+u}(y) = (1-\lambda)\ \mu_u(y) + \lambda\ \mu_s(y)$$

  and

$$\tilde{\mu}_{s+u}(\tilde{y}|\phi) = (1-\lambda)\ \tilde{\mu}_u(\tilde{y}|\phi) + \lambda\ \tilde{\mu}_s(\tilde{y}|\phi)$$

Note that these models noted $\gamma_\phi \in \{\gamma_u, \gamma_s, \gamma_{s+u}\}$ are parameterized by ϕ via the samples $\tilde{y}^{(l)}$ in $\widetilde{\mathscr{S}}$ and $\widetilde{\widetilde{\mathscr{S}}}$. The bandwidths h and $\tilde{h}$ are user defined and using $h = \tilde{h} = 0$ enables the recovery of the empirical pdf estimates with Dirac kernels.

3.2 Robust cost $c_G(y, \tilde{y})$

Concave functions g to define costs c of the form $c(y, \tilde{y}) = g(|y - \tilde{y}|)$ have been suggested for robustness [8]. Here, we go further by proposing the following robust cost:

$$c_G(y, \tilde{y}) = A - \mathcal{N}(y; \tilde{y}, h_c^2 \mathrm{I}) \tag{8}$$

where A is a constant that can be added if one need to enforce a positive cost c_G. Our cost c_G is convex near the origin $\|y - \tilde{y}\| \sim 0$ and then becomes concave as the difference $\|y - \tilde{y}\|$ increases. We also note that:

$$\langle c_G | \gamma \rangle = A - \int\int \mathcal{N}(y; \tilde{y}, h_c^2 \mathrm{I})\ \gamma(y, \tilde{y})\ dy\ d\tilde{y} \tag{9}$$

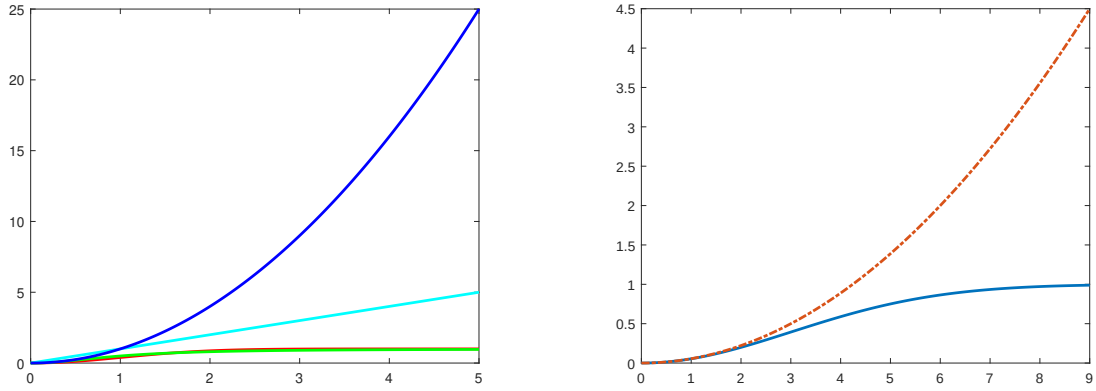
since γ integrates to 1 by definition. In practice, for estimation of ϕ that minimizes this cost, the constant A does not matter and can be set $A = 0$.

3.2.1 Relation to M-estimators

With the more familiar notation for error $\epsilon = \|y - \tilde{y}\|$, our robust cost c_G is proportional to the Welsch-Leclerc loss ρ_G [2]:

$$\rho_G(\epsilon) = 1 - \exp\left(-\frac{1}{2}\left(\frac{\epsilon}{\sigma}\right)^2\right) \tag{10}$$

which is a well-known hard redescending M-estimating function with scale parameter $\sigma = h_c$ [13, 15, 7, 2]. The more the chosen function ρ penalises large errors ϵ, the more it is robust to outliers. See for instance in Fig. 3(a) how the hard redescending functions ρ_{GM} (for Geman-McClure loss [6, 2]) and ρ_G have an upper finite limit (equal to 1) when $\epsilon \to +\infty$ and thus prevent high residuals (outliers) to overly contribute too much when estimating $\hat{\phi}$. The non-robust Least Square function ρ_{LS} is also shown and corresponds here to the quadratic Wasserstein cost $c(y, \tilde{y}) = \|y - \tilde{y}\|^2$ that is not robust to gross errors.



(a) (b)

Figure 3: Left (a): Functions $\rho_{ls}(\epsilon) = \epsilon^2$ (blue), $\rho_{abs}(\epsilon) = |\epsilon|$ (cyan), Welsch-Leclerc $\rho_G(\epsilon) = 1 - \exp\left(-\epsilon^2/2\right)$ (red) and Geman-McClure $\rho_{GM}(\epsilon) = \frac{\epsilon^2}{\epsilon^2+1}$ (green). Right (b): comparison ρ_G (solid blue line) with its Taylor approximation (orange dash dot) with scale $\sigma = 3$ for a range $\epsilon \in (0; 3\sigma)$ - the approximation is good for $\left|\frac{\epsilon}{\sigma}\right| << 1$.

3.2.2 Relation of robust cost c_G to Wasserstein distance

When the bandwidth h_c (or scale parameter σ) is very very large compared to ϵ, using Taylor approximation of the cost shows that (cf. Fig. 3(b)):

$$\rho_G(\epsilon) = 1 - \exp\left(-\frac{\epsilon^2}{2\sigma^2}\right) \sim \frac{\epsilon^2}{2\sigma^2} \quad \text{for} \quad \epsilon << \sigma \tag{11}$$

making our cost c_G proportional to the one used in the quadratic Wasserstein distance. The bandwidth h_c allows for the modulation of the cost from the non robust Euclidean distance ($h_c \to \infty$) to a more robust cost (h_c small) for penalising high differences $\|y - \tilde{y}\|$ (or outliers).

3.3 OT perspective for terms $\mathscr{T}_2$ and $\mathscr{T}_3$

Using the definitions of our cost $c_G = -\mathcal{N}(y; \tilde{y}, h_c^2 \mathrm{I})$ and our joint probability density functions $\gamma_\phi \in \{\gamma_u, \gamma_s, \gamma_{s+u}\}$ (cf. Sec. 3.1), we note that:

$$\langle c_G | \gamma_u \rangle = \frac{1}{n\tilde{n}} \sum_{i=1}^{\tilde{n}} \sum_{j=1}^{n} \mathcal{N}\left(0; \tilde{y}^{(i)} - y^{(j)}, (h^2 + \tilde{h}^2 + h_c^2)\ \mathrm{I}\right) \tag{12}$$

hence it is equivalent to the term $\mathscr{T}_2$ (since the bandwidths are user defined). Likewise we note

$$\langle c_G | \gamma_s \rangle = \frac{1}{n_c} \sum_{k=1}^{n_c} \mathcal{N}\left(0; \tilde{y}^{(k)} - y^{(k)}, (h^2 + \tilde{h}^2 + h_c^2)\ \mathrm{I}\right) \tag{13}$$

which is equivalent to $\mathscr{T}_3$ (Eq. 1) introduced by Grogan et al to take advantage of correspondences [11]. Since the weight λ_1 was chosen in an ad hoc fashion, we can propose a more elegant alternative form combining $\mathscr{T}_2$ and $\mathscr{T}_3$ into a new term $\mathscr{T}$ using the estimate γ_{s+u}:

$$\langle c_G | \gamma_{s+u} \rangle = (1 - \lambda)\ \langle c_G | \gamma_u \rangle + \lambda\ \langle c_G | \gamma_s \rangle \qquad (\mathscr{T}) \tag{14}$$

With the OT formulation (Eq. 4), Grogan et al's estimation (terms $\mathscr{T}_2$ and $\mathscr{T}_3$, Eq. 1) can be rewritten:

$$\hat{\phi} = \arg\min_{\phi} \left\{ \mathcal{W}(\mu_{s+u}, \tilde{\mu}_{s+u}) = \langle c_G | \gamma_{s+u} \rangle \right\} \tag{15}$$

to which entropic terms on the marginals μ and $\tilde{\mu}$ ($\mathscr{T}_0$ and $\mathscr{T}_1$) can be added along with other constraints on ϕ (e.g. $\mathscr{T}_4$ and $\mathscr{T}_5$).

When setting $h = \tilde{h} = 0$ for simplicity (i.e. using empirical pdf estimates with Dirac kernels, Sec. 3.1), Grogan et al's terms $\mathscr{T}_2$ and $\mathscr{T}_3$ are robust OT distances where the parameter h_c in the robust cost c_G controls the influence of outliers when performing estimation of the mapping function ϕ in the same way as the scale parameter for M-estimation.

3.4 Parametric Modelling of the transfer function ϕ

In practice, a parametric form of ϕ is used: Thin Plate Splines (TPS) have been used for colour transfer and shape registration [14, 11, 12]. The term $\mathscr{T}_5$ in Eq. 1 corresponds to a smoothness constraint on the TPS solution [14, 11]:

$$\mathscr{T}_5 = \lambda_3 \int \left\| \frac{\partial^2 \phi(x)}{\partial^2 x} \right\|^2 dx \tag{16}$$

However TPS is not a convenient formulation when modelling transfer functions in high dimensional spaces and Deep Neural Networks are now providing more powerful formulations for ϕ.

3.5 Interpretation and Generalization of the cost c_G

Our formulation of OT is equivalent to :

$$\hat{\phi} = \arg\max_{\phi} \int \int \mathcal{N}(y; \tilde{y}, h_c^2 \mathrm{I})\ \gamma_\phi(y, \tilde{y})\ dy\ d\tilde{y} \tag{17}$$

where more generally $\mathcal{N}(y; \tilde{y}, h_c^2 \mathrm{I})$ can be understood as a conditional pdf (y given $\tilde{y}$ or vice versa since the Normal distribution is symmetric w.r.t. its mean). Using a flat prior for $\tilde{y}$ (e.g. $\tilde{y} \sim \mathcal{N}(\tilde{y}; 0, a\mathrm{I})$ with bandwidth

a very large to approximate a flat prior), then a model for the joint probability density function is available $\gamma_m(y, \tilde{y}) = \mathcal{N}(y; \tilde{y}, h_c^2\mathrm{I}) \times \mathcal{N}(\tilde{y}; 0, a\mathrm{I})$ and our OT formulation (Eq. 17) is equivalent to:

$$\hat{\phi} = \arg\max_{\phi} \langle \gamma_m | \gamma_\phi \rangle \tag{18}$$

which has the same form as the cross product $\langle \mu | \tilde{\mu} \rangle$ appearing in $\mathscr{L}_2$ (cf. Eq. 2): as indicated in [11], the main difference between the two frameworks lies in the modelling of one r.v. (y in $\mathscr{L}_2$, with notation $\langle .|. \rangle$ indicating integration over this one vector) or two r.v. (y and $\tilde{y}$ in OT, $\langle .|. \rangle$ indicating integration over these two vectors). These scalar products between probability densities functions (joint, marginals or conditionals) are frequent for robust estimation including for instance the Hough transform widely used in image processing [5, 7, 9]. While some robust costs can be identified as a negative log likelihood [6, 2], we identify directly our robust cost c_G as a negative Multivariate Normal distribution instead.

4 Final remarks

We have proposed a new generic formulation for Optimal Transport with the following advantages:

- it is robust: our new robust cost $c_G(y, \tilde{y}) = -\mathcal{N}(y; \tilde{y}, h_c^2\mathrm{I})$ is parameterised by a bandwidth h_c that acts like the scale parameter of M-estimators. This bandwidth enables the control of the level of robustness and when chosen very large, it makes our cost converge towards the standard (non robust) quadratic Wasserstein distance.
- Our formulation can seamlessly consider various scenarios e.g. unsupervised, supervised (with correspondences) or semi-supervised depending on the dataset(s) available.
- Grogan et al [11] propose the use of entropy terms for the marginals (e.g. $\tilde{\mu}$) that can be used in addition to (or instead of) an entropy on the joint pdf γ [16].
- More generally, we have shown the commonality of these formulations ($\mathscr{L}_2$ and OT) in using scalar products between two p.d.fs. The main difference between $\mathscr{L}_2$ and OT is then in the number of random vectors used in the formulation of this scalar product. We believe this thinking extends to the Gromov-Wasserstein formulation which defines 4 random vectors [20].

Beyond the impact of our formulation for colour transfer [12, 11], future work will investigate shape registration with correspondences (e.g. for user interactions) and with kernels other than Gaussian better suited to directional data [10]. Finally this analogy between L2, M-estimators and Optimal transport also points to the suitability of well known optimisation algorithms such as the Iterative Reweighted Least Squares (IRLS) [13, 6] for the robust estimation of a transfer function ϕ as a solution to optimal transport.

Acknowledgments

This work is partly supported by a scholarship from Umm Al-Qura University, Saudi Arabia, and in part by a research grants from Science Foundation Ireland (SFI) (Grant Number 15/RP/2776), and the ADAPT Centre for Digital Content Technology (www.adaptcentre.ie) that is funded under the SFI Research Centres Programme (Grant 13/RC/2106) and is co-funded under the European Regional Development Fund.

References

[1] C. Arellano and R. Dahyot. Robust ellipse detection with gaussian mixture models. *Pattern Recognition*, 58:12 – 26, 2016.

[2] J. T. Barron. A more general robust loss function. *CoRR*, abs/1701.03077, 2017.

[3] A. Basu, I. R. Harris, N. L. Hjort, and M. C. Jones. Robust and efficient estimation by minimising a density power divergence. *Biometrika*, 85(3):549–559, 1998.

[4] N. Courty, R. Flamary, D. Tuia, and A. Rakotomamonjy. Optimal transport for domain adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(9):1853–1865, Sep. 2017.

[5] R. Dahyot. Statistical hough transform. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(8):1502–1509, Aug 2009.

[6] R. Dahyot, P. Charbonnier, and F. Heitz. A bayesian approach to object detection using probabilistic appearance-based models. *Pattern Analysis and Applications*, 7(3):317–332, Dec 2004.

[7] R. Dahyot and J. Ruttle. Generalised relaxed radon transform (gr2t) for robust inference. *Pattern Recognition*, 46(3):788 – 794, 2013.

[8] J. Delon, J. Salomon, and A. Sobolevski. Local matching indicators for transport problems with concave costs. *SIAM Journal on Discrete Mathematics*, 26(2):801–827, 2012.

[9] A. Goldenshluger and A. Zeevi. The hough transform estimator. *The Annals of Statistics*, 32(5), October 2004.

[10] M. Grogan and R. Dahyot. Shape registration with directional data. *Pattern Recognition*, 79:452 – 466, 2018.

[11] M. Grogan and R. Dahyot. L2 divergence for robust colour transfer. *Computer Vision and Image Understanding*, 181:39 – 49, 2019.

[12] M. Grogan, R. Dahyot, and A. Smolic. User interaction for image recolouring using $\mathscr{L}_2$. In *Proceedings of the 14th European Conference on Visual Media Production (CVMP 2017)*, CVMP 2017, pages 6:1–6:10, New York, NY, USA, 2017. ACM.

[13] P.J. Huber. *Robust Statistics*. John Wiley and Sons, 1981.

[14] B. Jian and B. C. Vemuri. Robust point set registration using gaussian mixture models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(8):1633–1645, Aug 2011.

[15] S. Mittal, S. Anand, and P. Meer. Generalized projection-based m-estimator. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(12):2351–2364, Dec 2012.

[16] G. Peyré and M. Cuturi. Computational optimal transport. *Foundations and Trends in Machine Learning*, 11(5-6):355–607, 2019.

[17] F. Pitié, A. C. Kokaram, and R. Dahyot. Automated colour grading using colour distribution transfer. *Computer Vision and Image Understanding*, 107(1):123 – 137, 2007. Special issue on color image processing.

[18] Y. Rubner, C. Tomasi, and L. J. Guibas. The earth mover's distance as a metric for image retrieval. *Int. J. Comput. Vision*, 40(2):99–121, November 2000.

[19] D. W. Scott. Parametric statistical modeling by minimum integrated square error. *Technometrics*, 43(3):pp. 274–285, 2001.

[20] J. Solomon, G. Peyré, V. G. Kim, and S. Sra. Entropic metric alignment for correspondence problems. *ACM Trans. Graph.*, 35(4):72:1–72:13, July 2016.